PCFGによる派生語処理手法の比較と検討

市丸, 夏樹
九州大学大学院システム情報科学研究科知能システム学専攻

中村, 貞吾
九州工業大学情報工学部

日高, 達
九州大学大学院システム情報科学研究科知能システム学専攻

https://doi.org/10.15017/1500395

出版情報：九州大学大学院システム情報科学紀要. 4 (1), pp.63-68, 1999-03-26. 九州大学大学院システム情報科学研究院
バージョン：
権利関係：

九州大学大学院
システム情報科学研究科報告
第4巻　第1号　平成11年3月

Research Reports on Information Science and
Electrical Engineering of Kyushu University
Vol.4, No.1, March 1999

PCFGによる派生語処理手法の比較と検討

市丸夏樹* ・ 中村貞吾**・ 日高　達*

Comparative Study of Japanese Derivative Processing Using PCFG

Natsuki ICHIMARU, Teigo NAKAMURA and Toru HITAKA

(Received December 21, 1998)

Abstract: In Japanese language, derivative word consists of a noun and some following suffixes, and those components are concatenated into a sequence without separators. This construction often cause ambiguity in parsing or *Kana-Kanji* conversion. Some methods to treat derivatives have been developed; 1) recognizing arbitrary combination of any noun and any suffix, 2) registering collected derivative words directly into the word dictionary, and 3) using semantic category to enable selectional restriction. However, these methods have too simple mechanism to derive correct analysis. In our previous paper, we proposed 4) an example-based method in which collected sample words are generalized for wider coverage of derivative words. In this paper, we compare these methods through experiments. To realize fair comparison, all methods were represented in Probabilistic Context Free Grammar (PCFG) and equally tuned with the same training method – Maximum Likelihood Estimate. The results show that our method is superior to the methods 1) and 3), under the condition that the grammar learned more than 80,000 generalized examples.

Keywords: Thesaurus, Semantic Category, Selectional Restriction, Example-Based Method, Probabilistic Context Free Grammar, Training.

1. はじめに

完成された技術のように思われる仮名漢字変換にも，未だ問題点が残されている．その中でも名詞と短い接尾辞からなる派生語は曖昧性を生じやすく，正解を得ることが難しい．そこで我々は，高精度な派生語処理の実現をめざして，シソーラスと一般化用例による手法[3]の利用を検討している．

本稿では，「一般名詞+接尾辞」という形の派生語を解析対象とした仮名漢字変換実験を行い，我々の手法といくつかの従来手法とを比較する．この実験にあたり，確率文脈自由文法(PCFG)[7]という共通した枠組みを用意して各手法を明確に記述し，公平なチューニングを施すことを心掛けた．その結果，用例の一般化は1段程度で充分であり，我々の手法が従来手法より優れた性能を発揮するためには，少なくともものべ80,000語以上という，これまでの研究で用いられていたよりもやや多くの用例を学習する必要があることが判明した．

2. 派生語処理の様々な手法

派生語は語幹と接尾辞の連接からなる．接尾辞を固定して考えると，各接尾辞が連接できる語幹は特定の意味を持つグループ内の語であることが多い．このことから，語幹の分類を利用して，各分類内の語と接尾辞との連接性を捉える手法などが有効だと考えられる．

派生語処理手法には様々なものがあり，主に分類の細かさの違いによって次のように分けられる．

1) 任意連接　任意の語幹名詞と任意の接尾辞を受理する．語幹名詞と接尾辞を使用頻度により互いに独立に優先付けする．

2) 用例の単語登録　派生語用例を1つの単語として辞書に直接登録する．辞書に含まれている語のみを認識し，含まれていない語は一切受理しない．

3) 意味分類　全ての名詞に数10～100程度の意味分類コードを付け，各接尾辞と連接可能な語幹名詞を，予め分類コードで指定しておく手法である[4]．これにより語幹名詞と接尾辞との修飾・被修飾関係について選択制約が課せられ，解析時の曖昧性が絞られる．

4) 用例とシソーラス　意味分類よりもさらに細かく単語を分類し，階層的に細分化したものはシソーラスと呼ばれる．この手法では，接尾辞と連接可能な語幹名詞の分類をシソーラス中のノードによって表す[3]．具体的な語幹と接尾辞の連接ルールは，テキストコーパスから収集した派生語の用例を一般化(語幹名詞をシソーラス中の上位ノードで置き換えたルールを作成)することによって得られる．解析時には，用例に意味的に類似した解が優先解となる．

