日時：平成 24 年 3 月 2 日（金曜日）15 時 00 分より 17 時 30 分まで
場所：東京外国語大学 AA 研 304 室（マルチメディア会議室）
報告者名（所属），報告タイトル：

1) 小山内優子（東京外国語大学），「『楞厳経諺解』中の朝鮮語の連体形及び名詞形について」（仮題）
2) 村田寛（ＡＡ研共同研究員、九州大学），須賀井義教（ＡＡ研共同研究員、近畿大学），「15 世紀朝鮮語の形態素解析とその周辺」

報告内容：別途添付

『楞厳経諺解』巻四中の名詞形と「連体形＋依存名詞」について

小山内優子
（東京外国語大学大学院 博士後期課程）

要旨

本研究の目的は、中期朝鮮語に於いて、名詞形と「連体形＋依存名詞」との間に如何なる棲み分けが存在するかを明らかにすることである。調査資料としては『楞厳経諺解』の巻四を用いた。この資料中に名詞形と「連体形＋依存名詞」の両方で現れる用言について、両者の違いを考察した。

動詞nirAda（言う）の場合、処格名詞形nir‘omaiで現れる例は、いずれも漢文原文の「云」の朝鮮語訳であった。これは、「所謂」がnirAdaの連体形nir‘onで固定して現れるのと同様に、副詞的に用いられていると考えられる。

形容詞mArgda（きれいだ）は、「連体形＋依存名詞」で現れる例が全てmArgAn dAi byteで現れ、いずれも「黏湛」の朝鮮語訳であった。

動詞‘arda（知る）は、次の一例を除き、全て名詞形‘aromで現れている。

（1） 其分劑頭數i 又非阿難‘Ai 所知者ini
gy 分劑頭數i sdo 阿難‘Ai ‘arorx gesi ‘anini（楞嚴4:104a_3）

王力（1980[2008]: 344-345）、伊藤丈（1995: 148）によると、結構助詞「所」は常に他動詞の前に置かれ、述語形式や文に、連体修飾語的性質を持たせる働きをする。連体修飾語の後ろは名詞が来たり、「者」が用いられたりする。この「者」は名詞が人や事物を指す時は用いられないことがある。このとき、連体修飾語は名詞性を持つ。上の（1）は、ちょうど結構助詞「所」が「者」と共に用いられている例で、依存名詞gesが「者」に対応していると考えられる。

得られた用例の中には、例（2）のように、連体形用言の動作の対象や、動作の行われる場所を依存名詞が担っているものがあった。このような例は、依存名詞が具体的な事物や場所を示しているという点で、名詞形との意味の違いが明らかである。

（2） a. 宛轉虛妄hA‘ia 無可憑據hAnira
duruhirhue 虛妄hA‘ia ‘eru byturx dAi ‘ebsynira（楞嚴4:043b_4）
b. 氣附rAr 曰合‘ini 合濕hA‘ia 而生也ira

氣 bytumyr nir‘odAi ‘e‘urumini 濕‘yr ‘e‘ure narssira（楞嚴4:028b_5）

一方、名詞形と「連体形＋依存名詞」との間に大きな差異が見られない例もあった（例3）。

（3） a. 當知是明‘i 非日‘inmie 非空‘imie 不異空日‘iroda

‘i bArgomi hAi ‘animie 空 ‘animie 空goa hAi‘oa darAdi ‘anihAn dAr bandAgi ‘arriroda（楞嚴4:041a_5）

b. 若悟本**真**hAmien 則知諸妄‘i 了不相關hArini

hAdaga 本來s **真**‘Ar ‘armien han 妄‘i god sery byddi ‘anihomAr ‘arrini（楞嚴4:068b_3-069a_1）

例（3a）は、依存名詞dAが具体的な事物や場所を指しているとは考えにくく（cf. 例2a）、意味上では名詞形と接近していると考えられる。本発表では、このような名詞形と「連体形＋依存名詞」との間の違いが小さいと考えられる例について、十分な考察が出来なかったため、今後の課題としたい。

参考文献

伊藤丈（1995）『仏教漢文入門』東京: 大蔵出版
王力（1980 [2008]）『漢語史稿』北京: 中華書局

中期朝鮮語の形態素解析とその周辺

村田寛（AA 研共同研究員）・須賀井義教（近畿大学）

1. はじめに

本稿は，15 世紀朝鮮語の文献の自動形態素解析についてその方法を検討するとともに，その解析結果をどのように活用することができるのか，具体的な方法を提示することを目的とする．形態素解析エンジンとしては，オープンソースの形態素解析エンジンである MeCab（めかぶ）を用いる．

形態素解析の技術は言語研究，言語教育だけでなく，機械翻訳や構文解析などの前処理において重要な役割を果たす．特に現代語の形態素解析については，形態素解析の技術を応用してインターネットのウェブページ上からキーワードを抽出したり，検索や入力の補助に利用したりすることも可能である．インターネットやコンピュータ技術の進展とともに，自然言語を処理するための技術が果たす役割も，さらに重要度を増していると考えられる．

