

## 用例に基づく派生語の確率的解析

市丸 夏樹\*, 中村 貞吾†, 宮本 義昭‡, 日高 達\*

\*九州大学, †九州工業大学, ‡日本ユニシス

シソーラスを意味的な分類として利用し、コーパスから収集した用例と、確率文法を用いて派生語を解析する方法と、大量データに対する仮名漢字変換実験によるその方法の評価について述べる。派生語の語幹名詞と接尾語の接続性を大量の用例データから学習し、学習した用例に含まれていない派生語についても取り扱うことを可能とした。用例そのものを学習に使用した場合には、生成能力が不十分であることが判明したため、用例の語幹をシソーラス上の上位語で置き換えた、一般化サンプルを使用した仮名漢字変換実験を行った。その結果、用例の一般化によって十分な生成能力を得ることができ、正解率を向上させることができることを示した。

## Example-Based Stochastic Analysis of Derivatives

Natsuki Ichimaru\*, Teigo Nakamura†, Yoshiaki Miyamoto‡, Toru Hitaka\*

\*Kyushu University, †Kyushu Institute of Technology, ‡Nihon Unisys

We are engaged in a research on using a thesaurus to analyze Japanese derivative words, and presented Example-Based Approach with Stochastic Grammar, which gives preference to words of high frequency. Using only raw examples for learning, however, turned out to be failure to produce enough derivative words out of given examples. Our solution to this problem was to use generalized examples for learning; i.e. to propagate the suffixes to synonyms of the body of the examples. In an experiment on kana-kanji conversion, we confirmed that there was a great improvement in number of correct answers.

## 1 はじめに

仮名漢字変換システムを使用して文章を作成していると、思いがけない漢字に変換されて驚くことがある。変換の失敗の多くは、派生語や複合語など、名詞の接続部分に起因する。とくに派生語は、助詞と紛らわしい読みをもつ接尾語が多く、また文章中に頻繁に使用されるため、変換の正解率を下げる大きな原因となっていた。そのために派生語を意味的に取り扱う手法が望まれていた。本稿では統計的な手法を用いた派生語の処理について述べる。

## 2 派生語の文法

### 2.1 言語表現としての派生語

派生語とは、単語の前または後に接辞あるいは造語成分が接続して、元の単語と若干異なった意味や品詞を持つようになった転成語である。なぜ日本語にはこのような派生語が必要であったのだろうか。その理由としては以下のような事が考えられる。

まず、新語を作るのではなく汎用性のある接辞や造語単位を単語に接続させて新しい語を派生させる仕組みを持つことは、単語の数を爆発させることなく多様な意味表現を可能にする。また、接辞や造語単位の多くは漢字1文字の表記を持っていることから、それらを語幹に接続するだけで作られる派生語というものは、複雑な修飾句の付随した名詞句と比較してはるかに簡潔な表現形態を取ることができるという利点も持っている。

従って派生語は、多様な意味表現の需要に応じ、覚えやすさと使いやすさを兼ね備えた言語表現として普及したものであろう。

### 2.2 派生語処理の問題点

一方、この派生語の持つ意味的な多様性と言語としての簡潔性が機械による処理を取り扱い難いものにしてきた。すなわち、以下のような事が問題となる。

まず第1に、接辞や造語単位の多くが非常に簡潔に表記されるため、同音異義語や「同表記異音語」さら

には「同義異語」などのように、高い曖昧性が現われがちである。第2に、派生語の語幹と接尾語や造語単位との接続は多分に慣用的なものであり、語幹の好む接尾語について明確な規則性は認められない。ただそれぞれの接尾語について、特定の意味内容を持つ語幹に選択的に接続するという漠然とした傾向が見られるのみである。従って意味分類を用いて語幹を分類する従来の手法では、解析の精度を向上させることが困難なものとなっていた。そして第3に、派生語の多様性と膨大さのために、派生語全体を網羅した辞書を作成することも難しいのである。

そのため派生語の解析や生成の処理を行うためには、語幹と接尾の妥当な接続を網羅しつつ、解候補の曖昧さを絞り込む手法が必要になる。よって我々は派生語処理の研究を行うにあたり、大量の派生語用例とシソーラスを利用して語幹と接尾<sup>1</sup>との意味的な接続性を捉える方法を採用した。