平成10年12月21日受付
* 知能システム学専攻
** 九州工業大学情報工学部

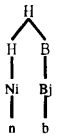
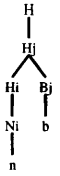
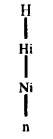
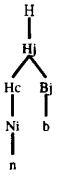
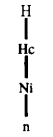
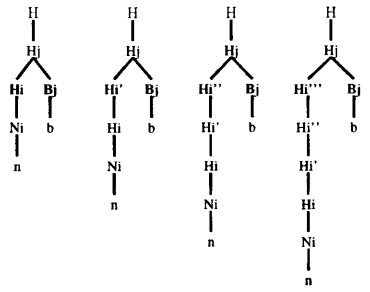
Method	Production	Sample Trees for Training	
		For Each Example "n b"	For Each Noun "n"
Arbitrary Combination	$\mathcal{H} \longrightarrow \mathcal{H}\ \mathcal{B}$ $\mathcal{H} \longrightarrow N_i$ $N_i \longrightarrow n$ $\mathcal{B} \longrightarrow B_j$ $B_j \longrightarrow b$		for any combination of n and b. 
Example without Thesaurus	$\mathcal{H} \longrightarrow H_j$ $H_j \longrightarrow H_i\ B_j$ $H_i \longrightarrow N_i$ $N_i \longrightarrow n$ $B_j \longrightarrow b$		
Semantic Category	$\mathcal{H} \longrightarrow H_c \mid H_j$ $H_j \longrightarrow H_c\ B_j$ $H_c \longrightarrow H_i \mid N_i$ $N_i \longrightarrow n$ $B_j \longrightarrow b$	if a semantic category of i is c. 	
Thesaurus and Generalized Example	$\mathcal{H} \longrightarrow H_0 \mid H_j$ $H_j \longrightarrow H_i\ B_j$ $H_{i'} \longrightarrow H_i$ $H_i \longrightarrow N_i$ $N_i \longrightarrow n$ $B_j \longrightarrow b$	k steps generalized examples. $k=0$ $k=1$ $k=2$ $k=3$ 	for each pathes. 

Fig.1 The Grammar and Trees of each Method;
Where, n, b are terminal strings. $\mathcal{H}$ is a nonterminal symbol which represents a single noun or its derivative concatenated with some suffixes. Another literal nonterminal symbol $\mathcal{B}$ represents a suffix. Nonterminal symbols N_i, H_i, $H_{i'}$, $H_{i''}$, $H_{i'''}$, H_c, H_0, H_j, B_j are subcategories of H or B with one-to-one correnspondence to thesaurus nodes or concepts. Small letters i, j of N_i, B_j coincide with identification numbers of leaf words n, b. $H_{i'}$ represents a parent node of H_i. H_0 is root node of the thesaurus. H_c is a node which correnponds to a rough semantic category.

5) その他　この他にも単語を構成する漢字を1文字ごとに分け，文字間の連接として捉える手法[1]なども研究されているが，ここでは扱わない．

上記 1) ～ 4) の手法を確率文脈自由文法(PCFG)を用いて表現した例を**Fig. 1** に示す．これらは，どれも開始記号が$\mathcal{H}$であり，語幹名詞に任意個の接尾辞が連接した形の派生語を受理するような文法である．添字が付いた非終端記号を含んだ生成規則は1つではなく，実際の文法にはこの形の規則が多数含まれる．

PCFGで記述する利点としては，

- 導出木の生起確率によって解の優先付けができ，
- その計算に必要な確率パラメータを統計的な学習によって自動的に獲得することができる，

ということが挙げられる．本稿では最尤推定法を用いて，学習サンプル構文木の生起確率の積を最大とするように学習する．各々の生成規則の適用確率値は，学習する構文木とその頻度によって決まる．

なお，形態素解析の都合上，1)，2) の構文木において，単語の概念番号を使った中間ノードを使用している．これには，表記や送り仮名の曖昧さを吸収するという利点がある．

3. 実 験 条 件

それではまず，我々が上記の手法を比較するために行った実験について述べる．

この実験の目的は，我々の手法がどの程度の量の用例

Table-1 The Total Number of Production Rules in each Grammar;
SemCat(100) means the number of semantic categories used in the grammar is 100. Thesaurus+Gen{k} means the number of generalization steps in training samples is k. Thesaurus+Gen{0,...,k} is a grammar which is trained with all samples generalized from 0 to k steps.