現代朝鮮語の形態素解析についてもさまざまな成果が見られ [1)]，中でも国立国語院などが進めてきた 21 世紀世宗計画の成果物として，1000 万語節規模の形態素解析済みコーパスが公開されている．この成果物にはコーパスだけでなく自動形態素解析器も含まれており，利用者が自由に解析を行い，結果を得ることができる．形態素解析済みのコーパスは朝鮮語研究においても多く利用されているようであり，研究の上で有益な資料といえる．また，形態素解析器を使って，利用者が目的に合わせ，新たなテキストをコーパスに追加することができるという点も長所といえるだろう．

一方，15 世紀の朝鮮語文献を対象とした形態素解析は，解析対象の内容があらかじめ固定されているため，ユーザーのさまざまな入力をリアルタイムで解析して結果を得る必要はなく，一度解析を行えばそれで終わりである．ただし，手作業で解析を行うには膨大な時間が必要であり，解析の時間短縮という点で，自動形態素解析の方法を検討することは有用であると考える．また，中期朝鮮語について形態素解析を行った電子データはこれまでのところ見られず，従来にないさまざまな検索の可能な資料が提供できる．

本稿では，特定の言語に依存しない，汎用的な形態素解析エンジンである MeCab を利用して，ユーザーが辞書を自由に定義することができ，さらにインターネット上でも利用することのできる自動形態素解析の方法について提案したい．MeCab を用いることで，利用者のニーズに合わせた出力結果のカスタマイズ，新たな語句の登録，インターネットを通じたサービスの提供などが可能となり，さまざまな応用が可能になると考える．

---

1) 現代朝鮮語の形態素解析に関して朝鮮語学の側から概説したものに유혜원(2004)，황화상(2006)がある．いずれもコンピュータによる形態素解析，構文解析について説明しており，特に前者は構文解析，後者は形態素解析についてより詳しい解説がなされている．

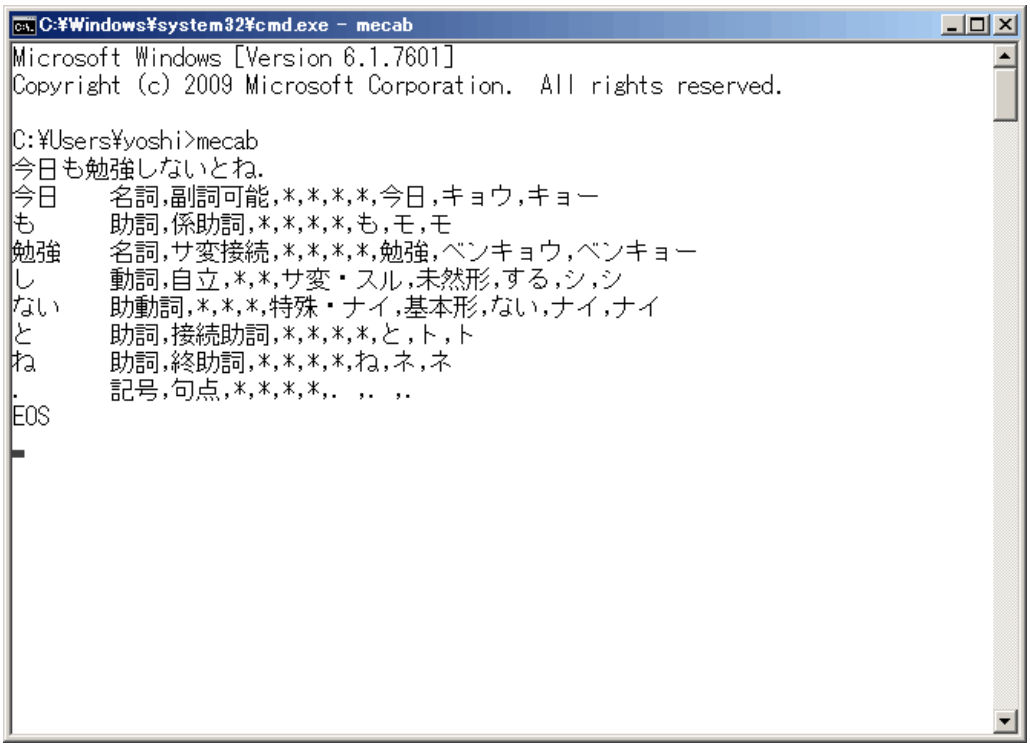
なお，MeCabを用いた 15 世紀朝鮮語の自動形態素解析に関する先行研究[2]としては，村田寛(2010)と須賀井義教・村田寛(2011)がある．前者はMeCabによる 15 世紀朝鮮語の形態素解析が可能であることを明らかにし，後者はその延長として解析率をいかに向上させるか，検討したものである．特に後者では，辞書の登録項目数が多いほど解析率が向上し，解析ミスにも一定のパターンが見られることが分かった．こうした知見を生かし，須賀井義教(2011)では現代朝鮮語の解析を試みた．結果として，後述する「連接コスト」の値を修正することで，解析率がより向上することが明らかになり，さらにインターネットを利用した解析についても例を示した．

