様式 C-19

科学研究費補助金研究成果報告書

平成２３年　６月　１日現在

機関番号：３４３１０
研究種目：基盤研究（C）
研究期間：2007～2010
課題番号：１９５２０３７５
研究課題名（和文）　朝鮮語 Web 辞典の設計

研究課題名（英文）　Newly designed Korean-Japanese Web-dictionary

研究代表者
油谷　幸利（YUTANI YUKITOSHI）
同志社大学・言語文化教育研究センター・教授
研究者番号：５０１２２３６２

研究成果の概要（和文）：
１）見出し語が1万語を突破し，朝鮮語辞典としても十分な実用性を備えるにいたった。
２）句単位の形態素解析を行うことに加えて，文単位での形態素解析を明示的に提示することにより，朝鮮語Web辞典利用者に対する文法学習に大きく貢献することを新たな目標として設定した。従来等閑視されてきた同形異語に対する知見を形態素解析に導入してきたが，これを文単位に拡張することにより，形態素解析の精度を上げることが期待できる。

研究成果の概要（英文）：
1) Web dictionary exceeds 10,000 words. Our system can be used as a practically Korean-Japanese dictionary.
2) In addition to phrase-based morphological analysis, we have also set up a new sentence-based morphological analysis that will clearly provide a platform for those wanting to learn Korean through our system. Previously thought to be underestimated, we have introduced a new homonyms analyzer which uses several different structures and has produced positive results. We expect to improve the accuracy of the morphological analysis in the future.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	1,100,000	330,000	1,430,000
２００８年度	800,000	240,000	1.040,000
２００９年度	800,000	240,000	1.040,000
２０１０年度	700,000	210,000	910,000
年度			
総　計	3,400,000	1.020,000	4,420,000

研究分野：言語学
科研費の分科・細目：言語学・言語学
キーワード：外国語，辞書，インターネット，アルゴリズム

１．研究開始当初の背景

朝鮮語の代表的な Web 辞典としては，韓国国立国語院が構築・運営している『標準国語大辞典』および延世大学言語情報開発研究院が構築･運営している『延世韓国語辞典』，韓国ヤフーが構築･運営している各種の辞典類，ウィキペディアの韓国語版，などを挙げることができるが，いずれも韓国語母語話者のための韓国語辞典であり，日本語母語話者のための朝鮮語 Web 辞典は，油谷が実験的に構築・運営を始めたものしか存在しない．

２．研究の目的

助詞を取ることによって同形になったものを多数収録することによって初級段階の朝鮮語学習者に有益な情報を提供する朝鮮語 Web 辞典を設計し，10,000 語程度の見出し語とアルゴリズムによる形態素解析機能を備えた朝鮮語 Web 辞典を一般に広く公開することを目的とする．検索方法としては以下のようなものがある．

１）見出し語検索では用言の原形･体言･副詞・冠形詞・感嘆詞・助詞･語尾・接頭辞･接尾辞以外に，‘침이 마르도록’のような慣用句も検索できる．

２）全文検索は本辞典に収録されたデータの範囲内で全ての用例を検索するものであり，間に空白を含む任意の文字列を用いての用例検索が可能である．本格的な用例検索を行うにはコーパスを利用すべきであるが，用例の確認程度という限界をわきまえた上での利用であれば大きな支障はなかろう．

３）本辞典では一覧性の欠如という欠点を補うために，前後の項目を参照できるように，クリックするだけで当該の項目が検索できるジャンプ機能を装備している．

４）類義語や関連項目などが参照できるように解説参照機能を付加し，クリックするだけで当該の項目が検索できる．

５）見出し項目が同形異語を有する場合には，同形異語の形式を出力して該当項目にジャンプできる機能が組み込まれている．

６）アルゴリズムによる形態素解析は理論上の可能性を示したものに過ぎず，実際にそのような単語が存在するか否かは保証の限りではない．本辞典では参照機能を利用して解析結果が本辞典の見出し項目に存在するか否かを確認できる仕様である．