Grammar	Quantity of Newspaper							
	1 week	2 weeks	1 month	3 months	6 months	1 year	2 years	3 years
Arbitrary Combination	164,190	164,190	164,190	164,190	164,190	164,190	164,190	164,190
Example without Thesaurus	24,145	41,528	66,004	133,722	203,936	298,425	430,447	526,044
SemCat(100)	220,676	225,111	229,877	239,687	246,592	253,554	261,129	265,888
Thesaurus+Gen{0}	201,123	218,506	242,982	310,700	380,914	475,403	607,425	703,022
Thesaurus+Gen{1}	198,072	212,356	231,265	281,293	328,206	387,527	464,234	515,888
Thesaurus+Gen{2}	197,506	210,274	226,528	266,468	300,952	341,972	392,063	423,565
Thesaurus+Gen{3}	195,555	205,908	218,322	246,664	269,722	295,695	325,768	344,129
Thesaurus+Gen{0,1}	220,709	251,414	293,175	406,669	518,874	665,978	864,766	1,004,464
Thesaurus+Gen{0,1,2}	237,805	278,600	332,370	473,668	607,274	777,598	1,001,146	1,154,470
Thesaurus+Gen{0,1,2,3}	250,480	296,838	356,169	508,654	648,978	825,188	1,052,612	1,207,458
Total Frequency	6,911	13,905	27,659	83,728	168,392	338,762	728,536	1,124,123
Num. of Uniq. Original Examples	29,422	52,537	85,429	179,758	284,336	429,289	644,113	808,106

を学習すれば従来手法を越える性能が得られるか，また，その際用例はどの程度一般化しておけばよいか，を明らかにすることである．そのためには，用例の数を様々に変えながら，それぞれの手法を表す文法に学習させ，我々の手法においては更に一般化段数も変えて実験を行い，その結果を互いに比較することが必要である．この実験では1文単位の仮名漢字変換を想定し，仮名文字文の形態素解析において生じる全ての派生語候補に対して，それぞれの手法を表す文法による構文解析を行なう．

3.1 使用データ

実験には次のようなデータを使用した．A, B がシソーラスと単語辞書，C, D, E はテキストコーパスである．

A. EDR概念辞書 概念体系[8]**:**
上位-下位関係数 409,717．ノード数は 402,027.

B. EDR日本語単語辞書: 単語数 395,013 語.

C. CD-毎日新聞 '91〜'94[9]**:**
'91〜'94年の毎日新聞記事全文.

D. RWC-DB-TEXT-95-1[2]**:** データCを形態素解析したもの．原文表記に品詞情報がついた形態素列.

E. RWC-DB-TEXT-95-2:
CD-毎日新聞 '94 から3000記事を選び，形態素解析結果を人手でチェックしたもの．読み仮名つき.

3.2 学習サンプルの作成

コーパスデータD中の '91〜'93 分の形態素から一般名詞と接尾辞の連接を抽出して，文字列 "$n\ b$" で表わされる派生語用例を作成した．その際，このデータをシソーラスAと対応付けるため，単語辞書Bから可能性のある全ての単語の候補を割り出し，読みと概念番号の情報を補った．ここで，n, b の概念番号を各々i, jとおく．なお，語幹名詞と接尾辞のいずれかがデータBに含まれないものは除外し，曖昧性のある場合は頻度を等分して全ての候補を使用した．こうして作成された派生語用例は，新聞記事3年分の場合でのべ 1,124,123 語に達した．

この用例をPCFGに学習させるために，それぞれ **Fig. 1** の中央に示したような形の構文木に変換した．その際，n, bの概念番号i, jを非終端記号の添字に割り当てた．

1) 任意連接 全ての名詞と接尾辞を組み合わせた構文木を等しい小さな初期頻度で学習させた上に，用例の構文木を学習させることによって用例中に出現した単語を優先するようバイアスをかけた.

