# 15世紀朝鮮語の形態素解析について

須賀井義教・村田　寛

## 1．はじめに

本研究の目的は、オープンソース形態素解析エンジンであるMeCab[1]を用いて15世紀朝鮮語の形態素解析を行い、その解析率を考察し、実用に耐えうる解析率を得るための諸条件を探ることである。現在、15世紀朝鮮語の研究において用いられる電子データは、入力者によってその入力形式、方法が統一されておらず、非常に多くの問題を抱えているといえる。例えば既に須賀井義教（2009）で指摘したように、語節の途中で改行を入れる、漢字表記の語に漢字音も併記する、声調[2]を示す傍点に関わる情報が入力されていない、などといった点である。このようなデータでは計量的な研究を行うことは難しい。入力がまちまちであれば、検索してもヒットするものとしないものとが出てくるおそれがある。出現の頻度などを計算するにしても、その数を信用することができないことになる。言語史研究における電子データの有用性が高まっている現在、そのもととなる電子データ自体を整備することは緊急の課題である。

こうした現状をふまえ、村田寛（2010）では形態素解析エンジンMeCabを利用して、15世紀朝鮮語の形態素解析が可能であることを明らかにした。そこでは、15世紀朝鮮語のいくつかの文献データを用いて解析を試みており、その解析率については今後の課題としている。しかし、実際にMeCabを利用して15世紀朝鮮語の形態素解析を行うならば、実用性という面でその解析率にも注意を払わなければならない。そこで、本研究ではMeCabを用いて15世紀朝鮮語の形態素解析を行うと同時に、その解析率に注目し、解析率を向上させるために必要な諸条件を探ってみることにする。

## 2．考察の方法

MeCabで日本語を解析するための辞書は既に配布されているが、新たな言語の解析を行うためには、そのための辞書を用意しなければならない。さらに学習用データを用意して、パラメータ値の学習を行う必要がある。このようにして構築した辞書を用いてデータを解析し、その結果を「正解」と比較することで、解析率を求めることができる。「正解」とはデータを手作業で解析した結果である。MeCabの配布パッケージには、評価のためのスクリプトが同梱されている。評価については後述する。

ところで、解析率を上げるために考えられるのは、(a)辞書の登録項目を増やす、(b)学習

用データを増やす、(c)連接コスト[3]を修正する、といった方法である。このうち(c)については筆者の側で検討が不足しているうえ、コスト値の修正を繰り返してみる必要があり時間がかかるため、今回は検討を見送った。残る(a)と(b)であるが、これは並行して行うことのできる作業である。具体的には、村田寛（2010）で行った作業を説明する必要があるだろう。

村田寛（2010）では、『釈譜詳節』（1447年刊）巻6の電子データを用意し、それを手作業で形態素片[4]に分析して解析することで、MeCab用の辞書を作成した。さらに同じデータを学習用データとして利用し、解析用の辞書を構築した。そのため、文献データを解析して辞書の登録項目を増やせば、同時に学習用データも増やすことができるのである。本研究でも基本的に同じ方法で作業を進める。辞書の登録項目はそのままで学習用データを増やしたり、学習用データは増やさないで辞書の項目だけを増加させる、という実験も可能であるが、本研究では、まず辞書の分量と解析率との関係を探ることにした。

## 3．解析用辞書の構築

さて、村田寛（2010）でも既に説明しているが、ここでMeCabの辞書構築について簡単に概要を紹介しておく[5]。用意するものは、次の通りである：

(1) Seed辞書（品詞ごとにファイルを作成する）
(2) 設定ファイル（文字カテゴリの定義、未知語処理の動作定義など）
(3) 学習用コーパス

### 3．1．Seed辞書の準備

最終的な辞書を構築するための元になるのがSeed辞書である。コンマ区切りのテキストで記述し、品詞ごとにファイルを作成する。MeCabに付属のIPA日本語辞書から、一部を抜粋すれば以下の通り：

(4) a. 進学校,0,0,0,名詞,一般,*,*,*,*,進学校,シンガクコウ,シンガクコー
明日,0,0,0,名詞,副詞可能,*,*,*,*,明日,アシタ,アシタ
b. 受ける,0,0,0,動詞,自立,*,*,一段,基本形,受ける,ウケル,ウケル
受け,0,0,0,動詞,自立,*,*,一段,未然形,受ける,ウケ,ウケ
(5) 表層形,左連接状態番号,右連接状態番号,コスト,素性1,素性2,素性3,…

（4a）は名詞の記述例、（4b）は動詞の記述例である。(5)はそれぞれの素性を説明したものであるが、「コスト」までは必須の素性で、「素性 1」以降はその数も割り当ても任意である[6]。MeCab では学習用コーパスを用いて学習させ、左連接状態、右連接状態、コストなどを決定する[7]。素性の部分を柔軟に定義できるのが MeCab の特徴であり、この部分に 15 世紀朝鮮語に関する情報を付与することができる。

