# シソーラス作成のための辞書関係語の抽出

1G-1 荻野 孝野 ・ 横山 晶一 ・ 荻野 綱男

(日本電子化辞書研究所*) (電子技術総合研究所) (筑波大学)

## 1. はじめに

本研究は、辞書[1]の見出し[2]と語義文中の単語との間の意味的関係をとらえ、シソーラスを作成することを目指すものである。辞書を用いてシソーラスを作成する研究は、すでに行われている[3]。これは、計算機による関係語の抽出を出発点として、自動的なシソーラス作りを行ったものである。我々の研究は、関係語の抽出および関係付け部分を人手によって行い、階層付け部分を計算機によって行った[4]。ここでは、言語データを作ることを第一義とし、基礎データをできるだけ確実なものとするため、計算機処理が有効な部分と人手による判断が有効な部分とをうまく組み合わせて処理を行おうとした。本稿では、これらの研究の概要、および手作業による関係語抽出時に生じた言語的な諸問題とその扱いについて述べる。

## 2. 研究手順

1)関係語の抽出 名詞に相当する見出しについて、所定の関係に該当する単語を拾いだして、関係の種類を記入する。

例1

```
0000810  @アウト@①〔out〕
0000820  1  (1)外(ソト)。
                =
0001020  [-プット]④〔output〕
0001030  1  (1)〔電気の〕出力。
0001040  2  (2)企業の売り出す資材。産出。
0001050  3  (3)電子計算機が処理して出す結果。
0001060  4  (4)レコードプレーヤーや録音機を拡声器につなぐ装置。+インプット。
```

2)関係語レコードの入力 関係語1語につき、見出し(子見出しも対象となる)のレコード番号と、関係語が出現したレコード番号、関係語、関係の種類からなるレコードを作成する。

例2

| 見出し行番号 | 関係語相対番号 | 関係種類 | 関係語 |
|---|---|---|---|
| 6けた | 2けた | 1けた | 71けた |
| 000102 | 01 | R | 電気 |
| 000102 | 01 | = | 出力 |
| 000102 | 02 | < | 資材 |
| 000102 | 02 | < | 産出 |
| 000102 | 02 | < | @産出物 |
| 000102 | 03 | R | 電子計算機 |
| 000102 | 03 | < | 結果 |
| 000102 | 04 | < | 装置 |

3)計算機による見出し語の抽出 関係語レコードの見出し行番号を見出しに置き換える。つまり、見出しファイル[2]に付与された見出し行番号をリンクすることによって、見出し語への置き換えができる。

例3

| | | |
|---|---|---|
| アウトプット(1) | R | 電気 |
| アウトプット(1) | = | 出力 |
| アウトプット(2) | < | 資材 |
| アウトプット(2) | < | 産出 |

4)2段階の階層関係の作成 3)の関係語レコードを計算機処理して、まず、2段階の階層関係をとらえる。3),4)の詳細は、別稿[4]に述べる。

5)文脈付関係語表の作成 階層関係の信憑性などのチェックに使用するためのものである。これは、階層関係を築きあげていく上で、疑問が生じた関係について、元データに戻って見直しができるようにしておくための資料である。

例4

| 関係語 | 関係 | 見出し | 出現文脈 |
|---|---|---|---|
| 電灯 | < | 明かり | 電灯、灯火など。 |
| 電灯 | > | アーク灯 | 向いあった二本の炭素棒に電流を通じその間に白熱した光を出させる電灯。 |
| 電灯 | > | 懐中電灯 | 乾電池を使った、携帯用の小型電灯。 |

6)全体階層(シソーラスの作成) 4)で作成した2段階の階層リストを、関係語ごとに1方向に固定し、複数の階層リストに拡大していくものである。関係語とに1方向に固定するとは、「<」の関係のものは、左右の単語を入れ換えて、「>」の関係のみとすることである。ここでは、計算機処理による階層のリンクは、人手による全体関係の構築の資料という位置付けで行う。

例5

明かり > 電灯 > (アーク灯、懐中電灯、...)

## 3. 関係語の抽出

3.1 関係の種類 関係の種類は、表1に示す12種類とし、このうち、反義関係については、語義文の記号によりまったく一義的に抽出できるので手作業対象外とし、手作業で付与する記号は、反義を除いた11種類とする。

3.2 関係語抽出時の言語的な諸問題とその取り扱い

1)関係の追加 関係語抽出作業の初めにおいては、[上下、全体部分、同義、関連]の6種類の関係で始めた。その後、[上下、関連]として扱っている関係の一部について、意味的な関係の違いを認め、[和集合、例示関係、兄弟関係]などを追加した。追加した項目の内容と、関係の境界の判断の揺れた部分について説明する。

(a)[上下関係]と[例示関係] [上下関係]とは、意味的特徴を共通にもつ部分の増減によって決まってくるもの

---

* 当所入社以前の研究を主とする。

---

Extraction of Related Words from a Japanese Dictionary for Implementation of a Thesaurus

Takano OGINO[1*], Shoichi YOKOYAMA[2], and Tsunao OGINO[3]

[1].Japan EDR Institute Ltd., 2. Electrotechnical Laboratory, 3.University of Tsukuba

ainly based on the study performed before she joined the company.