### 2.3 用例に基づく派生語処理

近年、機械可読辞書や大規模コーパスが開発されてきたことから、大量の用例を元に用例と類似した入力を処理する手法、つまり用例に基づく言語処理(Example-Based Approach)、についての研究が盛んになってきている[5]。用例に基づく手法は、言語現象の多様性や慣用性に対処する際に特に有効である。それは次のような利点のためである。

1. 用例が信頼できる確実な正解として働く。特に使用頻度の高い用例を揃えて高い優先度を与えれば、不確実な解が優先されることが少なくなり、最尤解の正解率を向上できる。
2. 用例の追加によって精度の向上を簡単に行える。
3. 用例との類似度を使って解候補に優先付けできる。

また実際、用例処理を例えば機械翻訳へ応用することによって、従来のように処理規則や文法のみ reliant 方法と比べて高精度の処理が実現できるということが実験により確認されている[5]。

<sup>1</sup>本稿では単語の後に接続する漢字1文字の接尾語と造語単位を指す。

従って、用例処理的な考え方をうければ、従来の手法では扱いきれなかった派生語も簡単に機械処理できるようになる。ここで、派生語の用例に基づく処理システムを実現するには、大量の派生語用例、それに語幹の類似性の根拠となる階層的な意味分類(シソーラス)が必要である。

## 2.4 名詞の上位下位関係

名詞の上位下位関係<sup>2</sup>は単語同士の2項関係として捉えられ、通常1つの上位語に多数の下位語が対応する。上位語はこの下位語の集合を包括する一般的な概念を表す単語である。したがって、シソーラスを細かな意味分類として捉えることもできる。

派生語を作る接尾は、特定の意味を持つ語幹に選択的に接続するものと考えられる。よって、派生語の用例から、ある接尾がシソーラス中のある単語に接続可能であることがわかった場合、その接尾はその単語の類義語や下位語にも接続できる可能性が高いものと思われる。このことから、名詞の上位下位関係を辿って派生語用例の語幹部を書き換えることにより、用例以外の派生語を派生させることができる。

## 2.5 確率文法

言語の解析に用例を用いる場合、用例そのものや用例から派生した解候補を如何に優先付けするかということが重要な問題である。我々は派生語解析に確率文法を導入することにより、用例の頻度情報に基づいた優先付けを自然な形で実現することができた。本節ではこの確率文法について概説する。

### 2.5.1 確率文脈自由文法

確率文法とは、通常の文脈自由文法の生成規則  $\alpha \rightarrow \beta$  に、適用確率  $p(\alpha \rightarrow \beta)$  を付与したものである。確率文法により生成される導出木の生起確率は、導出に使用する確率生成規則の適用確率の積で与えられる。確率文法を用いた解析は、入力に対応する導出木を求め、生起確率の降順に出力する問題となる。

<sup>2</sup>本稿では以後これを単にシソーラスと呼ぶ。

我々の研究室では、文節内の単語や付属語の接続をマルコフモデルで表わした確率文法で取り扱うことについて研究されてきた。よって派生語文法の基盤として確率文法を用いれば、文全体の解析とも親和性の良いシステムを作成できるものと期待される。

### 2.5.2 学習について

文法規則の適用確率は通常大量のサンプルデータから学習される。その計算方法としては、サンプルデータの生起確率の総和を最大にする方法、または総積を最大にする方法が知られている。

前者の方法では生成、学習の繰り返し計算が必要である。その収束計算の結果、学習サンプルは生起確率の特に大きなもの小さいものとの2極に分離されてしまいがちである。そして派生語の用例は低頻度語がほとんどを占めるため、この方法では、用例に含まれていても非常に小さな生起確率を与えられるものが生じてしまうのではないかと考えられた。よってこの総和を最大にする計算法はあまり適さないものと思われる。一方後者の総積を最大にする方法では、用例に頻度を直接的に反映した生起確率を与えることができる。さらに、生成規則の適用確率を簡単に計算できるため、巨大なシソーラスと膨大な用例を取り扱うのに適しているものと思われる。

したがって、我々は確率文法を派生語処理に応用するにあたり、サンプルデータの生起確率の総積を最大にする簡便な方法を用いている。ここでこの方法での生成規則の適用確率の計算方法を簡単に述べる [1][2]。