また、用言の活用については（4b）のように基本形、未然形、さらにはその他の活用形も項目として登録する必要がある。MeCab では特定の言語、文法に依存せず、文法情報に基づいた活用の処理を行わないためである。なお、本稿では 15 世紀朝鮮語の用言活用について、「語基[8]」の概念を用いて記述することとする。

表層形を除き、素性は村田寛（2010:18-19）に従い、次のように記述することとした。品詞分類は暫定的なものである：

(6) 分類 1, 分類 2, 分類 3,*, ムード , 形式 ,*

分類 1：Verb（動詞と形容詞）、Sonzaisi（存在詞）、Siteisi（指定詞）、Noun（名詞）、Adverb（副詞）、Postposition（後置詞）、Ending（語尾）、Prefix（接頭辞）、Suffix（接尾辞）、Symbol（補助記号）

分類 2 ～ 3：分類 1 の下位分類

ムード：平叙、疑問、命令など

形式：用言は第何語基か（第Ⅰ語基、第Ⅱ語基、…）、語尾は統辞的位置（終止形、接続形、…）

*：予備（将来の拡張用）

村田寛（2010）では『釈譜詳節』巻 6 を手作業で解析し、重複するレコードを取り除いた上で Seed 辞書用のファイルを作成した。また前述の通り、学習用コーパスとしても『釈譜詳節』巻 6 の本文を利用した。『釈譜詳節』を手作業で解析した例は次の通り：

(7)『釈譜詳節』の解析例（[　　] 内は筆者による注記）

| | |
|---|---|
| 釋譜詳節［tab］ | Noun, 固有名詞 , 文献名 ,*,*,* ［改行］ |
| 第 | Prefix,*,*,*,*,* |
| 六 | Noun, 数詞 ,*,*,*,* |
| EOS | ［文末を示す］ |
| 世尊 | Noun, 普通名詞 , 一般 ,*,*,* |

```
'il             Ending, 体言語尾 , 主格 ,*,*,*
象頭山          Noun, 固有名詞 , 地名 ,*,*,*
'ail            Ending, 体言語尾 , 処格 ,*,*,*
ga1             Verb, 自立 ,*,*, 語基 2,*
sia1            Suffix, 尊敬 ,*,*, 語基 3,*
龍              Noun, 普通名詞 , 一般 ,*,*,*
goa1            Ending, 体言語尾 , 共同格 ,*,*,*
鬼神            Noun, 普通名詞 , 一般 ,*,*,*
goa1            Ending, 体言語尾 , 共同格 ,*,*,*
'ui2h@1'ia1     Postposition,*,*,*,*,*
說法 h@1        Verb, 自立 ,*,*, 語基 1,*
de1             Suffix, 回想 ,*,*, 語基 2,*
si0             Suffix, 尊敬 ,*,*, 語基 1,*
da1             Ending, 用言語尾 ,*,*, 平叙 - 下称 , 終止形 ,*
.               Symbol, 句点 ,*,*,*,*
EOS
…
```

上の例は巻 6 の冒頭、「釋譜詳節第六 / 世尊이 象頭山애 가샤 龍과 鬼神과 위ᄒᆞ야 說法ᄒᆞ더시다.」（釈譜詳節第六／世尊が象頭山へお行きになって龍と鬼神とのために説法なさった）という部分をローマ字転写し[9]、解析したものである。用言については、左端の表層形が同じであっても、素性中の語基に関する情報で区別される。

### 3．2．設定ファイルの準備：未知語の処理について

未知語の処理であるとか、学習の際の連接 ID 付与であるといったさまざまな動作の設定を、設定ファイルで自由に変更することができる。ここでは未知語の処理についての修正点を述べておく。

MeCab では、辞書に登録されていない項目を未知語として処理するが、どのような素性を付与するかという処理は、あらかじめ定義された字種に基づいて行われる。例えば、漢字の連続であれば「名詞」として処理し、素性列を付与する、といった具合である。字種の定義については後から自由に定義することができ、またそれらの字種をどのように処理するかを定義することができるため、新たな辞書を構築する際に非常に有益である。

村田寛（2010）では未知語の処理に関し、すべて素性列に「未知語,*,*,*,*,*,*」と記述するようにしている。この方式ではどの部分を解析できなかったかを明らかにすることができるが、正解のファイルと突き合わせた場合、解析率が落ちてしまうというのが難点である。そこで本稿では、字種の定義や、未知語の処理方式についても再検討を行った。

まず字種の定義であるが、漢字を一つにまとめて「HANJA」カテゴリとし、さらにアルファベットとアラビア数字を一つのグループにまとめて「HANGEUL」カテゴリとした[10]。アルファベットとアラビア数字については、いずれもデータ上でハングルと声調を表すのに用いられるため、区分せず一つにまとめたものである。

次にグループごとの処理であるが、「HANJA」カテゴリについては未知語の場合、ひとまず名詞として処理し、一般名詞と同じく「Noun,普通名詞,一般,*,*,*」の素性を与えることとした。15 世紀の朝鮮語文献に現れる漢字語の大部分は名詞であり、漢字語のみで用言となることはない。一部の漢字語は助数詞であったり、固有名詞であったりということもあるが、大部分は一般名詞とみなして問題はないと思われる。