本稿では MeCab とその辞書構築について概観した後，解析結果を利用する方法について検討する．

2. MeCab とは

本稿で利用するMeCab[3]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．コマンドラインから実行するプログラムで，直接文を入力するか，入力ファイルを指定して解析を行う．MeCabとともに配布されている日本語解析用の辞書を用いて，「今日も勉強しないとね.」という文を解析すると，図 1 のような結果が得られる：

図 1：MeCab の実行画面



2) 朝鮮語の解析ではないが，MeCab で古典中国語を解析する試みとして守岡知彦(2008)がある．

3) 本稿執筆時点でのバージョンは 0.993 である．詳細については以下のホームページを参照のこと．
http://mecab.sourceforge.net/

解析した結果は，表層形に続き「素性」をコンマで区切った形式で出力される．この素性は自由に定義することができ，いくつでも連ねることができる．この素性を定義した辞書ファイルを用意することで，日本語以外の言語の解析に利用することが可能である．

なお，解析結果はテキストファイルに保存することができ，それを他のツールで利用することが可能である．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を使えば，ChaSen（茶筌）[4]やMeCabによる形態素解析済みテキストを読み込んで，形態素タグなどを組み合わせた条件によるKWIC検索や統計処理を行うことができる．また係り受け解析済みのテキストを用いて検索などを行える．こうしたツールの存在も，MeCabの利用を積極的に後押ししてくれるものである．

さらに，PerlやRubyといった他のプログラム言語からMeCabを扱うためのライブラリも用意されているため，単なる形態素解析だけでなく，より複雑な処理を行うことが可能になっている．例えば，須賀井義教(2011)で紹介したインターネットでの形態素解析は，PerlによるCGIとMeCabとを組み合わせたものである [5]．

3. 解析用辞書の構築

3.1. 辞書構築に必要なもの

さて，MeCabで形態素解析を行うための辞書を構築するには，Seed辞書，学習用データが必要である．Seed辞書と学習用データを用いて生起コスト，連接コスト [6]などを学習させ，解析用の辞書を構築する．なお，村田寛(2010)，須賀井義教・村田寛(2011)と同様に，本稿でもハングルの表記にローマ字転写 [7]を用いた．MeCabは文字を単位として解析を行うため，さまざまな形態素が文字のなかにとけあってしまう朝鮮語のような場合，その中から形態素を抽出することが難しい [8]．処理に際してローマ字転写を用いることで，ハングルの文字単位ではとらえることのできない形態素を抽出することができる．アルファベットは内部的に用いるため，入力，出力においてはハングルで表示するようにプログラムを作成すれば，ローマ字転写の部分は目に触れずに済む．

---

4) 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座松本研究室により開発された日本語形態素解析器．詳細については，以下のホームページを参照のこと．
`http://chasen-legacy.sourceforge.jp/`

5) インターネットを通じた形態素解析を以下のサイトで公開している：
`http://porocise.sakura.ne.jp/korean/morph/analyzer.html`

6) 「生起コスト」とはその形態素の出現しやすさを表し，「連接コスト」とは二つの形態素のつながりやすさを表す．コストが低いほど現れやすいことを示す．MeCab はこれらのコストの和が最小になる場合に正解と判定する「接続コスト最小法」（松本裕治ほか 1997:62–63）を採用している．特に連接コストについては「日本テレビ東京で学ぶ MeCab のコスト計算」が分かりやすい（`http://www.mwsoft.jp/programming/munou/mecab_nitteretou.html`）．

7) 本稿で用いたローマ字転写法は，アクセントを除き福井玲(1989)の転写法に従う．本稿で用いた転写法は付録 1 の通りである．実際のハングル表記の例とそのローマ字転写の例を資料 1 に挙げておく．

8) ハングル表記を利用して，文字を単位として解析を行う場合，例えば「주금괴라」（殺すことである：三綱行実図・忠臣図）という語節から名詞形成接尾辞の「-로」や共同格語尾の「-과」を取り出すことは不可能である．

3.2. Seed 辞書の準備

本稿では『釈譜詳節』巻六をデータとして辞書の構築と解析を行った[9]．15 世紀朝鮮語の辞書を構築する手順は，次の通りである．最初に，学習用コーパスを作成し，そのコーパスに現れる形態素片[10]をもとにSeed辞書を構築する．形態素片が辞書の表層形である．そして，そのSeed辞書を学習用コーパスで学習させ，配布用辞書を構築する．

なお，Seed 辞書の素性は次のように記述する：

(1) 品詞 1，品詞 2，品詞 3，活用語基，語基接続の情報，基本形，表層形

品詞 1 は大分類であり，品詞 2，品詞 3 はその細分類である．ただし，今回は作業時間の関係で基本形と表層形は入力していない．

また，本稿で設定した品詞体系を示せば次の通りである：

(2) 本稿での品詞体系概要

品詞 1	品詞 2
名詞（Noun）	普通名詞，代名詞，数詞，指示詞，固有名詞，接尾語
動詞（Verb）	自立
指定詞（Siteisi）	非自立
存在詞（Sonzaisi）	自立
副詞（Adverb）	一般，不可能
語尾（Ending）	語尾
接頭辞（Prefix）	
接尾辞（Suffix）	尊敬，謙譲，回想，支えのイ，名詞形成
後置詞（Postposition）	