生起頻度情報付きのサンプルデータを  $S = \{(t_i, m_i) | i = 1, 2, \dots, M\}$  とおく。ただし  $t_i$  は与えられた文脈自由文法  $G$  における学習サンプルの導出木で、 $m_i$  はその頻度情報である。また  $M$  は異なった学習サンプルの総数である。このとき文脈自由文法  $G$  を用いて  $t_i$  を生成する際に生成規則  $\alpha \rightarrow \beta$  を使う回数を  $C(\alpha \rightarrow \beta, t_i)$  とすると、生成規則  $\alpha \rightarrow \beta$  の適用確率  $p(\alpha \rightarrow \beta)$  は、

$$p(\alpha \rightarrow \beta) = \frac{\sum_{i=1}^M m_i \cdot C(\alpha \rightarrow \beta, t_i)}{\sum_{\gamma} \sum_{i=1}^M m_i \cdot C(\alpha \rightarrow \gamma, t_i)} \quad (1)$$

と求められる。

## 2.6 確率派生語文法

確率文法に基づいて用例に類似した派生語を生成できるようにするためには、シソーラスを階層的意味分類として取り扱えなくてはならない。そこでシソーラスを確率文法で記述し、シソーラスの頂点からシソーラス中の各単語を導出した際の全導出木を学習サンプルに含めることとした。これにより、用例の語幹部を下位語で置換した派生語の生成が可能となった。

### 2.6.1 文法規則

#### 定義 2.1 (確率派生語文法)

仮名漢字変換用の派生語文法  $G$  は、 $G = \langle N, \Sigma, P_s, S \rangle$  と表される。ただし、 $N$  は非終端記号の有限集合、 $\Sigma$  は終端記号の有限集合、 $S$  は開始記号、 $N \cap \Sigma = \phi$  (空集合)、 $S \in N$  である。また、 $P_s$  は確率生成規則の有限集合であり、

$$P_s = \{A_i \xrightarrow{p_{ij}} \alpha_{ij} \mid 1 \leq i \leq m, 1 \leq j \leq n_i\}$$

ただし、 $A_i \in N, \alpha_{ij} \in (N \cup \Sigma)^*$  であり<sup>3</sup>、 $p_{ij}$  は  $A_i$  から  $\alpha_{ij}$  への書換えの生起確率で、

$$0 \leq p_{ij} \leq 1, \sum_{j=1}^{n_i} p_{ij} = 1$$

を満たす。また  $m$  は非終端記号の有限集合  $N$  の要素数、 $n_i$  は  $A_i$  を左辺に持つ生成規則の数である。

そして  $P_s$  は以下のような 5 種類の生成規則からなる。

$$N \xrightarrow{p_1} H \mid \overline{w_0} \quad (2)$$

$$H \xrightarrow{p_2} \overline{w} \mid \overline{B} \quad (3)$$

$$\overline{w} \xrightarrow{p_3} \overline{w'} \quad (4)$$

$$\overline{w} \xrightarrow{p_4} w \quad (5)$$

$$\overline{B} \xrightarrow{p_5} B \quad (6)$$

<sup>3</sup> $V$  を記号の有限集合とすると  $V^*$  は、 $V$  の要素を並べた任意の長さの記号列全体の集合であり、空列  $\epsilon$  を含む。

ただし、 $N, H, \overline{w_0}, \overline{w}, \overline{w'}, \overline{B} \in N, w, B \in \Sigma, S = N$  で、 $p_{1 \sim 5}$  は生成規則の適用確率である。

ここで  $N, H, \overline{B}$  はそれぞれ名詞、派生語、接尾語のカテゴリーであり、また  $\overline{w_0}$  はシソーラスの頂点にある単語で、 $\overline{w}$  はシソーラス上のノードや葉となっている単語を表す。ただし、 $\overline{w_0}, \overline{w}$  にはどちらも語義番号が付与されている。そして、 $w, B$  は各々  $\overline{w}, \overline{B}$  の読みである。

規則 (2),(3) はそれぞれ、名詞が派生語かシソーラス中の任意の単語を生成すること、派生語が語幹単語と接尾語を生成することを表わす。規則 (4) は上位語が下位語を生成することを表わす。そして規則 (5),(6) は、単語、接尾語が、仮名漢字変換での終端記号としての読み仮名を生成することを表している。