３．研究の方法

（１）用例収集に際しての方針を以下のとおりに立てた．

１）用例は新聞や雑誌･文学作品などで実際に使用されたものに限り，作例や既存の辞書・対訳書からの引用は採用しない．

２）用例は必ず出典を明示する．

３）用例には日本語訳を添える．

４）体言については主要な助詞を伴う用例や典型的な動詞との組み合わせ，慣用句などの用例を収集する．

５）用言については連体形や終止形，主要な語尾を伴う用例，典型的な名詞との組み合わせ，慣用句などの用例を収集する．

（２）アルゴリズムによる形態素解析は以下のような手順で段階的に行った．

１）初級段階の文法項目における形態素解析を中心に作業を行う．

２）中級段階の文法項目における形態素解析を中心に作業を行う．

３）形態素解析の一般的なアルゴリズムでは記述できない例外的な同形異語を中心に作業を行う．

４）形態素解析の再帰的なアルゴリズムを中心に作業を行う．

４．研究成果

（１）2011年3月段階で，見出し語が1万語を突破し，朝鮮語辞典としても十分な実用性を備えるにいたった。

（２）形態素解析

１）未登録の見出し語に対する形態素解析

筆者が目指している朝鮮語 Web 辞典の完成までには恐らく 20 年以上の歳月を要することが予想される。見出し項目が 800 に満たない段階で本辞典の公開に踏み切ったのは，(ⅰ)動作確認をしながら開発を進める必要があったことと，(ⅱ)用例収集に対する協力依頼を呼びかけたかったという理由以外に，(ⅲ)未登録語に対する検索が要求された際に，アルゴリズムによる形態素解析を行って理論的にありうる単語を提示することにより，見出し項目が少なくてもある程度実用に耐えうるであろう，という見通しがあったからである。

アルゴリズムによる形態素解析は，理論的に存在しうる形態素列を予測するための手順を示したものである。実際にそのような単語が存在するか否かは保証の限りではない。実際にありえないものを排除しないのは，学習者にさまざまな可能性が存在することに気付いてもらいたいためである。

２）語節単位での形態素解析

2007 年 9 月末時点で実装した，アルゴリズムによる形態素解析結果の一例を以下に示す。

①ａが먹는の場合はａを먹(ｂ)と는に分割する。

②ｂは母音で終わらないので「먹(動詞語幹)＋는(現在連体形語尾)」と表示して終了

する。

③ａが파는の場合はａを파(ｂ)と는に分割する。

④ｂは母音で終わっているのでｃを팔とする。

⑤以下の３行を表示して終了する。

파(動詞語幹)＋는(現在連体形語尾)

팔(動詞語幹)＋는(現在連体形語尾)

파(体言･副詞)＋는(助詞「は」)

３）解析規則の再帰的な適用

周知の如く朝鮮語は膠着語であるために，助詞や語尾がいくつも繋がることがある。4.(2)2)⑤で示した파는では３通りの可能性しか存在しなかったが，文字列が長くなれば幾何級数的に可能性が増えていく。連体形の는であれば，판다는，팠다는，파느냐는，파자는，파라는，파려는，などがすぐに思い浮かぶし，助詞の는であれば，파끼지는，파에는，파로는，파에까지는，파하고는，など枚挙に暇がない。これらを網羅的に数え上げるためには解析規則を再帰的に適用するようにアルゴリズムを設計しておく必要がある。

例えば，‘사기까지는’の解析結果は以下のようになる。

사기까지(動詞語幹)＋는(現在連体形語尾)
【参照】사기까지다
사기까질(動詞語幹)＋는(現在連体形語尾)
【参照】사기까질다
사기까지(体言･副詞)＋는(主題助詞 「 は 」)
【参照】사기까지
사기(体言)＋까지(助詞「まで」)＋는(主題助詞「は」)　　【参照】사기
사(用言語幹)＋기(こと)＋까지(助詞「まで」)＋는(主題助詞「は」)　【参照】사다