なお、「HANGEUL」カテゴリについては名詞あるいは用言として処理することとし、「Noun,普通名詞,一般,*,*,*」、「Verb,自立,*,*,*,*」という素性を与えた。これは本研究での暫定的な措置である。用言が辞書に網羅されているならば、名詞としてのみ処理するよう設定するべきであろうが[11]、ここで構築する辞書は項目も少ないため、まずは名詞あるいは用言として判定するように設定した。こうしたハングルの未知語処理については、今後さらに検討していく必要があるだろう。

以上のような設定で辞書を構築し、形態素解析を行ったところ、漢字語についてはおおむね望ましい結果となった。例えば『釈譜詳節』巻 6、巻 9 のデータで構築した辞書で巻 23 を解析した場合、「信心」「貧窮海」など辞書に存在しない漢字語についても、一般名詞として結果的に「正しい」解析結果を出力した。ハングルについては思うような結果を得られず、そもそも形態素片の抽出、解析自体がうまくできていない。

なお、字種の定義、未知語処理の定義のほかは、IPA 日本語辞書の設定ファイルを元にいくつかの変更を行う程度にとどめた。その他の設定を修正することでどのように解析率が変化するかについては、今後の課題としたい。

### 3．3．学習用コーパスを用いた学習と辞書の構築

以上で説明した辞書ファイル、設定ファイル、さらに学習用コーパスを用いて、辞書を構築する。基本的には MeCab ホームページなどの手順に従うが、学習用バイナリ辞書、配布用バイナリ辞書の作成に際して、辞書の文字コードを UTF-8 に指定する必要があ

る[12]。村田寛（2010）では漢字語の解析がうまくいかなかったのであるが、それはこの文字コードの設定に起因するものと思われる。

このようにして構築した辞書を利用して解析を行うが、学習した内容に基づいて解析処理を行うため、学習用データに現れなかった語節でも解析を行ってくれる。例えば動詞「가-」（行く）に〈現在〉を表す時称接尾辞「-ᄂᆞ-」、さらに接続形語尾の「-니」（…だが）がついた「·가ᄂᆞ·니」[13]（行くが）という語節は『釈譜詳節』巻6に現れるが、接尾辞「-ᄂᆞ-」の代わりに〈尊敬〉の接尾辞「-시-」がついた「·가시·니」（いらっしゃったが、お行きになったが）という語節は巻6には見られない。接尾辞「-ᄂᆞ-」は用言の第Ⅰ語基につくが、接尾辞「-시-」は用言の第Ⅱ語基に接続する。いずれの場合も用言「가-」の声調は去声となり、用言の声調では活用語基の違いを判断できない。さて、学習用データにない「·가시·니」は解析できるであろうか。

結論から言えば、こうした活用語基の接続に関する解析処理はかなり正確に行ってくれる。上記の例を、『釈譜詳節』巻6のデータで構築した辞書で解析すると、次のような結果となる：

| 「·가ᄂᆞ·니」を解析 | | 「·가시·니」を解析 | |
|---|---|---|---|
| ga1 | Verb, 自立 ,*,*,*, 語基1 ,* | ga1 | Verb, 自立 ,*,*,*, 語基2 ,* |
| n@0 | Suffix, 現在 ,*,*,*, 語基2,* | si0 | Suffix, 尊敬 ,*,*,*, 語基2,* |
| ni1 | Ending, 用言語尾 ,*,*,*, 接続形 ,* | ni1 | Ending, 用言語尾 ,*,*,*, 接続形 ,* |

表層形は「ga1」のように同じであっても、それぞれ活用語基を正しく判断していることが分かる。

なお、接尾辞自体も活用語基を持つと考え、素性の記述に反映させているが、いずれの例でも接続形語尾「-니」に前接する接尾辞は第Ⅱ語基となっており、正確に解析されていることが分かる。

以下では実際に構築した辞書で形態素解析を行い、その解析率について検討してみる。

## 4．解析率について

### 4．1．辞書の構築と解析の際に用いたデータ量

MeCabは辞書と学習用コーパスから学習した連接の強度などから形態素の解析を行う。ここでは辞書に登録された項目の量と、学習用コーパスのデータ量によって形態素解析の結果がどのように変わるか、検討してみたい。辞書構築のデータとして、『釈譜詳節』の巻6、巻9の全部と、同じく『釈譜詳節』巻23の始め10張分、さらに種類の異なる資料

として『法華経諺解』（以下『法華経』とする）巻 3 の始め 10 張分のデータを利用した。各文献の「正解」ファイルを作成し、そこから重複を取り除いて Seed 辞書用のデータにする。また「正解」ファイル自体は学習用コーパスのデータとして用いる。ここで「正解」ファイルから抽出した Seed 辞書の項目数（形態素片の異なりの数）と、「正解」ファイルの形態素片の総数[14]（＝学習用コーパスのデータ量。形態素片の延べ数）を示せば以下の通りである：