## 3 派生語の解析

用例に基づく言語処理の問題点として、入力と用例データベースとのマッチングにかなりの計算量が必要とされるということが挙げられてきた。シソーラスは下位語方向に大きく枝を広げた構造になっているため、上位語から下位語方向へ探索を行なったりすると、非常に多くのノードを検索しなければならなくなる。一方下位語から見て直接の上位語は高々 1~3 語程度であるため、下位語から上位語方向への探索を用いれば、探索空間を小さくすることができる。そこで我々の派生語解析の試験システムでは、用例中の接尾をシソーラスノードに属性として格納しておいて、入力に漢字表記と読み仮名が一致する語幹ノードからシソーラスを遡って、接続可能な接尾の検索を行うようにした。

### 3.1 派生語用例の一般化

第 4 節で述べるように、大量データに対する仮名漢字変換実験により、テキストコーパスから収集した用例そのものを学習に使用しても、十分な生成能力が得られないことが判明した。そこで我々はまずサンプルデータを増加させようとした。しかし、新たに用例を収集してもそのほとんどは元の用例と重複した派生語になってしまい、新規の派生語を獲得するには莫大な量のテキストが必要となることがわかった。シソーラ

スの中で類義語は、DAG 構造中の親ノードを共有する兄弟ノードに相当する。したがって、用例の語幹部を類義語で置換した派生語を導出できるようにするためには、用例の語幹部を上位語で置き換えた派生語を、仮想的なサンプルデータとして使用すれば良い。このように一般化したサンプルを用いる利点として、以下のようなことが挙げられる。

まず第1に、一般化後の仮想的なサンプルデータは、確率文法における通常の学習サンプルとして使用できる。そして第2に、一般化したサンプルデータを用いれば、類義語を検索する際に語幹部の親ノードから下位語を探索する必要がないため、計算量を増大させることなく生成能力を拡張することが可能なのである。ところで、この一般化の操作は確率文法の枠組を超えるものであるから、一般化サンプルに対して正確に頻度を設定することは難しい。しかし、後述の実験結果を見る限り、一般化サンプルの頻度を実際の用例よりも十分小さく設定しておけば実用上問題とはならないようである。

### 3.2 学習

確率派生語文法規則の学習には、シソーラスの頂点からのシソーラス中の全単語に対する導出木、派生語用例の構文木、それに一般化サンプルの構文木が学習サンプルとして用いられる。この例を表1に示す。なおこの全てに頻度情報が必要である。

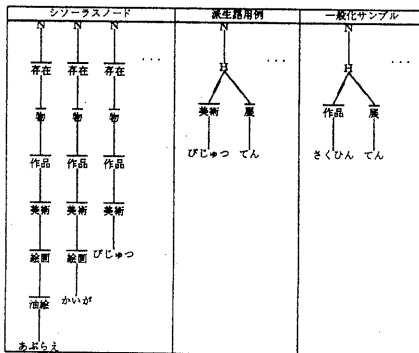


表 1: 学習サンプルの例

## 4 仮名漢字変換への応用

### 4.1 実験

名詞のシソーラスとして「現代日本語名詞シソーラス」、サンプル派生語データとして表2のような、語幹部が漢字2文字、接尾が漢字1文字からなる3漢字語データを使用して確率派生語文法を構成し、仮名漢字変換実験を行なった。

なおその際、コーパスから収集した派生語の用例には語義番号が付いていないため、用例と同じ漢字表記と読みを持つシソーラス中の語義が複数あった場合は、その全ての語義に対して頻度を等分している。また一般化サンプルの頻度としては、同じ接尾の接続した下位語のサンプルの頻度平均に、任意の接尾がある単語に接続した時にその上位語にも接続する確率約0.07を掛けた値を用いている。シソーラスの全単語の頻度としては、日経新聞コーパスから求めた、派生語頻度総和と一般の名詞の頻度総和との比、約7.78を派生語学習サンプルの頻度総和に掛けた値を、全単語数で平均した値を一律に設定している。