この中で実際に存在するものは「士気(史記/死期/沙器/邪気/詐欺)までは」と「買うことまでは」の２通りである。これらを全て‘는’の処理ルーチンに記述しておくことは非効率的であり，プログラムにバグが発生する可能性も高くなる。本辞典では網羅性と効率性を両立させるために助詞または語尾‘는’の前に出現可能な助詞や語尾を網羅的に記述した後，そこから先の処理はそれらの助詞や語尾のルーチンに委ねるように設計する予定である。なお，これらの用例を確認するには，【参照】マークに続く理論上の原形や体言･副詞をクリックすればよい。もちろん，사기あるいは사다をクリックすれば適切な用例が出力されるが，사기까지다や사기까질다という用言あるいは사기까지という体言･副詞は実際には存在しないので，これらをクリックしても「未登録語です：사기까지다」のように出力されるだけである。

４）同形異語の見出し語への登録

朝鮮語 Web 辞典においては，初級段階の学習者が見落としがちな同形異語を網羅的に数え上げるためにアルゴリズムを開発したわけであるが，中にはこのアルゴリズムだけでは分析できないものが存在する。

たとえば，3.1.3.1.のアルゴリズムに従えば‘가는’は①가(体言)＋는(助詞「は」)，②가(動詞語幹)＋는(現在連体形語尾)，③갈(動詞語幹)＋는(現在連体形語尾)，の３通りに分析できることが予測され，事実これらは全て実際に存在しうる。しなしながら‘가는’にはさらに가늘다(細い)という形容詞の現在連体形である可能性が存在するが，このアルゴリズムではそれを示すことができない。このような，通常のアルゴリズムでは予測不可能なものに関しては，アルゴリズムに例外処理を組み込むとともに，見出し語にも登録しておく必要がある。

また，‘ａ＋은’という文字列はａがパッチムを有していれば①ａ(体言)＋은(助詞「は」)，

②a (用言語幹)＋은(連体形語尾)，の３通りに分析できる．通常は a が母音で終わることはありえないが， a が모あるいは지であれば모으다の過去連体形あるいは짓다の過去連体形という通常のアルゴリズムでは得られない例外的な解析結果が要求される．

上記の例はほんの一例に過ぎず，アルゴリズムによる形態素解析が実用的であるためにはこれらの例外的処理を必要とする文字列を細大漏らさず数え上げ，適切な位置に処理装置を実装しておく必要がある。

（３）文単位での形態素解析

2010 年 11 月に至り，句単位の形態素解析を行うことに加えて，文単位での形態素解析を明示的に提示することにより，朝鮮語 Web 辞典利用者に対する文法学習に大きく貢献することを新たな目標として設定した。これに伴い，プログラミング言語を perl から php および SQLite に変更した。

本辞典では，従来等閑視されてきた同形異語に対する知見を形態素解析に導入してきたが，これを文単位に拡張することにより，形態素解析の精度を上げることが期待できる。現在すでに，インターネットを通して，韓日翻訳が利用できる状況が存在するが，同形異語に対する認識が低く，解釈の柔軟性に欠ける。たとえば，エキサイト翻訳[1]では，「이건 먹이고 저건 종이예요.」という入力に対して「これは食べさせてあれは鐘です。」という解釈しか提示されなかったが[2]，「これは墨で，あれは紙です。」という解釈が文脈に最も適した翻訳であろう。

まだ実験中ではあるが，本辞典で同じ文を入力すると，以下のような分析結果が表示される。

形態素解析の開始　【이건 먹이고 저건 종이예요】
【解析 1】
이건　〔句〕主題　〔意味〕これは＝이건
먹이다　〔他動詞 8〕먹이（母音語幹）
〔意味〕食べさせる
＝먹이（用言語幹）＋고（並列語尾：て）＝【参照】먹이다
저건　〔句〕主題　〔意味〕あれは＝저건
종이　〔名詞〕物質
〔意味〕紙＝종이（母音体言）＋이（指定詞語幹）＋에요（略式
丁寧体語尾＝平叙文・疑問文）
平叙文
【解析 2】
이건　〔句〕主題　〔意味〕これは＝이건
먹이다
먹이（母音体言）＋이（指定詞語幹省略）
〔意味〕餌＝먹이（用言語幹）＋고（並列語尾：て）＝【参照】먹이
저건　〔句〕主題　〔意味〕あれは＝저건
종이　〔名詞〕物質
〔意味〕紙＝종이（母音体言）＋이（指定詞語幹）＋에요（略式
丁寧体語尾＝平叙文・疑問文）
平叙文
【解析 3】
이건　〔句〕主題　〔意味〕これは＝이건
먹이다
먹（子音体言）＋이（指定詞語幹）
〔意味〕墨＝먹이（用言語幹）＋고（並列語尾：て）＝【参照】먹
저건　〔句〕主題　〔意味〕あれは＝저건
종이　〔名詞〕物質