9) 村田寛(2010)，須賀井義教・村田寛(2011)でも『釈譜詳節』をメインの資料として利用したが，本稿ではその後の検討を元に，設定ファイルや辞書の記述などを大幅に見直している．なお，本稿で用いた『釈譜詳節』初刊本の電子データは，東京外国語大学大学院の趙義成先生に提供して頂いた．この場を借りて感謝申し上げる．

10) 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，分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．

Seed辞書の例を一部挙げれば，以下のとおりである[11]：

(3) 'a0d@r1,0,0,0,Noun,普通名詞,一般,*,*,*
'a0ra1,0,0,0,Verb,自立,*,語基 3,*,*,*
'ai1,0,0,0,Ending,語尾,処格,*,*,*,*
ge0dyn1,0,0,0,Ending,語尾,接続形,*,接続 1,*,*
sia1,0,0,0,Suffix,尊敬,*,語基 3,*,*,*

学習用データを用いて学習を行うと，Seed 辞書に登録された項目についてコストの設定が行われる．上記(3)の項目は，学習後の解析用辞書では次のように記述されている：

(4) 'a0d@r1,41,41,3093,Noun,普通名詞,一般,*,*,*,*
'a0ra1,86,86,1596,Verb,自立,*,語基 3,*,*,*
'ai1,8,8,2155,Ending,語尾,処格,*,*,*,*
ge0dyn1,15,15,1975,Ending,語尾,接続形,*,接続 1,*,*
sia1,61,61,1411,Suffix,尊敬,*,語基 3,*,*,*

なお，学習用データは MeCab の出力と同じ形式で記述したテキストファイルを用いる：

(5) 世尊 Noun,普通名詞,一般,*,*,*,*
'i1 Ending,語尾,主格,*,*,*,*
象頭山 Noun,固有名詞,地名,*,*,*,*
'ai1 Ending,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,語尾,共同格,*,*,*,*

学習用データの作成においては，本文と割注部分とを分けて作成した．連接という点で無関係の要素が文中に現れると，学習に影響を及ぼす恐れがあると考えたためである．本文と割注とを分けることについては，実際に影響があるかどうかを検証する必要があるが，本稿ではいったん分ける措置を取ったことを断っておく．

さて，MeCab では文法に基づく活用の展開を行わない．そのため，辞書に全ての活用形を記

11) コンマで区切られた素性のうち，2 番目は左文脈 ID，3 番目は右文脈 ID，4 番目は生起コストである．Seed 辞書の段階では学習が行われていないため，これらの値に 0 を設定しておく．また，該当する素性が空白の場合，「*」で示される．

述しておく必要がある．例えば日本語の場合，MeCab とともに配布されている日本語 IPA 辞書から「書く」の項目を一部挙げれば，以下の通りである：

(6) 書く,679,679,7325,動詞,自立,*,*,五段・カ行イ音便,基本形,書く,カク,カク
書か,683,683,7745,動詞,自立,*,*,五段・カ行イ音便,未然形,書く,カカ,カカ
書こ,681,681,7299,動詞,自立,*,*,五段・カ行イ音便,未然ウ接続,書く,カコ,カコ
…
書きゃ,677,677,7123,動詞,自立,*,*,五段・カ行イ音便,仮定縮約１,書く,カキャ,カキャ

朝鮮語の場合，語基の概念を用いて活用語基を記述しておくことで，日本語の場合と同様に扱うことができる．「가-」（行く）を例に挙げれば以下の通りである[12]：

(7) ga0 Verb,自立,*,語基 1,*,*,*
ga1 Verb,自立,*,語基 1,*,*,*
ga0 Verb,自立,*,語基 2,*,*,*
ga1 Verb,自立,*,語基 2,*,*,*
ga1 Verb,自立,*,語基 3,*,*,*
ga1'a1 Verb,自立,*,語基 3,*,*,*
ga2 Verb,自立,*,語基 4,*,*,*

こうした記述の方式は，用言だけでなく体言にも適用される．例えば「·몸」（体）は以下のような項目が考えられる：

(8) a. mom1 Noun,普通名詞,一般,*,*,*,*
b. mo1m Noun,普通名詞,一般,*,*,*,*
c. mo0m Noun,普通名詞,一般,*,*,*,*

(8a)は単独で用いられた「·몸」の場合，(8b)は「·모·믈」（体を）の場合，(8c)は「모·매」（体に）の場合の登録項目である．

以上のように，15 世紀朝鮮語の辞書項目作成に当たっては，ローマ字転写に加えアクセントも表示しているため，登録すべき項目が非常に複雑である．アクセント変動などはある程度予測がつくため，あらかじめ機械的に変換した形式を辞書項目として登録しておくことができるかもしれない．ただし，文献に現れる形式は錯誤も含めて様々であるため，基本形からの機械的な変換はなかなか難しい．(8)に示したように，こうした異表記の問題は動詞に限ったことではない．

なお，解析後のデータを用いて単語の出現頻度など，統計的処理を行うような場合には，様々

12) 以下の例示では，煩雑を避けるため MeCab の出力形式で表示する．

な異表記を一つの単語としてくくるための素性が必要になる. そのため, (1)で示した素性のうち, 「基本形」のような項目は必須となる. これをローマ字で転写するか, ハングルで記述するかはともかくとして, 「基本形」の追加は今後の課題といえる.