実験の際には3回まで一般化を行い、(0)表2の学習サンプルデータ、(1)それに一般化したデータを加えたもの、(2)1に一般化したデータを加えたもの、(3)2に一般化したデータを加えたもの、の4種類それぞれのデータについて学習を行って正解率の変動を観察した。

	派生語データ源	単語数	頻度総和	語義数
a	九大公用データベース	13786	—	—
b	広辞苑	948	—	—
c	現代用語の基礎知識 (1989年版)	9939	42800	—
d	日本経済新聞 (1982年1~3月)	4497	16299	—
0	a ∪ b (dによる頻度付き)	14733	25022	23427
1	0 ∪ {0の一般化}	—	25199	39664
2	1 ∪ {1の一般化}	—	25200	47536
3	2 ∪ {2の一般化}	—	25200	51141
A	c(語幹 ∈ シソーラス)	9190	40379	—
B	A ∩ 学習サンプル0	3580	12333	—
C	A - B	5610	28046	—

表 2: 実験に使用した用例データの一覧

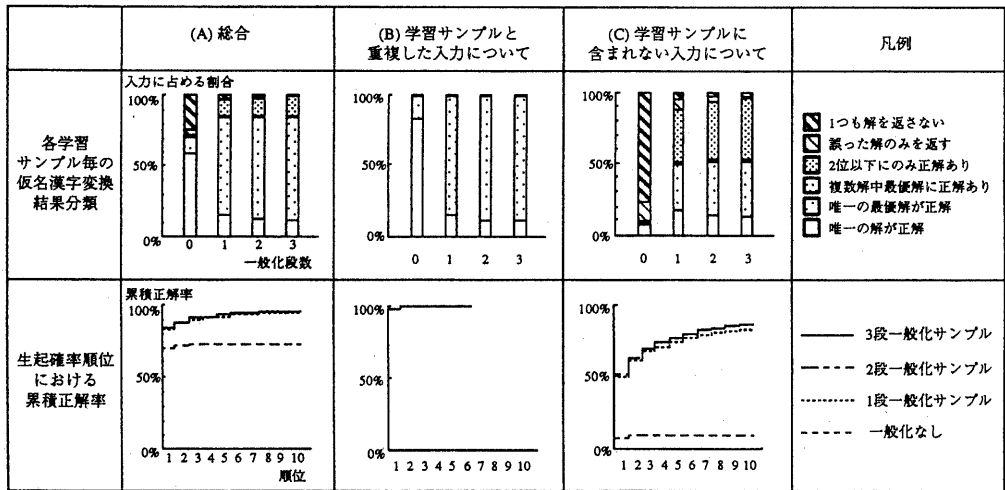


図 1: 仮名漢字変換実験結果

## 4.2 結果

仮名漢字変換の正解率と最尤解からの累積正解率を図1に示す。これは全て入力の種類を反映した確率を割り出したものである。(A)がその結果であり、(B)(C)はその内訳を、(B)入力派生語がサンプルデータに元々含まれていた語、(C)それ以外の語、に分類して集計したものである。

この結果から、一般化した用例を用いない場合、学習サンプルに含まれていなかった語のうち約80%を解析できていないことが分かる。一方、用例を一般化した場合には接続可能な語幹と接尾語の組合せをほぼ網羅できることがわかった。

また実験を行う前には、一般化を用いたことで信頼性の低い単語が優先されて累積正解率が悪化するのではないかという懸念もあった。しかし、この結果を見る限り、累積正解率は一般化につれて単調に向上していくと考えてよいものと思われる。

またそれぞれの一般化サンプルを用いて学習した文法が生成可能な派生語の総数を集計したものを表3に示す。シソーラス中の単語と接尾語の全組み合わせの数が53793168であることから考えると、特に2~3段の一般化では派生語として妥当でないものも大量

に生成している恐れが強い。それにもかかわらず図1のように高い正解率を得ることができたのは、一般化サンプルの仮想的な頻度を非常に小さく設定することによって、不確実な解の生起確率を低くすることができたためであると考えられる。

一般化段数	0	1	2	3
生成語数	217922	9852574	15277858	17522206

表 3: 一般化による生成派生語語義数の変化

## 5 考察

第4節で述べたように、一般化サンプルで学習した確率派生語文法では、十分な生成能力が得られる反面、派生語として妥当でない解候補も大量に生成してしまうことが分かった。これは仮名漢字変換の正解率をさらに向上させたい時や、翻訳などへ応用するため入力単語の語義番号まで解析したい際には大きな問題となる。そこでなぜこのように生成能力が過剰になってしまうのかを考える。