2) 用例の単語登録 単純に用例を構文木に変換した.

3) 意味分類 意味分類コードは本来人手で設定されることが多いものであるが，ここでは簡単のため，シソーラスAから子孫数の多い順に100ノードを取り出し，それらを意味分類コードを表す非終端記号とみなすことにした．各用例から構文木を作成する際には，N_iの先祖ノードのうち，この上位100ノード中で最も深い位置にあるものを，N_iの意味分類コードH_cとして採用した．また，意味分類中の全ての単語を導出するために，意味分類に対応するノードからその全ての子孫を導出する構文木も用意した.

4) 用例とシソーラス この手法では，N_iからシソーラスA中を何段上位へ辿ったノードを語幹の分類H_iとするかによって，いくつもの構文木が生じる．このようなルールの抽象化を用例の一般化と呼び，サンプル構文木の導出過程において生成規則 $H_{i'} \longrightarrow H_i$ が使用される回数を一般化段数と呼ぶ．ここでは，一般化段数による違いを調べるため，0〜3段のうち1つの段数を選んで学習させたり，あるいはそれらを組み合わせ，$0, 1, \cdots, k$段を累積して学習させた場合について調べた．また，この文法では語幹部でシソーラスを辿った導出が必要なため，シソーラスの

ルートから全名詞への導出パスを表す構文木を微少頻度で学習させた．

学習サンプルの作成において，複数の構文木が得られる場合は頻度を等分して全ての候補を使用した．また，構文木数が非常に多くなる部分は，等価な適用確率値を直接計算によって求めた．

こうして作成した学習サンプルは，用例の量を1週間分～3年分の範囲で分けて8種類，文法を手法の違いと一般化段数の違いによって10種類で，合計80通りとなった(**Table–1**)．これらの学習サンプルをPCFGに学習させることによって，それぞれの手法を表す確率派生語文法が出来上がるわけである．

3.3 評価基準

良い派生語文法が満たすべき条件として，

1. 正しい派生語候補を受理し，高い優先順位を与えること，
2. 派生語でない候補を受理しないこと，

が挙げられる．ここでは，これらを評価する指標として情報認識や検索の分野で用いられる適合率(Precision)と再現率(Recall)[4]を使用することにした．

ここで用いられる試験入力は，正解が存在する H 個の正入力と，本来派生語ではない負入力からなる．一方これを解析した解析結果は，上位n位までに正解が出力されるもの C_n 個と，n位までは不正解のみが出力されるもの E_n 個と，全く出力がないものの3種類に分けられる．

これらそれぞれの頻度の総和を計算することにより，次のようにしてn位までの適合率と再現率を求めることができる．

$$\overset{Precision}{\text{適合率}}\ [n] = \frac{C_n}{C_n + E_n},$$

$$\overset{Recall}{\text{再現率}}\ [n] = \frac{C_n}{H}.$$

適合率は1文単位の解析において出力に含まれるノイズの少なさを表し，再現率は派生語を区切って入力したときにもれなくn回の変換で正解する割合を表すことになる．

3.4 試験入力の作成

適合率を求めるためには，派生語のようで派生語でない，負入力を与えた実験が必要である．ここでは実際にテキストコーパスを1文単位で形態素解析し，その際に生じる派生語候補を試験入力として，その中の誤り，すなわち正解に含まれないものを負入力として用いる．

試験入力の作成時には，まずデータEの1/5から仮名表記のみを抽出して連結した仮名表記文と，それに対応する正解を作成した．次に，その仮名表記文を文節数最小法で形態素解析し，派生語として認識される可能性のある全ての一般名詞と接尾辞の候補を取り出した．

形態素解析時には，辞書B付属の品詞間接続表(+α) を使用し，その際，形態素解析結果にセグメンテーションなどの曖昧さがあれば可能性のある候補を全て含め，「か」「や」など助詞と紛らわしい接尾辞を落さないように留意した．

Table-2 Total Frequency of Test Input:

Quantity of Samples for Learning	True Derivatives in Key to The Test		Candidates Not in The Key
	Learned	Unknown	
1 week	697	1,351	52,262
2 weeks	843	1,205	52,262
1 month	999	1,049	52,262
3 monthes	1,131	917	52,262
6 monthes	1,199	849	52,262
1 year	1,234	814	52,262
2 years	1,269	779	52,262
3 years	1,280	768	52,262