今のところ, こうした異表記は文献に現れた形式だけをデータとして利用しているため, とりあえず問題はないが, 今後様々な文献に対応できる辞書を作成しようとする場合, あらかじめ辞書の項目を増やしておかなければならない. その際にどう解決するか, 検討が必要である.

3.3. 融合する語形について

「·고ㅎ」(鼻) のようないわゆるㅎ末音の名詞に「ㄷ, ㄱ」が続く場合, 「·고·쏴」(鼻と) のように, 後続する子音が激音となる (安秉禧·李珖鎬 1990:148-149 など参照). こうした場合にも表層形をそれぞれ「·고」と「·쏴」に分け, それぞれ登録することとした. これらの辞書項目は次の通りとなる :

(9) go1 Noun,普通名詞,一般,*,*,*,*
koa1 Ending,語尾,共同格,*,*,*,*

また, 用言「ᄒᆞ-」と語尾との結合においても同様に激音で現れる場合があり, これも(9)と同様に扱うこととする :

(10) 正 Noun,普通名詞,h@語根,*,*,*,*
kei1 Ending,語尾,接続形,*,接続 1,*,*
＜「正·케」

(10)の例では接続形語尾の「-게」について, 「·케」という異表記を辞書の項目として登録している.

なお, 「h@語根」とは, 「ᄒᆞ-」に前接することができるが単独では名詞として用いられない要素を含める. 「正ᄒᆞ-」などを一つの用言として辞書に登録してしまうと, (10)のような場合の記述が困難になる. 「ᄒᆞ-」に前接する要素が単独で名詞として用いられる場合も含め, 全て分離して項目を登録することで, 「正」と「正ᄒᆞ-」の両方を辞書に登録する手間を省くことができ, 結果として辞書のサイズを抑えることができると考える.

3.4. 構築した辞書の解析率

このようにして作成したSeed辞書の項目数は 1,709 項目である. このSeed辞書から構築した解析用辞書の解析率 [13]は以下のとおりである :

13) 解析率の計算には, MeCab に付属の評価用プログラムを用いた. これは出力結果と正解とで一致する形態素片の数を, 出力結果の形態素片の数, 正解の形態素片の数を分母として比率を計算し, 両者の平

(11) 解析率一覧（単位は%）

	品詞 1 のみ	全素性
『釈譜詳節』巻六本文	99.0216	96.2521
『釈譜詳節』巻六割注	95.4600	93.9068

以上のように，全ての項目が辞書に登録されているにも関わらず，解析率が 100%となることはない．しかし，形態素片の切り出しと品詞の付与についてはほぼ 100%に近いといえる．

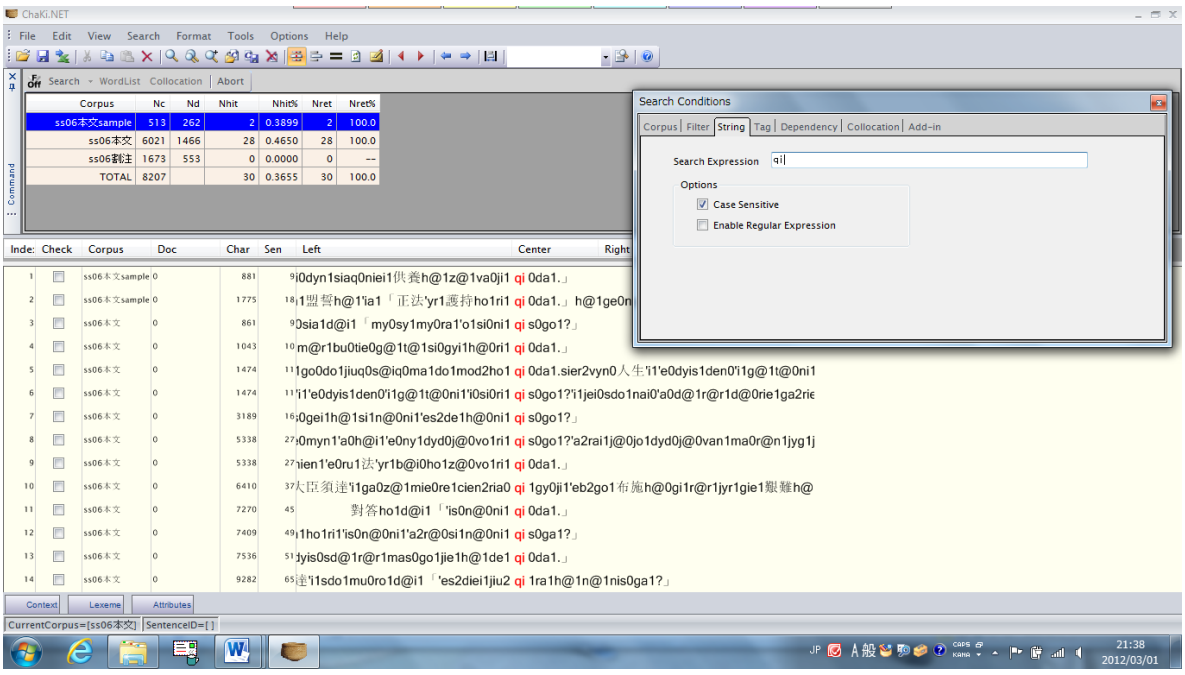
解析の誤りを見ると，母音語幹の用言において，第 I 語基と第 II 語基の判定を誤っているケースが多くみられた．須賀井義教(2011)でも行ったように，辞書構築後に連接コストを修正することで，こうした誤りの大部分を抑えることができると考えられる．本稿では作業の都合上この修正を行うことができなかった．