[1] http://www.excite.co.jp/world/korean/
[2] 종이예요という綴り字は母音体言종이に指定詞이다が接続した形ではありえないので，子音体言の종(鐘)という解釈はそもそも誤訳である。

〔意味〕紙＝종이（母音体言）＋이（指定詞語幹）＋에요（略式丁寧体語尾＝平叙文・疑問文）
　平叙文
分析完了。

먹이고に対して3通りの可能性が存在するので，3通りの解析結果を出力している画面である。以下に該当する部分だけを取り出してみる。

【解析1】
먹이다　〔他動詞8〕먹이（母音語幹）
〔意味〕食べさせる
＝먹이（用言語幹）＋고（並列語尾：て）＝【参照】먹이다

【解析2】
먹이다
먹이（母音体言）＋이（指定詞語幹省略）
〔意味〕餌＝먹이（用言語幹）＋고（並列語尾：て）＝【参照】먹이

【解析3】
먹이다
먹（子音体言）＋이（指定詞語幹）
〔意味〕墨＝먹이（用言語幹）＋고（並列語尾：て）＝【参照】먹

現時点では文脈に応じたものだけを抽出する機能はごく一部しか組み込まれていないが，本辞典が完成すれば、韓日機械翻訳の精度が一段階向上することが期待できる。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計５件）
①油谷幸利，朝鮮語問題バンクについて，朝鮮語教育，朝鮮語教育研究会，査読有，6，2011，1-27
② 金 亨 貞，선행명사구의 유정성과 조사「에게/에」의 선택(先行名詞句 の 有情性 と 助詞「ege/e」 の 選択），언어사실과 관점(言語事実と観点），査読有，26集，2010，141-196
③油谷幸利，多言語問題バンクについて，言語文化，同志社大学言語文化学会，査読有，13-2，2010，145-169
④油谷幸利，朝鮮語 Web 辞典について－用例辞典から学習辞典へ－，朝鮮学報，朝鮮学会，査読有，第211輯，2009，(1)-(40)
⑤油谷幸利，朝鮮語 Web 辞典の設計，朝鮮学報，朝鮮学会，査読有，第206輯，2008，(1)-(37)

〔学会発表〕（計４件）
①油谷幸利，「WEB基盤朝鮮語学習辞典について」，第２回国際シンポジウム「韓日・日韓辞典と朝鮮語教育」，2010.12.18，同志社大学寒梅館
②油谷幸利，「朝鮮語問題バンクについて」，朝鮮学会，2010.10.3，天理大学
③油谷幸利，「朝鮮語Web辞典について」―用例辞典から学習辞典へ―，第41回朝鮮語教育研究会，2009.3.28，キャンパスプラザ京都第４講義室
④油谷幸利，「朝鮮語Web辞典の設計について」朝鮮学会第58回大会，2007.10.7，天理大学

〔図書〕（計１件）
①油谷幸利、同形異語をめぐって，野間秀樹編著『韓国語教育論講座』第１巻，くろしお出版，2007，655-674

〔その他〕
ホームページ等
http://paranse.la.coocan.jp/PHP/KLDicV05.php

６．研究組織
(1)研究代表者
油谷　幸利（YUTANI　YUKITOSHI）
同志社大学・言語文化教育研究センター・教授
研究者番号：５０１２２３６２

(2)研究分担者
金　亨貞（KIM　HYEONG-JEONG）
同志社大学・言語文化教育研究センター・助教
研究者番号：２０４５７４１９

(3)連携研究者
なし