試験入力となった派生語候補の文字列が，学習サンプルや正解に含まれる数を **Table–2** に示す．このように学習サンプルが増加するに従って，学習済み派生語の割合が増加していった．

また，実験の簡素化のため，派生語候補の読みの長さを5文字以上に限定し，負入力データから500語をランダムに選択して実験して結果を 52262/500 倍して集計するものとした．従って，実際に使用した試験入力数はのべ 2,548 語であった．

3.5 実験手順

以上の学習サンプルの各々について，共通した試験入力を与えて次のような実験を行った．

1. 文法別に学習サンプルを与え，確率パラメータのトレーニングを行った．具体的には各生成規則とその左辺の非終端記号の使用回数をカウントした．
2. PCFGパーザを用いて試験入力データを構文解析し，生起確率の高い順にn位までの導出木を求めた．
3. 構文解析結果を正解と比較し，適合率と再現率を求めた．

なお，条件によっては生成規則数が 1,000,000 を越える場合があるため，構文解析には生成規則数を制限しないPCFGパーザを作成して使用した．

4. 結 果 と 考 察

4.1 手法間の比較

順位 $n = 1$ すなわち最尤解について，適合率をx軸，再現率をy軸上の値としてプロットしたグラフを**Fig. 2** に示す．このグラフでは，右上に近い点ほど適合率・再現率ともに良い結果であると言える．