本来派生語として妥当でない解候補を生成する原因の一つは、コーパスから派生語用例を抽出する際に語義番号を設定できないことである。そのため用例の

語幹部とシソーラスノードとの対応をとって学習サンプルを作り出す際に、用例と同じ漢字表記と読みがなを持つシソーラスノード全てに用例中の接尾が接続できるものとして取り扱われてしまう。その帰着としてコーパス中の文脈における用例の意味とは掛け離れたものが学習サンプルに含まれてしまうのである。また、全ての学習サンプルに対して無差別に一般化を行っていることも、怪しげな派生語を過剰に生成する原因となっているだろう。

したがって、学習に用いる用例へ語義番号を設定したり、一般化すべきかすべきでないかの判定を行い、語幹と接尾の接続を出来る限り妥当なものに限定しなくてはならない。本節ではこれらを一括して、学習サンプルへ語義番号を設定する問題として捉え、その解決法について考察する。

## 5.1 サ変動詞の格フレームによる語義設定

荻野シソーラスには動作や行為を表す動詞性の名詞も記載されている。このような名詞は動詞としての働きを併せ持ち、隣接する単語を格支配することがある。例えば、「輸入品」という派生語を論理式で記述してみると、“「輸入する」という動詞を用いて、輸入品(y) ⇔ ∃x[輸入する(xが, yを) ∧ 品(y)]”，のように表わせるだろう。つまり、「輸入品」とは語幹部の動詞「輸入」の目的格にあたる「品」となっている。このように、語幹部分がサ変動詞の語幹部としても使用される動詞性名詞であった場合、その動詞と接尾の間に何らかの格関係が成立すれば、派生語として妥当なものとして判定して良いものと思われる。

この方法の実現にはサ変動詞語幹の格フレームを予め用意しておかなくてはならないという問題がある。現在のところ我々には大規模な格フレームが手に入らないため、この方法の有効性は未判定である。さらに、有効性が確認できたとしても、膨大な派生語のうち語幹部がサ変動詞の語幹となっているもののみしか扱えないという限界がある。そこで次に現在 CD-ROM 等で普及してきた機械可読国語辞典を利用する方法について考察してみる。

## 5.2 国語辞典の語義文と用例による語義番号の設定

国語辞典には見出し語の語義を説明した文の後に「」で括った用例が提示されている。この用例から派生語であるものを抽出すれば、語幹部である見出し語の語義文も同時に得られることになる。見出し語が名詞である場合、この語義文には上位語が特に文末に記述されていることが多い。そこで、語義文を解析して上位語を求め、用例の語幹部と同じ漢字表記と読みを持つシソーラスノードの候補のうちシソーラス中の上位語と語義文中の上位語が近くに位置しているものを選択することで、語義番号付きの学習サンプルデータを得ることが可能である。シソーラス中で見出し語の近くのものだけが選択されることから、語義文中の上位語の候補が複数得られてしまう場合にも、シソーラス中の語幹の候補選択と同時に、語義文中の上位語の候補選択が自動的に行なわれる。

例えば、三省堂の新明解国語辞典(第三版)の「機関」の項目には、“㊦ 火力・電力などを利用して機械を動かす装置。「内燃—士」”と記載されている。一方荻野シソーラスには、

- |  |
|--|
| (1) /方法/組み立て(102)/仕組み(100)/装置(100)/機関(100) |
| (2) /方法/組み立て(102)/仕組み(100)/機関(200)         |
| (3) /体系/組織(102)/団体(300)/機関(201)            |
| (4) /体系/組織(102)/団体(300)/機関(300)            |

という4種類の「機関」が存在する。このうち「機関士」という用例の語義文に含まれる名詞「装置」と意味的な距離が近いのは(1),(2)に限られるため、用例として「機関(100)-士」「機関(200)-士」を採用することになる。

しかし、この方法は一般的にはあまり有効とは思えない。なぜならば、機械可読辞書の使用には高額の著作使用料が必要となるにもかかわらず、派生語の語義文付きの用例は1冊の国語辞典に約2000程度しか含まれていないからである。つまり語義番号を設定できる用例の数は語義文の数を越えることが出来ない。従って、若干精度は劣っても、もっと大量の用例を取り扱う方法が望まれる。