(8) 各文献のデータ量

| | 『釈譜詳節』 | | | 『法華経』 |
|---|---|---|---|---|
| | 巻 6 | 巻 9 | 巻 23 | 巻 3 |
| Seed 辞書の項目数 | 1746 | 1489 | 531 | 420 |
| 形態素片の総数 | 7813 | 6304 | 1611 | 1149 |

なお、解析率の計算には MeCab の配布パッケージに含まれる評価用スクリプト mecab-system-eval を利用した[15]。

## 4．2．解析の結果

ここで比較に用いたのは、『釈譜詳節』の巻 6、巻 9、巻 23（一部）、『法華経』巻 3（一部）のデータを使って構築した辞書である。辞書構築に際して用いたデータ、さらにその辞書を用いて解析した各文献の解析率をあわせて提示すれば、以下の表(9)の通りである。複数の文献から構築した辞書については重複する形態素も含まれるため、Seed 辞書の項目数も合わせて提示した：

(9) a. 『釈譜詳節』のみで構築した辞書による解析率と Seed 辞書の項目数

| 辞書データ | 解析データ | | | | Seed 辞書の項目数 |
|---|---|---|---|---|---|
| 『釈譜詳節』 | 『釈譜詳節』 | | | 『法華経』 | |
| | 巻 6 | 巻 9 | 巻 23 | 巻 3 | |
| 巻 6 | 97.3984 | 65.5565 | 72.9338 | 66.2298 | 1746 |
| 巻 9 | 62.4811 | 96.4859 | 76.4817 | 65.6042 | 1489 |
| 巻 6, 9 | 96.7483 | 95.4431 | 82.6336 | 71.4852 | 2617 |
| 巻 23（一部） | 47.6575 | 52.6818 | 98.6718 | 51.7658 | 531 |
| 巻 6, 9, 23（一部） | 96.6366 | 95.0758 | 96.4216 | 71.3781 | 2768 |

(9) b.『釈譜詳節』『法華経』で構築した辞書による解析率と Seed 辞書の項目数

| 辞書データ | | 解析データ | | | | Seed 辞書の項目数 |
|---|---|---|---|---|---|---|
| 『釈譜詳節』 | 『法華経』 | 『釈譜詳節』 | | | 『法華経』 | |
| | | 巻 6 | 巻 9 | 巻 23 | 巻 3 | |
| − | 巻 3（一部） | 36.0871 | 43.0325 | 45.6242 | 98.1216 | 420 |
| 巻 6 | 巻 3（一部） | 97.3528 | 66.5049 | 75.1496 | 96.0250 | 1949 |
| 巻 9 | 巻 3（一部） | 64.0344 | 96.0346 | 76.3625 | 97.0076 | 1693 |
| 巻 6, 9 | 巻 3（一部） | 96.7220 | 95.2863 | 82.1735 | 95.5317 | 2787 |
| 巻 6, 9, 23（一部） | 巻 3（一部） | 96.4187 | 94.9421 | 96.0676 | 94.6429 | 2933 |

上の表から分かるように、『釈譜詳節』巻 6、巻 9 ともに、それぞれのデータを用いて構築した辞書で解析した場合、90% 以上の解析率となっているが、辞書を入れ替えた場合、60% 台の解析率にとどまっている。また、辞書構築のデータとして利用していない『釈譜詳節』巻 23 については、『釈譜詳節』巻 6 のみ、『釈譜詳節』巻 9 のみのデータで構築した辞書ではそれぞれ 70% 台の解析率だが、両方を合わせた辞書では 80% を超える解析率となった。この傾向は『法華経』のデータを追加した場合も同様で、結果的に Seed 辞書の項目数が多いほど、解析率が高くなるといえよう。

次に『法華経』の解析結果を見ると、辞書構築に『法華経』のデータが使われている場合、いずれの辞書においても 90% 以上の解析率となっている。しかし、『法華経』のデータが使われていない辞書での解析結果は、70% 前後にとどまっている。

以上の解析結果を単純に考えるならば、辞書に解析する文献に現れる形態素片が登録されていれば、その解析結果は精度の高いものとなることが分かる。つまり、辞書に登録する形態素片の数を増やせば、それだけ解析精度が高まるということである。

## 5．解析ミスの傾向について

今回の解析結果を見てみると、解析のミスは多々あるものの、そのミスに一定の傾向が見られることが分かった。ここでは解析ミスの傾向について、一点だけまとめておく。

解析ミスのなかで最も目についたのが、用言の第Ⅰ語基に接続形語尾「－ユ」が続く場合に、この用言を第Ⅱ語基と解析する例である。『釈譜詳節』巻 23 を、巻 6 と巻 9 のデータで構築した辞書で解析した例について見てみよう：