MeCab で解析した結果はテキストファイルとして保存できる．このファイルを他のプログラムで利用することが可能である．以下では，そのような事例の一つとして「ChaKi（茶器）」による解析結果の利用について紹介する．

4. 応用の事例：ChaKi（茶器）の利用

ChaKi(茶器)は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である。

図 1　文字列 qi で検索した KWIC 索引



均を算出するものである．詳細については須賀井義教(2011)で述べているので参照されたい．

図 2　係り受け処理した文

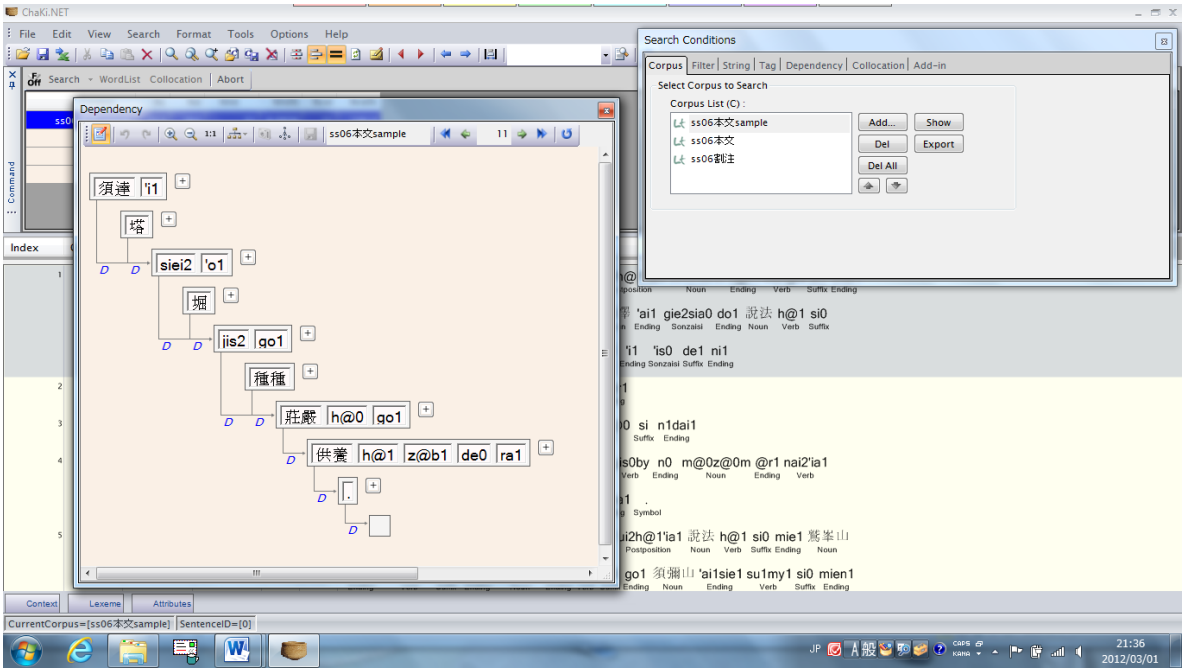
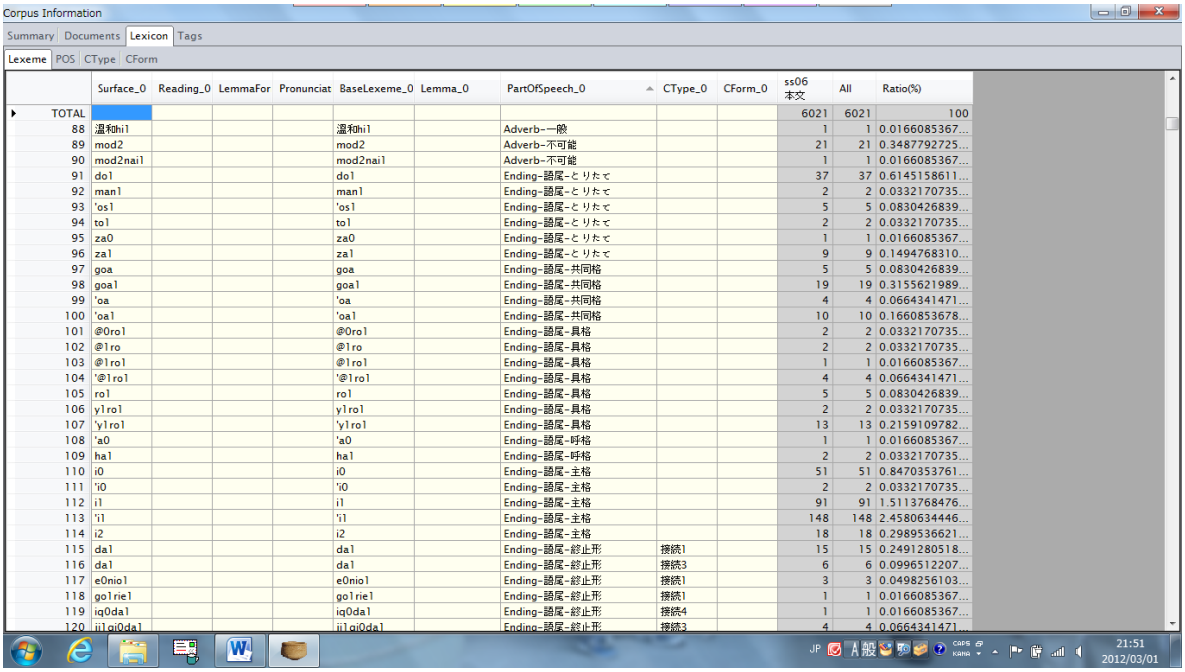