表1　関係の種類

| 関　係 | 記号 | (例)見出し | 関係 | 関係語 |
|---|---|---|---|---|
| 上下1 | ＞ | 明かり | ＞ | 電灯 |
| 上下2 | ＜ | アーク灯 | ＜ | 電灯 |
| 全体部分1 | ( | 手 | ( | 動物 |
| 全体部分2 | ) | あご | ) | 口 |
| 同義 | ＝ | アーク灯 | ＝ | アークライト |
| 反義 | ＋ | 年上 | ＋ | 年下 |
| 和集合1 | 〉 | 七道 | 〉 | 東海道 |
| 和集合2 | 〈 | 恵比須 | 〈 | 七福神 |
| 例示(値)1 | ≧ | 人名 | ≧ | 佐藤 |
| 例示(値)2 | ≦ | おはよう | ≦ | 挨拶 |
| 関連 | R | 栽培 | R | 野菜 |
| 兄弟(類義) | G | 油揚げ | G | 豆腐 |

である。共通部分を維持しながらも意味的特徴が減少していく関係（含む範囲が広がる）を上位関係、意味的特徴が付加されていく関係（含む範囲が狭まる）を下位関係とする。［例示関係］とは、上位語は単なるラベルに過ぎず、上位語の意味的特徴を引き継いでいないものとする。「呼び名」や「量」などの具体的な値などが例示関係になる。

例6　例示関係

学士　≦　称号　　九月　≦　秋

サー　≦　称号　　3歳　≦　年齢

例7　上下関係

屈折　＜　現象　　危機　＜　場合

虹　＜　現象

例示関係と上下関係の区別が難しい部分もあり、ゆれが出てきているので、最終的な調整が必要と思われる。

(b)［和集合関係］と［上下関係］　［和集合関係］とは、上下関係のうち、構成する下位語が限定されている語同士の兄弟関係をいう。「あわせよび関係」[5]と呼ぶ場合もある。

例8

七道　〉　(山陰道、山陽道、東海道、．．．)

なずな　〈　七草　　男女　〉　(男、女)

合否　〉　(合格、不合格)

(c)［兄弟関係］と［和集合関係］　［兄弟関係］とは、同じ上位語に属すると思われる下位語同士の関係をいう。

例9

挨拶　G　返答　　あす　G　今日

一見和集合関係に見えるようなものでも、「A OR B」の関係にある上位語Cと、AやBの関係は兄弟関係とした。

例10

多少　G　(多い、少ない)　　手足　G　(手、足)

「多少」、「手足」は、それぞれ「多いか少ないか」、「手や足」であって、必ずしも「A AND B」を包括する単位ではない。

(d)［全体部分関係］と［関連関係］　［全体部分関係］の範囲をどのように設定するかは抽出段階でかなり迷う部分であった。一応の目安として、部分を場所や位置として認定できる範囲を全体部分関係とした。

例11

帯芯　(　帯

うな重　R　(かば焼き、ご飯)　　掌紋　R　手のひら

(e)［関連関係］　［関連関係］とは、［上下、全体部分、同義、和集合、例示、兄弟］のいずれにも入らない意味的関係であるが、関係語として抽出しておきたいものを、とりあえず雑多な関係として抽出したものである。およその目安として、見出し語と共起しやすい用言などを重点的に抽出するようにした。用言の抽出に際しては、格関係に基づいた分析に役立つように、機能語も含め抽出した。

例12

愛嬌　R　をふりまく

愛情(1)　R　人に対して

愛情(2)　R　異性に対して

Rのついた単語関係については、今後さらに細分化の基準の決定と分析が必要である。

2)関連語の追加　ここでは、上記の(e)の扱いを除き、基本的に単語と単語のみの意味的関係を捉えることを前提としている。語義文全体で一つの概念のイメージを説明しているような場合、語義文から適切な関係語が抽出できないことがある。特に、抽象的な概念の説明的な語義文の場合には、採用できる関係語がなかったり、意味的関係が広すぎることがある。適切な関連語が抽出できない場合には、「@関連語」の形式で、抽出者が考えた関連語を付してよいものとした。特に、上位語、同義語などが一つの見出し内で皆無のような場合にはこの方法をとるようにした。これは、上下関係のリンク時において、階層関係の中断箇所を少なくするためである。上位語については、文献[6]のシソーラスの上位部分などを参考にした。

例13

アルバイト　＝　アルバイター

アルバイト　＜　@人

脂太り　＜　@状態

(語義文：からだの脂肪分が多すぎる・こと（人）)

3)同じ見出し内に同じ関連語が二つ以上ある場合

例14

メートル　①自動式の計量器。メーター。

②．．．の基本単位。メーター。

例14のように、関連語自体も多義の場合、異なる語義文に重複して出てくることになるが、それぞれ指示概念が異なるので、表面上は同じでも、別語として抽出した。これらの区別を反映するため、別稿[4]で述べる計算機処理においては、見出しに語義文番号を付加して多義語レベルの対応ができる形式としている。

## 4．おわりに

シソーラス作成の一段階として、本研究で行った関連語の抽出は、2段階の階層関係の出力結果を見る限り、かなり信頼性の高い意味的関係が付与された言語データを得ることができた。本報告で作成されたデータをもとに別稿[4]の研究も行った。また、別に作成されたシソーラス参照用エディタ[5]上での検討も行い、名詞に該当する単語の全体的な階層へと進めていきたい。本研究は、文部省特定研究「言語情報処理の高度化のための基礎的研究」の一環として行われたものである。関係語の抽出にご協力いただいた稲田順子さんに感謝します。

□参考文献　1．金田一他(編)：新明解国語辞典(第2版)、三省堂(1974)．2．横山、荻野(孝)：電総研彙報48-4(1984)．3．鶴丸、日高、吉田：情処全大27,2H-2(1983.10)．4．横山、荻野(孝)、荻野(編)：本大会予稿集(1989.10)．5．荻野(編)：文部省特定研究報告集(1989.3)．6．荻野(孝)：計量国語学16-3(1987)．