⑽　解析ミスの例・その 1

| 解析ファイル | | 正解 | |
|---|---|---|---|
| 修多羅 | Noun, 普通名詞 , 一般 ,*,*,* | 修多羅 | Noun, 普通名詞 , 一般 ,*,*,* |
| n@n1 | Ending, 体言語尾 , 題目 ,*,*,* | n@n1 | Ending, 体言語尾 , 題目 ,*,*,* |
| 經 | Noun, 普通名詞 , 一般 ,*,*,* | 經 | Noun, 普通名詞 , 一般 ,*,*,* |
| 'i1 | Siteisi, 非自立 ,*,*, 語基 2 ,* | 'i1 | Siteisi, 非自立 ,*,*, 語基 1 ,* |
| 'o1 | Ending, 用言語尾 ,*,*, 接続形 ,* | 'o1 | Ending, 用言語尾 ,*,*, 接続形 ,* |
| 毘那耶 | Noun, 普通名詞 , 一般 ,*,*,* | 毘那耶 | Noun, 普通名詞 , 一般 ,*,*,* |
| n@n1 | Ending, 体言語尾 , 題目 ,*,*,* | n@n1 | Ending, 体言語尾 , 題目 ,*,*,* |
| 律 | Noun, 普通名詞 , 一般 ,*,*,* | 律 | Noun, 普通名詞 , 一般 ,*,*,* |
| 'i1 | Siteisi, 非自立 ,*,*, 語基 2 ,* | 'i1 | Siteisi, 非自立 ,*,*, 語基 1 ,* |
| 'o1 | Ending, 用言語尾 ,*,*, 接続形 ,* | 'o1 | Ending, 用言語尾 ,*,*, 接続形 ,* |

上記は「脩多羅ᄂᆞᆫ 經이오 毗那耶ᄂᆞᆫ 律이오」（修多羅は経であり、毘那耶は律であり）を解析したものであるが、［　　］で囲んだ部分が誤りである。接続形語尾「- 고」（…して）は用言の第Ⅰ語基につく。ここでは用言の活用語基を第Ⅱ語基と解析している。このほか、用言と接続形語尾「- 고」の間に〈尊敬〉を表す接尾辞「- 시 -」が入る場合に、接尾辞「- 시 -」の活用語基をミスするという場合も見られた。

⑾　解析ミスの例・その 2

| 解析ファイル | | 正解 | |
|---|---|---|---|
| 世界 | Noun, 普通名詞 , 一般 ,*,*,* | 世界 | Noun, 普通名詞 , 一般 ,*,*,* |
| r@r1 | Ending, 体言語尾 , 対格 ,*,*,* | r@r1 | Ending, 体言語尾 , 対格 ,*,*,* |
| da2 | Adverb, 一般 ,*,*,*,* | da2 | Adverb, 一般 ,*,*,*,* |
| bi0cui1 | Verb, 自立 ,*,*, 語基 2,* | bi0cui1 | Verb, 自立 ,*,*, 語基 2,* |
| si0 | Suffix, 尊敬 ,*,*, 語基 2 ,* | si0 | Suffix, 尊敬 ,*,*, 語基 1 ,* |
| go1 | Ending, 用言語尾 ,*,*, 接続形 ,* | go1 | Ending, 用言語尾 ,*,*, 接続形 ,* |

上の例も⑽と同じく、巻 23 を解析した結果である。「世界ᄅᆞᆯ 다 비취시고」（世界をみなお照らしになり）という部分であるが、やはり接続形語尾「- 고」に前接する接尾辞「- 시 -」を第Ⅱ語基と判定している。

反対に、用言の第Ⅱ語基に接続する語尾の直前で、用言の活用語基を第Ⅰ語基と解析する例も見られた：

⑿　解析ミスの例・その3

| 解析ファイル | | 正解 | |
|---|---|---|---|
| gi0py | Verb, 自立 ,*,*,*, 語基 2,* | gi0py | Verb, 自立 ,*,*,*, 語基 2,* |
| n1 | Ending, 用言語尾 ,*,*,*, 連体形 ,* | n1 | Ending, 用言語尾 ,*,*,*, 連体形 ,* |
| m@0z@0m | Noun, 普通名詞 , 一般 ,*,*,* | m@0z@0m | Noun, 普通名詞 , 一般 ,*,*,* |
| @1ro1 | Ending, 体言語尾 , 具格 ,*,*,* | @1ro1 | Ending, 体言語尾 , 具格 ,*,*,* |
| 供養 h@0 | Verb, 自立 ,*,*,*, 語基 1 ,* | 供養 h@0 | Verb, 自立 ,*,*,*, 語基 2 ,* |
| mien1 | Ending, 用言語尾 ,*,*,*, 接続形 ,* | mien1 | Ending, 用言語尾 ,*,*,*, 接続形 ,* |

「기픈 ᄆᆞᅀᆞᄆᆞ로 供養ᄒᆞ면」（深い心で供養すれば）という部分で、接続形語尾「－면」（…すれば）は用言の第Ⅱ語基に接続するが、ここで「供養ᄒᆞ－」が第Ⅰ語基と解析されている。

