漢字シソーラスの構築と語句解析への応用*

4B－6

鈴木英二　中挾知延子　近藤邦雄　佐藤尚　島田静雄†
埼玉大学工学部情報システム工学科‡

1　はじめに

本研究の目的は、漢字１字単位の電子化漢字シソーラスを構築し、それを日本語文章の語句解析へ利用することである。

従来の日本語電子化シソーラスは、名詞を中心に単語別に分類したものが多い。人間の大人の平均的な語彙数は約４万語であり、大量の新語が毎年生ずることを合わせて考えてみると、そのシソーラスのサイズは莫大なものとなる。

そこで、我々は日本語の単語を構成する文字、特に漢字に注目した。漢字は表意文字であり、１字のみで最小の単語の役目を持っている。通常、文章で使われる漢字の総数はJIS第１水準で約3000字であり、これは英米語の基本単語数とほぼ一致する。同時に漢字は、日本語文章において仮名と組み合わせることによって、名詞・用言などの自立語を構成できる柔軟性がある。さらに、漢字には訓読みが与えられており、和語として日本語の語彙を広範に表現できる。その漢字の造語能力の高さが、大量の新語が生ずる原因ともなっているが、新しい漢字の発生とその利用の固定は滅多に起きず、安定した語の集合を保っている。この理由は、漢字の使い方に名詞・動詞・形容詞・副詞など、品詞別の用途に規則があるからである。漢字の有するこれらの特長を利用できれば、日本語文章の解析に役立つと我々は考えた。

また、外国人への日本語教育、とりわけ漢字を教育する時の利用も考慮している。漢字１文字に複数の読み方が与えられており、それが外国人が漢字を学習するに当たって困難さを増している。読み方が解らないために、辞書を引くこともままならないという事態が発生する。そのため、漢字仮名混じり文から読みだけでも解れば、有用なものとなると我々は考えた。

以下、第２章で今回構築した漢字シソーラスの概要を示し、第３章でそれを日本語文章の語句解析へ利用したものの一例を述べる。最後にまとめを第４章として示す。

*Construction of KANJI-thesaurus for Syntactic Analysis of Japanese Documents

†Eiji SUZUKI,Chieko NAKABASAMI,Kunio KONDO,Hisashi SATO,Shizuo SHIMADA

‡Department of Information & Computer Sciences,SAITAMA University

2　漢字シソーラスの概要

まず、JIS第１水準の漢字約3000字を対象にして、以下の３つに分類した。以下の分類において、上にいくほど優先度が高い。そのため、複数の分類にまたがる漢字でも、いずれか１つに属すようにしている。

(1) 形容詞漢字
形容詞・形容動詞の語幹に使う漢字を「形容詞漢字」とする。教科書文法でいう形容詞を「イ形容詞」、同様に形容動詞を「ナ形容詞」、双方合わせて「形容詞」とした。熟語の名詞に「～な」「～に」をつけてナ形容詞になるものを「ナニ名詞」としたが、これに使われる漢字は、本論文では名詞漢字に分類している。これらの分類は、形容詞漢字の読みに影響を与える。

形容詞漢字の総数は約200字である。

(2) 動詞漢字
動詞の語幹に使う漢字を「動詞漢字」とする。「思う」のように、送り仮名を伴って和語となる「和語の動詞」、「愛する」のように、漢字１字に「～する」をつけて動詞を構成する「スル動詞」、「運動する」のように、漢字２字以上で作る熟語に「～する」をつけて動詞を構成する「スル名詞」の３つに動詞漢字は使われる。これらの分類は、動詞漢字の読みに影響を与える。

動詞漢字の総数は約1000字である。

(3) 名詞漢字他
残りの漢字の大部分が名詞に使用する漢字である。なお、副詞、接続詞などは別に分類した。

以上の３つの分類に従って、シソーラスを以下の項目で構築した。

- 見出し語 (entry)
  漢字１字ごとに設ける。

- 下位分類
  上の３つの分類に該当するところだが、さらに漢字の下位概念をグループ化してある。形容詞漢字はイ形容詞とナ形容詞の２つに分類した。動詞漢字は、手の動作を表す動詞のグループや衣食住に関する動詞のグループといったように、同じよう