図 3　形態素片の統計的情報



	Surface_0	Reading_0	LemmaFor	Pronunciat	BaseLexeme_0	Lemma_0	PartOfSpeech_0	CType_0	CForm_0	ss06 本文	All	Ratio(%)
TOTAL										6021	6021	100
88	溫和hi1				溫和hi1		Adverb-一般			1	1	0.0166085367...
89	mod2				mod2		Adverb-不可能			21	21	0.3487792725...
90	mod2nai1				mod2nai1		Adverb-不可能			1	1	0.0166085367...
91	do1				do1		Ending-語尾-とりたて			37	37	0.6145158611...
92	man1				man1		Ending-語尾-とりたて			2	2	0.0332170735...
93	'os1				'os1		Ending-語尾-とりたて			5	5	0.0830426839...
94	to1				to1		Ending-語尾-とりたて			2	2	0.0332170735...
95	za0				za0		Ending-語尾-とりたて			1	1	0.0166085367...
96	za1				za1		Ending-語尾-とりたて			9	9	0.1494768310...
97	goa				goa		Ending-語尾-共同格			5	5	0.0830426839...
98	goa1				goa1		Ending-語尾-共同格			19	19	0.3155621989...
99	'oa				'oa		Ending-語尾-共同格			4	4	0.0664341471...
100	'oa1				'oa1		Ending-語尾-共同格			10	10	0.1660853678...
101	@0ro1				@0ro1		Ending-語尾-具格			2	2	0.0332170735...
102	@1ro				@1ro		Ending-語尾-具格			2	2	0.0332170735...
103	@1ro1				@1ro1		Ending-語尾-具格			1	1	0.0166085367...
104	'@1ro1				'@1ro1		Ending-語尾-具格			4	4	0.0664341471...
105	ro1				ro1		Ending-語尾-具格			5	5	0.0830426839...
106	y1ro1				y1ro1		Ending-語尾-具格			2	2	0.0332170735...
107	'y1ro1				'y1ro1		Ending-語尾-具格			13	13	0.2159109782...
108	'a0				'a0		Ending-語尾-呼格			1	1	0.0166085367...
109	ha1				ha1		Ending-語尾-呼格			2	2	0.0332170735...
110	i0				i0		Ending-語尾-主格			51	51	0.8470353761...
111	'i0				'i0		Ending-語尾-主格			2	2	0.0332170735...
112	i1				i1		Ending-語尾-主格			91	91	1.5113768476...
113	'i1				'i1		Ending-語尾-主格			148	148	2.4580634446...
114	i2				i2		Ending-語尾-主格			18	18	0.2989536621...
115	da1				da1		Ending-語尾-終止形	接続1		15	15	0.2491280518...
116	da1				da1		Ending-語尾-終止形	接続3		6	6	0.0996512207...
117	e0nio1				e0nio1		Ending-語尾-終止形	接続1		3	3	0.0498256103...
118	go1rie1				go1rie1		Ending-語尾-終止形	接続1		1	1	0.0166085367...
119	iq0da1				iq0da1		Ending-語尾-終止形	接続4		1	1	0.0166085367...
120	ji1qi0da1				ji1qi0da1		Ending-語尾-終止形	接続3		4	4	0.0664341471...

5.　おわりに

ここまで，15 世紀朝鮮語を MeCab で形態素解析するための辞書構築について述べてきた．解析率の向上という点で課題があるものの，本稿で構築した MeCab 用辞書を利用することで，手作業で解析をするよりは断然効率が良いと考えられる．おわりに，いくつか問題点と課題について述べておく．

まず，誤解析の問題が挙げられる．特に同音異義の形態素の判定に誤りがみられるケースがある．例として「耶輸ㅣ 부텻 使者 ·왯·다 드르·시·고」（耶輸が仏の使者が来ているとお聞きになって, 2a) についてみてみよう．このうち「使者 ·왯·다」（ローマ字転写では「使者 'oais1da1」）という部分の解析結果を示せば以下のとおりである：