これらの誤りは、いずれも第Ⅰ語基と第Ⅱ語基とが同じ形式、声調になる場合に起こっていることが分かるが、こうした場合の解析がなぜできないか、その理由を正確に知ることはできない。解決の方法として現在検討しているのが、(a)辞書構築後に、連接コストを修正する、(b)それぞれの接続形語尾について、接続する活用形のタイプを素性として記述する、といった点である。(a)については、「第Ⅰ語基―接続形」、「第Ⅱ語基―接続形」の連接コストを変えてやることで、どちらの結びつきが現れやすいか、また結びつきやすいかという情報を修正することができると考えられる。ただし、接続形語尾の中には用言の第Ⅱ語基に接続するものもあるため、一律に修正すると他の問題が生じる可能性がある。そのような点では(b)がより現実的な方策といえよう。今後実際に作業を行ってみて、どのような結果になるか検証したい。

## 6．おわりに

本研究では、形態素解析エンジンである MeCab を用いて 15 世紀の朝鮮語を解析し、その解析率の考察を行ったが、その結果は次のようにまとめることができる。

解析の際に、MeCab 用辞書に登録されている形態素片の数が多いほど解析精度が高くなる。そして、解析ミスにもある程度のパターンがあることが分かった。また、未知語の処理について、漢字の連続を一般名詞と判定することで、辞書にない語でも結果的に「正しい」解析ができるようになった。

本研究で試みた MeCab による形態素解析の結果は、他のツールで活用できるというメリットがある。例えば ChaKi（茶器）という自然言語コーパスの構築、検索、および言語要素へのタグ付けをサポートするツール群があるが、MeCab で形態素解析した結果を

ChaKi で取り込み、KWIC 検索やコロケーション情報などを取り出すことが可能である。

こうした研究の用途に供するためには、さらに解析率を上げて手作業による修正の手間を減らし、できるだけ大量のデータを作成していくことが必要であろう。そのための課題として、辞書、学習用データを拡充したり、解析ミスのパターンを分析してそのミスを減らす方法を検討するなどといった作業が今後必要である。

## 注

1　MeCab とは、言語、辞書、コーパスに依存しない汎用的な設計を基本方針とするオープンソース形態素解析エンジンで、京都大学情報学研究科－日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクトにより開発されたものである。MeCab の詳細については、ホームページ http://mecab.sourceforge.net/ を参照。

2　「声調」といっても中国語におけるそれとは異なり、高低のアクセントを示すものであったといわれる。15 世紀朝鮮語文献においてはハングルの横に「傍点」と呼ばれる点を打つことで、これらの「声調」を表示した。点が一つであれば「去声」（高調）、二つであれば「上声」（上昇調）、点がなければ「平声」（低調）を表す。

3　「連接コスト」とは、MeCab が解析の際に用いるパラメータの一つである。二つの要素のつながりやすさを判定する数値といえる。例えば日本語において「姓」→「名」というつながりは、「名」→「姓」というつながりよりも起こりやすいといえる。MeCab とともに配布されている日本語の IPA 辞書を見ると、「姓」→「名」の連接コスト値は「-7009」、「名」→「姓」の連接コスト値は「152」に設定されており、「姓」→「名」のコストが低くなるようになっている。コスト計算については「日本テレビ東京で学ぶ MeCab のコスト計算」（http://www.mwsoft.jp/programming/munou/mecab_nitteretou.html）を参照のこと。

4　形態素片とは、形態素と認めうる可能性のある最小単位の文字列を言う。あえて形態素片と呼ぶのは、言語学的に厳密な意味での形態素と言えないものもあるためである。本研究で使う形態素片という用語は、山下達雄・松本裕治（1998）で使われている形態素片とは若干異なる。山下達雄・松本裕治（1998: 19）では、形態素片を次のように定義している。

「形態素片とは形態素として認識される可能性のある最小単位の文字列である。わかち書きされていない言語では、その言語体系での文字一文字であり、わかち書きされている言語では、ブランク等で区切られた文字列である。」

15 世紀の朝鮮語は分かち書きされていないため、山下達雄・松本裕治（1998）の形態素片の定義に従えば、文字一文字を形態素片にしないといけないが、本研究では朝鮮語の一文字をローマ字転写し、一文字より小さい単位を形態素片として扱うので、本研究で言う形態素片は、山下達雄・松本裕治（1998）でのそれとは若干異なる。

5 辞書構築についての詳細は MeCab ホームページ「オリジナル辞書 / コーパスからのパラメータ推定」（http://mecab.sourceforge.net/learn.html）を参照。

6 IPA 辞書では素性として「品詞、品詞細分類 1、品詞細分類 2、品詞細分類 3、活用型、活用形、基本形、読み、発音」を定義している。素性中「*」はその素性を使用していないことを表す。

7 そのため、Seed 辞書の段階では 0 になっている。

8 「語基」については菅野裕臣（1997）などを参照のこと。第Ⅳ語基までを認める。

9 ローマ字転写の方法は福井玲（1989）にならった。ローマ字転写の一覧は本稿末の資料 1 を、『釈譜詳節』巻六冒頭部分の転写の実際については資料 2 を参照のこと。