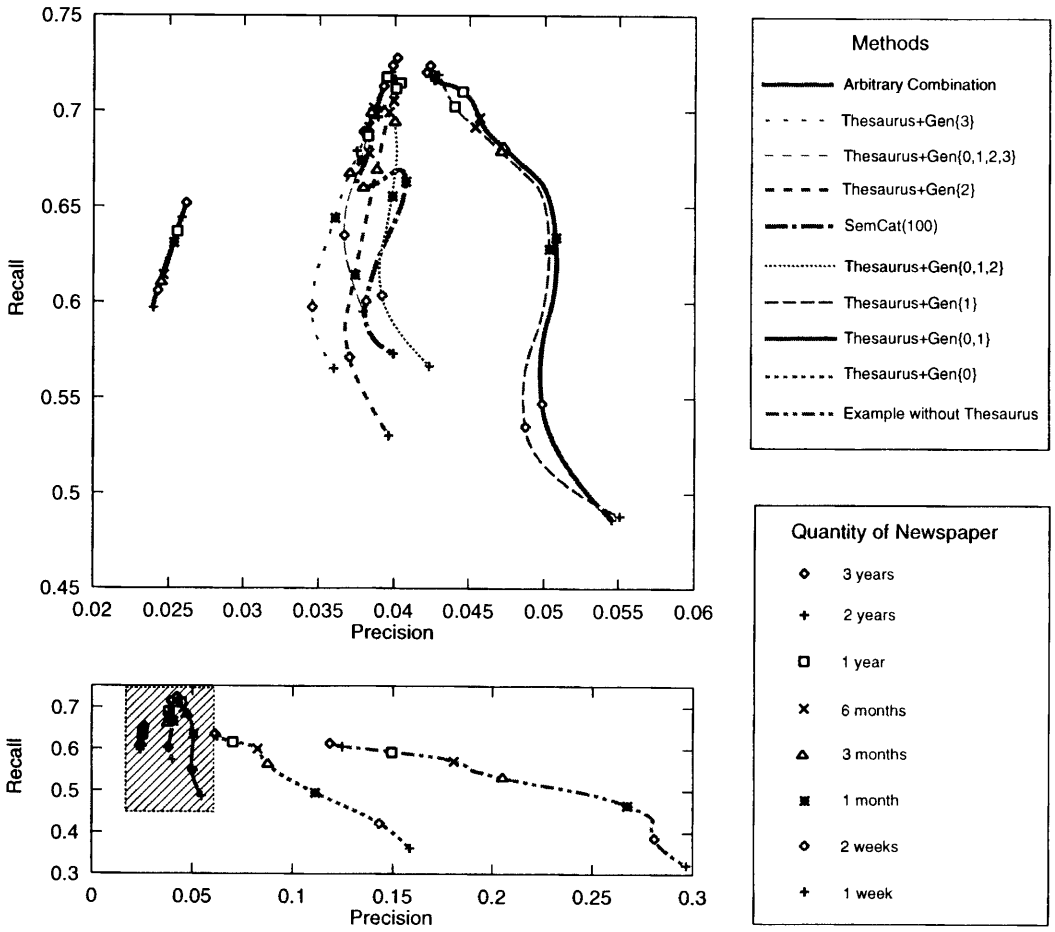


Fig.2 The Results of Experiments on *Kana-Kanji* Conversion of Japanese Derivative Words; **Training sample :** The results of morphological analysis of Mainichi Shimbun 1991 - 1993 (RWC-DB-TEXT-95-1). **Test input :** 2548 *kana* candidates found in morphologic analysis of 5000 sentences in 3000 articles from Mainichi Shimbun 1994. **Key to the test :** Manually post-edited results of morphological analysis of 3000 articles from Mainichi Shimbun 1994 (RWC-DB-TEXT-95-2). **Thesaurus :** EDR concept dictionary and EDR Japanese word dictionary (version 1.5).

4.1.1 適合率の比較

各手法間で適合率を比較すると，分類の数と同じく，

$$\text{任意連接} < \text{意味分類} \leq \begin{matrix}\text{一般化用例と}\\ \text{シソーラス}\end{matrix} < \text{単語登録}$$

という順序になっている．

新聞3年分の用例を学習した段階で，意味分類と2～3段一般化した本手法がほぼ同等である．1段一般化は全域に渡ってそれらより優れている．

意味分類の場合の適合率が1ヶ月～3ヶ月間で悪化しているのは，解の優先付け機能が弱いため，正解が2位以降に落ちてしまうことが原因のようである．また，適合率の値が全体的に大変低いのは，派生語と非派生語を選り分けることが本質的に難しいことを示している．

4.1.2 再現率の比較

再現率は**Fig. 3**のように学習用例数の増加に伴い，任意連接を除いておおむね上昇した．

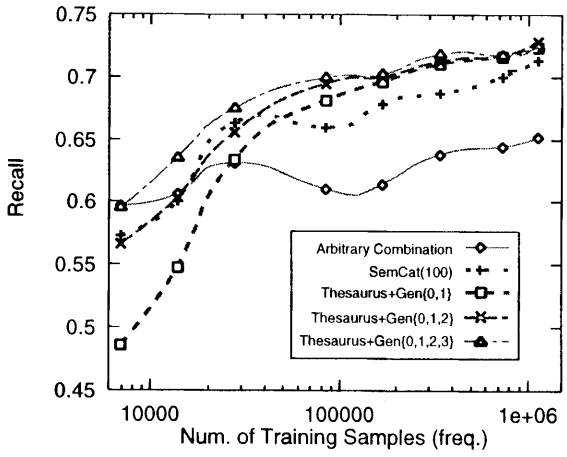


Fig.3 Maximum of Recall

この中で再現率が最も優れているのは，一般化用例を0～3段に累積した場合である．次点は学習用例数によって

入れ替わり，約 20,000 語までは任意連接，約 30,000 語付近ではカテゴリ数 100 項目の意味分類であり，約 80,000 語以降は0～2段の一般化用例を累積して学習した場合となった．

以上のように，適合率の優劣は分類の細かさを選んだ時点でほぼ決まり，再現率は学習する用例数によって変動する．

4.2 学習用例数と一般化段数の関係

シソーラスと用例に基づく手法において，最適な一般化段数は学習用例数によって変わる．適合率は前述のように一般化段数が少ないほど良い．一方再現率は，単一の段数の一般化用例を学習した場合には，始めは優れた順に 3段, 2段, 1段 の順であったものが，約100万語の用例を学習した段階では逆順に入れ替わっている(**Fig. 4**)．これは，十分に学習すれば適合率と再現率を両立できるということを示している．