### 5.3 シソーラス中の用例の近傍密度による語義推定

いくつかの接尾は特定の意味を持つ単語を語幹として好むことが多い。例えば、「丼」や「煮」が接続できる単語は食べ物が多い。よって、用例の頻度を語義毎に分割する際に、シソーラス中の近くに同じ接尾の接続可能なノードが多いものほど頻度の重みを大きくする方法が考えられる。つまり、用例の集合自体から各用例の語義番号を推定することになり、用例の頻度が語義選択時の信頼性の指標としての役割も担うことになる。この方法を用いれば、大量の用例の語義推定が可能であり、実現も比較的容易であると思われる。

今後は各語義への重みの計算方法などが課題となるだろう。

### 5.4 反例による語義の限定

「人」を表す接尾には「家<sup>か</sup>」、「人<sup>じん</sup>」、「人<sup>にん</sup>」、「者<sup>しや</sup>」、「者<sup>もの</sup>」、「民<sup>みん</sup>」等がある。そのうち、「発明」と接続するのは「家<sup>か</sup>」や「者<sup>しや</sup>」に限られる。同様に、「調子」には「者<sup>もの</sup>」、「遊牧」には「民<sup>みん</sup>」のみが接続可能である。このように語幹に、ある意味の接尾が接続される際には、似た意味の接尾のうちどれか1~2つしか用いられないようである。よって、ある意味を持つ接尾集合のうちいくつかがある語幹に接続することが分かった場合、その語幹は集合中の別の接尾とは接続しないことが多いのではないと思われる。このことから、大量の用例を基に接続しにくい語幹と接尾の組み合わせを反例として求めることができるかもしれない。ただし、こうして自動的に生成した反例から、曖昧さの絞り込みや生成能力の限定に役立つ有意義な反例がどの程度抽出できるかは今後の調査の必要がある。そのためにはまず、現在使用している748語の接尾を意味的に分類することが課題となる。

### 5.5 対話的学習

第2.3節で述べたように、用例に基づく言語処理では用例を加えることで簡単に正解率を向上させられる。よって、入力に与えられた派生語を全て用例として学

習していけば、同じ派生語がもう1度入力された際には、ダイレクトに正解を返すことができる。対話的な仮名漢字変換システムに応用する際には、このような学習機能を役立てられるだろう。

## 6 おわりに

派生語の用例と名詞のシソーラスをもとに確率派生語文法を構成する方法について述べた。また、仮名漢字変換実験を通してシソーラスを用いた派生語候補の絞り込みと、確率文法を用いた優先付けが有効に機能することを確認した。その際に用例を一般化して学習を行うことで文法の生成能力を拡張し、正解率を大幅に向上することができた。

## 謝辞

「現代日本語名詞シソーラス」を作成された、筑波大学の荻野綱男先生、「九州大学大型計算機センター公用データベース日本語単語辞書」の原データを作成された、九州芸術工科大学の稲永紘之先生、日経新聞コーパスを作成された、NTT情報通信網研究所の皆様へ深く感謝致します。

## 参考文献

- [1] 市丸夏樹, 中村貞吾, 日高達, 名詞シソーラスを用いた派生語の処理, 情報処理学会第45回全国大会講演論文集(3), pp71-72, 平成4年10月
- [2] 市丸夏樹, 中村貞吾, 日高達, 名詞シソーラスを用いた派生語の処理, 電子情報通信学会技術研究報告, [言語理解とコミュニケーション], NLC92-17, pp39-46, 平成4年10月
- [3] 杉本洋, 接辞の意味的結合性に基づく派生語文法, 九州大学大学院総合理工学研究所修士論文, 平成4年3月
- [4] 荻野綱男, 現代日本語名詞シソーラスから見た語彙の意味分類, 平成4年度科学研究費補助金 一般研究(C) 研究成果報告書, 平成5年3月
- [5] 隅田英一郎, 用例を使った言語処理, ATR Technical Report, TR-I-0374, pp33-40, 1993, (人工知能学会研究会資料 SIG-SLUD-9204-4(2/5))