10 実際には文字カテゴリの定義ファイル char.def において、カテゴリ名とそのカテゴリに含まれる文字を、文字コードを指定して定義するという作業である。例えば Unicode の 0x0061（小文字の a）から 0x007A（小文字の z）の範囲に含まれる文字を、「HANGEUL」カテゴリとして定義するには、char.def へ以下のように記述する：

0x0061..0x007A HANGEUL

11 例えば日本語 IPA 辞書などは、ひらがなやカタカナが連続する未知語は名詞あるいは感動詞として処理するように設定されている。

12 具体的には mecab-dict-index を用いてバイナリ辞書を作成する際、-c オプションと -f オプションとで「utf8」を指定すればよい。

13 ハングルの横につけた「·」は声調を表す傍点（注 2 を参照）を示すものである。ここでは「·가」が去声、「ᄂᆞ」は平声、「·ᄂᆡ」は去声であることを表す。

14 ここでの「形態素片の総数」は、文の終を示す「EOS」も含めた、ファイルの行数を提示している。

15 mecab-system-eval は解析結果の出力ファイルと正解ファイルを引数にとり、以下のような結果を出力する：

| | precision | recall | F |
|---|---|---|---|
| LEVEL 0: | 85.0784（1357/1595） | 87.9456（1357/1543） | 86.4882 |
| LEVEL ALL: | 81.5674（1301/1595） | 84.3163（1301/1543） | 82.9191 |

ここで「LEVEL」は比較に用いた素性のレベルを表し、「LEVEL 0」は最初の素性（＝表層形）を指す。「LEVEL ALL」行は全ての素性を比較した値である。（ ）内

の数値は分母が形態素片全体の数、分子は出力と正解とで一致する形態素片の数で、「precision」の列は正解ファイルについて、「recall」の列は出力ファイルについてのデータを示している。「F」列は左 2 列のパーセンテージの平均である。上の例でいえば、正解ファイルの形態素片は 1595 個、解析結果の形態素片は 1543 個で、表層形（LEVEL 0）が一致するのは 1357 個、表層形も含めた全ての素性が一致するのは 1301 個、ということになる。

本研究では解析率として、全素性を比較した場合の平均値、即ち上の表で右下に示される数値を用いることとする。

## 影印本など

"活字本 法華経諺解（巻之三）", 弘文閣（1997）
"석보상절 제 23・24 연구", 金英培（2009）, 동국대학교출판부
"역주 석보상절 제 6・9・11", 세종대왕기념사업회（1991）

## 〈参考文献〉

菅野裕臣（1997）「朝鮮語の語基について」『日本語の外国語との対照研究Ⅳ『日本語と朝鮮語』下巻 研究論文編』、くろしお出版

岸本貴之・高橋治久・堀田一弘（2009）「CRF による係り受け解析の結果を反映させた日本語形態素解析」『情報処理学会研究報告［自然言語処理］』、2009-2、情報処理学会

工藤拓・山本薫・松本裕治（2004）「Conditional Random Fields を用いた日本語形態素解析」『情報処理学会研究報告［自然言語処理］』2004-NL-161、情報処理学会

須賀井義教（2009）「中期朝鮮語文献の電子データ構築に関するいくつかの問題— XML の利用を中心に—」『近畿大学語学教育部ジャーナル』第 5 号、近畿大学語学教育部

平野善隆（1997）『用言の活用を考慮した韓国語品詞体系の提案とそれを用いた韓国語形態素分析』、奈良先端科学技術大学院大学情報科学研究科情報処理学専攻修士論文（NAIST-IS-MT9551092）

福井玲（1989）「中期朝鮮語文献の電子計算機による処理」『明海大学外国語学部論集』2、明海大学

村田寛（2010）「15 世紀朝鮮語の形態素解析の試み— MeCab を利用して—」『福岡大学研究部論集 A：人文科学編』Vol. 10 No. 3、福岡大学

守岡知彦（2008）「MeCab を用いた古典中国語の形態素解析の試み」『情報処理学会研究報告［人文科学とコンピュータ］』2008-73、情報処理学会

山下達雄・松本裕治（1998）「言語に依存しない形態素解析ツールキットの開発」『情報処理学会研究報告』98-99、情報処理学会

Lafferty, John, Andrew McCallum and Fernando Pereira（2001）*Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* In Proc. of ICML

## 関連 URL

ChaKi プロジェクト日本語トップページ
http://sourceforge.jp/projects/chaki/

MeCab: Yet Another Part-of-Speech and Morphological Analyzer
http://mecab.sourceforge.net/

日本テレビ東京で学ぶ MeCab のコスト計算
http://www.mwsoft.jp/programming/munou/mecab_nitteretou.html

本研究は、2009-2011 年度科学研究費補助金基盤研究(B)、研究課題番号 21320075「朝鮮語史の国際的共同研究のための研究資源基盤構築」の研究成果の一部でもある。