な概念を持つものをまとめて、８つのグループに分類した。これは、漢字１字が表意文字であることを利用したものであり、部首の意味を考えて、幾つかの部首ごとにまとめた。例えば、人の状態・性質・情感を表す動詞のグループには、にんべんやこころなどを部首に持つ漢字をまとめている。名詞漢字も同様に 42 のグループに分類した。ただし、動詞漢字と同じように部首ごとで分類すると偏りを生ずるので、天地・日時・人文などの分類をしている。

- 類義・対義・code・種類
  それぞれ、類義語、対義語、区点コード、漢字の学習区分を示している。漢字の学習区分を示すことによって、解析に制限を加えることが出来る。例えば、日本語学習の初級者が書いた文章を解析する際は、その初級者のレベルに応じた漢字だけを用意すればよい。この分類には、学習漢字・常用漢字などを当てた。

- 代表音・代表訓・送り品詞
  見出し語の音の中で最も使用頻度の高いものを代表音に、分類にふさわしい訓の中で最も使用頻度の高いものを代表訓とした。もし代表訓が用言ならば、送り品詞の箇所に活用ルールが示される。

- 他音・他訓
  代表音・代表訓の他にも見出し語に読み方があるのなら、この欄に示される。また、その他音・他訓で読む熟語も登録している。もし存在するならば、その送り品詞も示される。

- 訳
  その漢字１字の英訳を示している。

以下に漢字シソーラスの実際の画面を示す(図１)。

この漢字シソーラスの全体の容量は約 1M バイトであり、十分に小さいといえる。

3 語句解析への応用

この漢字シソーラスを利用して、漢字仮名混じり文の中の漢字に読みを与えるツールを構築した。

漢字の読みに音読みと訓読みが混在しているのにも関わらず、日本人が読みを間違えることはあまりなく、かなりの速度で文章を読むことが出来る。そこで、漢字の読みの選択は、一つ一つの単語から導き出されるものではなく、むしろ漢字の字形と文章中の配列から決定されるものだと我々は仮定した。

漢字に読みを与えるツールは、以下のようなルールから実現されている。

(1) 活用する動詞・形容詞の語幹に使われる１字の漢字は、訓読みに使われる。そこで、１字の漢字を抽出して、その漢字が動詞漢字・形容詞漢字であり、かつ訓読みがあるならば、訓読みにする。

(2) (1) で動詞漢字・形容詞漢字でなかった１字の漢字は、名詞として使われており、訓読みされることが多い。もし、訓読みがあれば、訓読みにする。訓読みがなければ、音読みにする。

(3) 漢字の前後が仮名でなければ、その漢字は熟語を構成している。漢字の直後の文字が平仮名であるならば、その平仮名を調べる。それが連用形の活用語尾であるならば、その直前の漢字は訓読みされる。そして、その訓読みされた漢字の直前の文字が１字の漢字であれば、それも訓読みされる。

(4) 残った漢字は音読みされる。

複数の読みがある場合、まず漢字シソーラス中の他音・他訓を調べ、該当する熟語がなければ代表音・代表訓を読みとする。

このツールによって、２万文字の漢字仮名混じりの文章に読みを与えた。その結果、84.7% の漢字に正しい読みを与えられた。

4 まとめ

漢字１字単位の読み・品詞に注目して、電子化漢字シソーラスを構築した。また、それを用いた語句解析への応用の一例として、漢字仮名混じり文に読みを与えるツールを構築した。

今回の漢字に読みを与えるツールは、湯桶読み・重箱読みといった読み方や、「売掛」といった送り仮名を伴わない複合語などには対応していない。それらが今後の課題となるだろう。また、漢字シソーラスの読みと品詞の情報しか活用しなかったので、他の情報も活用したツールの構築を目指していきたい。

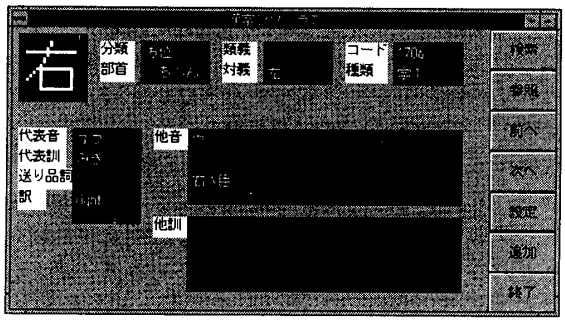

図 1: 漢字シソーラスの画面

参考文献

[1] 中挟知延子 島田静雄:「動詞辞書の提案とその利用についての一考察」 情処第 51 回全国大会,3H-4,1995

[2] Nelson: 漢英辞典 Tuttle