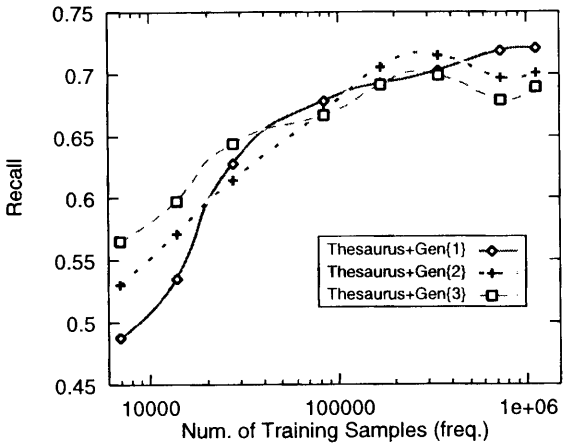


Fig.4　Recall with Generalized Examples

具体的な用例の一般化の段数としては，EDR概念体系では1段一般化用例の使用が最も適切だと思われる．ただし，この場合100項目の意味分類より高い再現率を得るためには，のべ80,000語程度以上の用例を学習させる必要がある．

一方，最も高い再現率が得られる0～3段の累積一般化では，適合率が意味分類よりも低くなることがあり，時間的効率も悪化してしまう．また，適合率が最も高い単語登録や0段一般化では，今後再現率が1段一般化を越えるまでに膨大な量の用例が必要とされる．よって一般化は1段程度が現実的である．

一般化サンプルを累積して学習した場合は，1つの段数のものを学習した場合よりも適合率・再現率共に若干良い結果が得られた．これは，様々な一般化段数のサンプルを学習したことによって，シソーラス中で用例に近い順，すなわち異表記，同義語，類義語の順に徐々に生起確率が落ちていくように生成規則の適用確率が設定され，優先付けの精度が向上したためだと考えられる．

5. おわりに

様々な派生語の処理手法を PCFG を用いて表現し，大量データによる仮名漢字変換問題についての比較実験を行った．この実験によって，用例数，分類カテゴリ数，優先付け手法が異なった様々な学習条件下における解析性能の違いが明らかになった．

各手法の性質をまとめると，次のようになる．

1) **任意連接**　最も多くの解候補が得られる．
2) **用例の単語登録**　不正解の出力数が最も少ない．
3) **意味分類**　実現が容易で高速．しかし複数解の優先付けの機能は弱い．
4) **用例とシソーラス**　充分な量の用例を学習すれば，適合率と再現率を両立した高精度な解析が可能である．一般化段数を調節することによって適合率と再現率のバランスを微妙にコントロールすることができる．実験によれば，最尤解の適合率と再現率の両方を従来手法である意味分類よりも高めるためには，一般化段数を1段に押さえて，少なくともものべ 80,000 語程度以上の用例を学習することが必要である．

適用対象に応じて，高速性が要求される対話的処理には 2) または 3)，精度が要求されるバッチ式処理には 4) というように，手法を選択して利用するとよい．

今後は用例辞書を強化し，さらに係り受けや共起関係など，派生語外部からの制約を導入することによって，適合率を改善することが望まれる．

参考文献

1) 稲永紘之; 新谷隆之: 意味的なつながりを考慮した接尾語辞書の作成について, 情報処理学会自然言語処理研究会研究報告 86-NL57-3, Vol. 86, No. 60, pp. 1–7, 1986.
2) 技術研究組合新情報処理開発機構: RWC テキストデータベース 第1版, CDROM, 1996.
3) 市丸夏樹; 中村貞吾; 宮本義昭; 日高達: シソーラスと確率文法による派生語解析, 情報処理学会論文誌, Vol. 36, No. 4, pp. 849–858, 1995.
4) 長尾真（編）: 自然言語処理, 岩波講座 ソフトウェア科学 15. 岩波書店, 4月 1996.
5) 徳永健伸: RWC テキストデータベースについて, 自然言語処理におけるコーパスの利用, pp. 1–17. 電子情報通信学会情報・システムソサイエティ 言語理解とコミュニケーション研究会, 1996.
6) 楠本達治: 日本語文の形態素解析における接辞処理, 九州大学工学部情報工学科卒業論文, 1984.
7) 日高達: 確率文法, 情報処理, Vol. 36, No. 2, pp. 169–176, 1995.
8) 日本電子化辞書研究所: EDR 電子化辞書 version 1.5, CDROM, 1993.
9) 毎日新聞社: CD-毎日新聞 '91–'94, CDROM, 1991–1994.