1M－1

# 既存のシソーラスを利用した漢字シソーラスの半自動生成

川村　和美　　宮崎　正弘
新潟大学・大学院工学研究科

## 1．まえがき

現在までに、分類語彙表[1]など、人間用のシソーラスが作成されてきた。既に電子化されているシソーラス等も少なくない。これらのシソーラスの多くは上位／下位関係を中心に作成されている。しかし、語（概念）の分類には、それ以外にも種々の観点がある。語と語の共通点を見いだしたり、それらをある観点からクラスタリングすることは自然言語理解において、非常に有効な手段である。これらを実現するために、種々の観点から語（概念）間の関係を表す、多次元のシソーラスの作成が要求される。多次元シソーラスの作成においては、語の持つ種々の観点を抽出することが重要となるが、これには、莫大な手間と時間が必要である。本稿では、観点の半自動抽出を行うために、日本語の持つ漢字という特徴に着目し、既存のシソーラスを利用して、漢字一字を基本的概念とした『漢字シソーラス』の半自動生成法について述べる。

## 2．二字漢字熟語の構成

日本語において、漢字は非常に重要な役割を占めている。漢字は表意文字であるため、造語能力に加え、伝達能力も優れている。また、文節の目安となったり、語意の把握に役立っている。

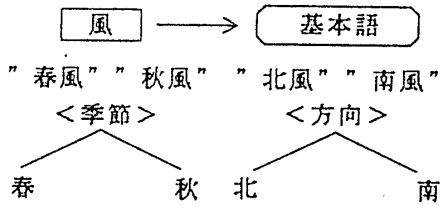
日本語の名詞は、大部分が二字の漢字で構成されている。以下にそのパターンを示す。

【名詞＋名詞】
・互いに類義／対義（例 河川／前後）
・連体修飾　　（例　春風、茶道）
【動詞＋名詞】
・連体修飾　　（例　食器、引力）
【形容詞＋名詞】
・連体修飾　　（例　強風、小川）

例のように、前の漢字が後ろの名詞性の漢字を修飾する、連体修飾関係の結合が二字の漢字熟語の約半数を占めている。[2]

「風」は一字漢字である。これが基本概念を表す基本語となる。「風」を含む二字漢字熟語”春風””秋風””北風””南風”は、複合概念を表している。「風」は’春’’秋’’北’’南’と結合することで修飾され、詳細化している。’春’’秋’からは＜季節＞、’北’’南’からは＜方向＞という上位概念が獲得され、これらが、「風」の観点と考えられる。（図1）

つまり、結合する漢字により、その語のもつ観点が抽出できる。漢字を利用した、多次元シソーラスの観点自動抽出が行える。




＜季節＞＜方向＞が「風」の観点

図1

## 3．漢字シソーラスの作成

### 3．1　一字漢字の抽出

一字漢字は既存の角川類語新辞典[3]から抽出する。このデータは名詞、動詞、形容詞と分割しやすい。また、動詞、形容詞のデータ数が比較的多く、分類の観点がかなり明示されている。

抽出条件は、名詞性の一字漢字の場合は「風」や「川」のように”単独で存在する”ことである。これに対して、動詞性、形容詞性の一字漢字の場合は、活用語尾があるため”一字目が漢字であり、かつ、二字目以降が平仮名の形で存在する”という条件で

---

A Semi-Automatic Construction of Kanji-Thesaurus from Existing Thesauruses

Kazumi KAWAMURA, Masahiro MIYAZAKI
Niigata University

抽出する。

この条件により、総数２６７０字の一字漢字を得た。結果を図２に示す。

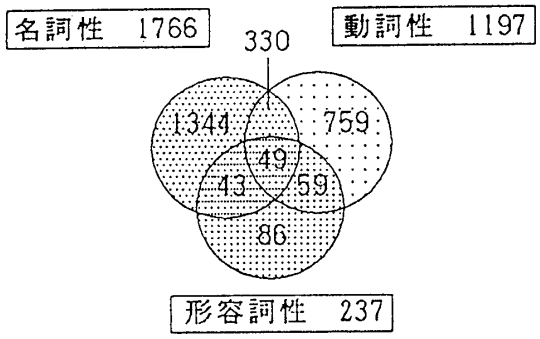



図２

３．２　漢字シソーラスの網羅性

漢字シソーラスの網羅性の点から、連用形名詞、類義漢字などの追加を行う。例えば、「川」を表す語の中には、「流」を含む二字漢字熟語が多い。しかし、”流れ”の形で存在するために一字漢字としては抽出されない。”流れ”は、動詞から名詞に転じた連用形名詞である。漢字シソーラスには、連用形名詞を追加する。また、名詞の例外である、最後の音節を送る語（例　幸せ、災い）も追加する必要がある。

前述の抽出条件では、抽出されない一字漢字が他にも存在する。「河」や「樹」のように、単独では存在しない漢字である。単独では存在しないが、他の漢字と結合する場合が多いため、追加する必要がある。これには、「河」と「川」、「樹」と「木」が類義であることを利用する。互いに類義である漢字で構成されている語を収集した類語ファイルを作成し、これを利用して類義漢字を獲得する。漢字シソーラス中の一字漢字について、類語ファイルを参照することにより、＜川＞というノードの中の一字漢字「川」からは「河」（河川より）、＜風＞というノードの中の一字漢字「風」からは「潮」（風潮より）の類義漢字が得られる。これをそのまま類義漢字としてシソーラスに追加するのは不適当である。そこで、判定を行う。判定条件は”角川類語新辞典の同ノードの中に類語ファイル中の語と同形のものがある”ことである。例えば、”河川”は、角川類語新辞典の＜川＞というノードの中に含まれているため、このノードに関しては、「河」と「川」は類義と見なすが、＜風＞というノードの中に”風潮”は含まれないため、類義とはみなさないということである。この判定条件を用いて、漢字シソーラスに類義漢字を追加する。

以上により、図３のような漢字シソーラスが作成される。

| ＜川＞ | ＜風＞ |
|---|---|
| 川 流 河 | 風 |
| \| | \| |
| 瀬 滝 淵 渚 | 嵐 颪 凪 |

図３

４．漢字の多品詞性と多義性

図２の抽出結果から、品詞の重複が非常に多いことがわかる。この事実は、漢字間の関係を得る際の問題となる。重複部分には、優先順位を付与し、品詞の絞り込みを行う。二字漢字の結合パターンから、形容詞性、動詞性の漢字による連体修飾が多いことを考慮し、優先順位を形容詞、動詞、名詞と仮定する。明らかに、この優先順位に該当しない漢字については、学習研究社漢和大辞典[4]の漢字の品詞情報を参考に変更を行った。

また、一字漢字はシソーラス上の複数のノードに分散している。このような漢字の多義性の扱いは、今後の課題の一つである。

５．あとがき

漢字に着目し、そのノード別の関係（上位、下位）を表す品詞別『漢字シソーラス』の半自動生成を行った。網羅性の点から、連用形名詞やノード別類義漢字を追加し、さらに品詞の優先順位を付与した。

今後は、多次元シソーラスの観点自動抽出、名詞の語義記述や複合名詞解析への応用を考えている。

［参考文献］

[1]国立研究所：分類語彙表　秀英出版（1965）
[2]野村：二字漢語の構造　日本語学 vol.7, no.5, pp.44-55（1988）
[3]大野、浜西：角川類語新辞典　角川書店（1981）
[4]藤堂：学研漢和大辞典　学習研究社（1978）