資料 1　15 世紀朝鮮語のローマ字転写表

＜子音＞

| ハングル | ㄱ | ㄴ | ㄷ | ㄹ | ㅁ | ㅱ | ㅂ | ㅸ | ㅅ | |
|---|---|---|---|---|---|---|---|---|---|---|
| 転写 | g | n | d | r | m | w | b | v | s | |
| ハングル | ㅿ | ㅇ | ㆁ | ㆆ | ㅈ | ㅊ | ㅋ | ㅌ | ㅍ | ㅎ |
| 転写 | z | ' | q | x | j | c | k | t | p | h |

＜母音＞

| ハングル | ㅏ | ㅓ | ㅗ | ㅜ | ㅡ | ㅣ | ㆍ |
|---|---|---|---|---|---|---|---|
| 転写 | a | e | o | u | y | i | @ |

＜重母音＞

| ハングル | ㅑ | ㅐ | ㅘ | ㅝ | ㅖ |
|---|---|---|---|---|---|
| 転写 | ia | ai | oa | ue | iuiei |

＜複子音＞

| ハングル | ㅺ | ㅼ | ㅽ | ㆅ | ㅄㅈ | ㅄㅅ |
|---|---|---|---|---|---|---|
| 転写 | sg | sb | bd | hh | bsg | bsd |

＜アクセント＞

| 平声 | 去声 | 上声 |
|---|---|---|
| 0 | 1 | 2 |

資料 2 『釈譜詳節』巻 6 の冒頭部分

（註や会話文などの記号を付加している。｢　｣ の中はローマ字転写する前のハングル表記。）

釋譜詳節 第六

世尊 'i1 象頭山 'ai1 ga1sia1 龍 goa1 鬼神 goa1 'ui2h@1'ia1 說法 h@1de1si0da1.

[龍鬼 'ui2h@1'ia1 說法 h@1sia0mi1 bu0ties0 na1hi1 sier0hyn1

dur2hi0re1si0ni1 穆王 'ie0sys1 cas0 h@i1 乙酉 i0ra1.]

ㅇ bu0tiei2 目連 'i1 d@0rie1 ni0r@0sia1d@i1 「nei2 迦毗羅國 'ei1

ga1'a1 'a0ba1nims2gyi1'oa1 'a0j@1ma0nims2gyi1'oa1

['a0j@1ma0ni2m@n1 大愛道 r@r1 ni0ry0si1ni1 大愛道 i0 摩耶夫人 s0 兄 ni2mi1si0ni1 'iaq0j@i1 摩耶夫人 man1 mod2 h@1sir0ss@i1 be0gyn1 夫人 'i1 d@0'oi0si1ni0ra1.]

'a0ja0ba1nim2nai1sgyi1 da2 安否 h@1z@b0go1 sdo1 耶輸陁羅 r@r1 dar0'ai1'ia1 恩愛 r@r1 gy0cie1 羅睺羅 r@r1 no0ha1 bo0nai1'ia1 siaq2jai1 d@0'oi0'ei1 h@0ra1. 羅睺羅 i0 得道 h@1'ia1 do0ra1 ga1za1 'e1mi0r@r1 濟渡 h@1'ia1 涅槃 得 ho1m@r1 na0 g@d1gei1 h@0ri1ra1.」

｢釋譜詳節 第六

世尊·이 象頭山·애 ·가·샤 龍·과 鬼神·과 : 위·ᄒᆞ·야 說法·ᄒᆞ·더시·다 .

[龍鬼 : 위·ᄒᆞ·야 說法·ᄒᆞ샤·미 부텻 ·나·히 셜·흔 : 둘히·러시·니 穆王 여·슷 찻 ·히 乙酉ㅣ·라 .]

○부 : 톄 目連·이 ᄃᆞ·려 니ᄅᆞ·샤·ᄃᆡ「: 네 迦毗羅國·에 ·가·아 아·바 : 닚·긔·와 아·ᄌᆞ마 : 닚·긔·와

[아·ᄌᆞ마 : 니·믄 大愛道·ᄅᆞᆯ 니르·시·니 大愛道ㅣ 摩耶夫人ㅅ 兄 : 니·미시·니 양·ᄌᆡ 摩耶夫人·만 : 몯 ·ᄒᆞ실·씨 버·근 夫人·이 ᄃᆞ외·시니·라 .]

아자·바 : 님·내·씌 : 다 安否·ᄒᆞᅀᆞᆸ·고 ·또 耶輸陁羅·ᄅᆞᆯ 달·애·야 恩愛·ᄅᆞᆯ 그·쳐 羅睺羅·ᄅᆞᆯ 노·하 보·내·야 : 샹·재 ᄃᆞ외·에 ᄒᆞ·라 . 羅睺羅ㅣ 得道·ᄒᆞ·야 도·라 ·가·ᅀᅡ ·어미·ᄅᆞᆯ 濟渡·ᄒᆞ·야 涅槃 得·호·ᄆᆞᆯ 나 ·ᄀᆞᆮ·게 ᄒᆞ·리·라 .」