(12) 使者　　Noun,普通名詞,一般,*,*,*
'oa　　Ending,語尾,共同格,*,*,*
is1　　Sonzaisi,自立,*,語基 1,*,*
da1　　Ending,語尾,終止形,*,接続 1,*,*

もちろん，実際の文脈上(12)の解析は誤りといえるが，文法的に誤りとは言いがたい．MeCab では学習されたコストを元に，基本的に一つの最適解だけが出力されるため，可能な候補がいくつかあったとしても，最終的な出力は一つだけである．. ただし，MeCab の実行時にオプションをつけることで，複数の解析結果を表示することはできる．例えば解析結果の表示を 2 つに増やすと，上記(12)の結果に次いで以下の「正解」が出力される：

(13) 使者　　Noun,普通名詞,一般,*,*,*
'oa　　Verb,自立,*,語基 3,*,*,*
is1　　Sonzaisi,自立,*,語基 1,*,*,*
da1　　Ending,語尾,終止形,*,接続 1,*,*

しかし，解析の途中で複数の候補からユーザーが一つを選ぶ，というような利用はできない．MeCab 単独の機能では実現できないため，他のプログラミング言語と合わせて出力を調整できるようにする必要があるだろう．また，学習用データの量を増やすことで，「確からしい」解析結果を得ることはできるようになると思われるが，最適でない結果が必要な場合，どのように出力させるか，やはり検討すべき問題である．

これと関連して，データ量の問題がある．ここでは『釈譜詳節』巻六だけを対象としているため，学習のためのデータ量としては圧倒的に少ない．MeCabの学習モデルでは少ない学習用データで効率的に学習できるようになっているが，生起コストの学習などといった面から考えれば，やはり学習用データは多ければ多いほどよいと思われる [14]．また，学習用データだけでなく，辞書データも非常に少ない．本稿で構築した辞書を用いて他の文献を解析しても，かなりの部分が誤りとなってしまう．試みに『釈譜詳節』巻九の冒頭，4 帳表の 2 行目までの本文のみ（形態素片の数 202）を解析したところ，全素性での解析率が約 77.4%であった．辞書に登録されていな

14) MeCab に同梱されている日本語 IPA 辞書の学習には，4 万文程度のデータを使っているという．「気まぐれ日記：Yahoo!の形態素解析を MeCab で無理やり再現してみる」（http://chasen.org/~taku/blog/archives/2007/06/yahoomecab.html）を参照.

い項目，すなわち未知語の処理において，形態素片の切り出しに失敗しているケースが多くみられた．これらの問題を解決するには，結局のところ現在の辞書で解析を行い，その結果を修正して辞書と学習用データに加え，さらに辞書を構築して文献を解析する…，という手順で作業を進めるほかない．それにつれて解析率も向上するといえよう．

以上を今後の課題とし，作業を進めていきたいと考える．

参考文献

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(2011)「MeCab を用いた現代韓国語の形態素解析」，第 50 回朝鮮語教育研究会例会（2011 年 6 月 12 日，京都女子大学）話題提供要旨
（http://porocise.sakura.ne.jp/archive/paper/mecabdic.pdf にて公開）

須賀井義教・村田寛(2011)「15 世紀朝鮮語の形態素解析について」『教養・外国語教育センター紀要』第 1 巻第 2 号，近畿大学教養・外国語教育センター，pp.41-56

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，pp.17-29

松本裕治ほか(1997)『単語と辞書』岩波講座　言語の科学 3，岩波書店

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，pp.17-28

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-CH-73，情報処理学会，pp.17-22

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』，情報処理学会，pp.98-99

山本和英(2000)「計算機処理のための韓国語言語体系と形態素処理」『自然言語処理』Vol. 7 No.4，言語処理学会，pp.25-62

강승식(2002;2003) “한국어 형태소 분석과 정보 검색”(수정판), 홍릉과학출판사

김한샘(2005) “현대 국어 사용 빈도 조사 2”(국립국어원 2005-1-33), 국립국어원（国立国語院ホームページより入手可能）

安秉禧·李珖鎬(1990) “中世國語文法論”, 學研社

유혜원(2004) “한국어 정보 처리의 이론과 실제”, 제이앤씨

조남호(2002) “현대 국어 사용 빈도 조사—한국어 학습용 어휘 선정을 위한 기초 조사—”(국

립국어연구원 2002-1-17), 국립국어연구원（国立国語院ホームページより入手可能）
조남호(2003) “한국어 학습용 어휘 선정 결과 보고서” (국립국어연구원 2003-1-4), 국립국어연구원(国立国語院ホームページより入手可能)
황화상(2006) “한국어와 정보”, 박이정
John Lafferty, Andrew McCallum, and Fernando Pereira(2001) *Conditional random fields: Probabilistic models for segmenting and labeling sequence data.* In Proc. of ICML

付録　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	ㅅㄱ	ㅅㅂ	ㅂㄷ	ㆅㆅ	ㅂㅅㄱ	ㅂㅅㄷ
転写	sg	sb	bd	hh	bsg	bsd

＜アクセント＞

平声	去声	上声
0	1	2

資料１　『釈譜詳節』巻六の冒頭部分（{} 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.]
○ 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·ᄒᆞ숩·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 ·곧·게 ᄒᆞ·리·라.」｝