出現状況の包含関係による語彙の階層構造の構築

山本英子[†] 神崎享子[†] 井佐原均[†]

本論文では，コーパスから語彙の階層構造を自動構築するために，コーパス中の出現状況の包含関係に基づく手法を提案する．本研究は，ユーザが扱うデータからユーザの思考に合ったシソーラス構築を目的とし，データに依存した人間の直感にも沿った語彙の意味的な階層構造を抽出する手法を検討している．そこで，対象とするコーパスにおける語彙間の出現状況の包含関係をその語彙間の意味的階層関係としてとらえ，その関係から語彙の階層構造を構築することを試みる．本研究では，出現状況の包含関係を測定するために，補完類似度を適用した．補完類似度は文字認識の分野で劣化印刷文字を認識するために開発され，ベクトルで表現された文字画像とテンプレート文字との重なり度合いを測る尺度である．実験では，多種のテキストコーパスから共起関係にある抽象名詞と形容詞・形容動詞を収集した言語データから抽象名詞の階層構造を構築することを試みた．この実験によって，語彙の階層構造を構築する問題における補完類似度に基づく提案手法の適用可能性を検証する．そして，構築された階層構造の EDR 電子化辞書との一致度を測定した結果，人間の直感に近く，コーパスに依存した階層構造が得られたことを報告する．また，共起頻度を考慮することの有効性を検証した．その結果，より人間の直感に近い階層構造を得ることができたことを報告する．

Hierarchical Word Structure Construction Based on the Inclusion of Appearance Patterns

Eiko Yamamoto,[†] Kyoko Kanzaki[†] and Hitoshi Isahara[†]

In this paper, we propose a method of automatically extracting word hierarchies based on the inclusion relation of appearance patterns from corpora. We applied the complementary similarity measure (CSM) to measure the inclusion relation. CSM is a similarity measure developed for the recognition of degraded machine-printed text in the field. As the initial task, we attempted to extract hierarchies of abstract nouns co-occurring with adjectives in Japanese. Then, we evaluated the extracted hierarchies by measuring the degree to which they agreed with the EDR electronic dictionary. As the result, we found that our CSM-based method could extract the hierarchies which were more in accordance with human intuition and depended to the corpora. We also verified the effectiveness which the co-occurrence frequency information for co-occurring words are considered to extract word hierarchies.

1. はじめに

語彙の階層構造は言語資料として有用であり，さまざまな応用に用いられるが，語彙の階層構造は唯一のものではなく，分野や利用目的によって異なると考えられる．たとえば，医療分野の文書内での語彙の階層構造は，新聞記事全体といった一般的な文書内での階層構造とは異なるであろう．また，利用目的によっては，切り出される語の長短に差があり，階層を構成する語の集合が異なる場合もある．したがって，ユーザが対象とする分野および目的に合致した階層構造を作成し，利用することが好ましい．

これまでにさまざまな観点から階層構造を含むシソーラスが構築され，公表されているが，それらは分野や利用目的を限定しないものであり，辞書編集者に大きく依存しているため，ユーザの特定の目的と一致しない場合もある．一方，これらのシソーラスでは，語彙のカテゴリ化が人手で行われ，人間の直感に基づいて，語彙は分類される．この方法は語彙データベースを作成する良い手法であるが，高いコストがかかる．それぞれのユーザが，多くの人手を投入して，自分のためのシソーラスを構築することは現実的ではない．

ユーザはその目的に応じて，対象文書群を選択し，形態素解析などの前処理をする．このような対象文書の選択と前処理の選択がユーザの（たとえば，医療分野での翻訳といった）目的を示しており，必要とされる階層構造の特徴は，この前処理済の対象文書群に含

† 独立行政法人情報通信研究機構
National Institute of Information and Communications Technology

まれている．このため，本研究では，ユーザが自分が利用しようと思うコーパスから語彙の階層構造を自動抽出できる手法を提案する．我々はコーパスからシソーラスを自動的に抽出するので，得られる概念階層は入力となるコーパスによって異なる．もしユーザのコーパスが特定分野に限った特殊なコーパスであるならば，抽出される語彙のシソーラスは特殊なものとなる．さらに，既存のシソーラスにおいては，上下関係に加えて，同義語や類義語が列挙されているが，これらの間にある意味的または統計的な関係が明記されていないことがある．本論文では，語彙階層における単語間の意味的な関係を数値で表すことを試みている．

これまでに，階層関係を自動抽出する手法として，辞書にある定義文を利用する手法[20),22)]や，コーパスから語彙の意味関係を確率モデルによって得る手法[10)]，辞書とコーパスから得られる単語の依存関係の両方を利用する手法も提案されている[13)]．また，既存のシソーラスと文字情報，コーパス中の共起情報を利用し，現シソーラスに未登録語を追加していく手法[15)]もある．また，英語を対象として，コーパスから “a part of” や “is-a”，または “and” などを含むパターンを利用して抽出する手法がよく知られている[2),3)]．日本語においても同様に，「の」，「などの」，「という」などを含むパターンが利用される[1)]．本論文では，階層中で上下関係にある語に共起する語の集合は包含関係にあるという考えを用い，コーパス中の語の出現状況の包含関係を利用する手法を提案する．

本論文での実験においては，分野を特化せず，一般的なテキストデータである新聞記事データを使用した．これは，本論文の目的が，提案手法が語彙の階層構造をコーパスから自動抽出できるという能力を測り，適用性を示すことにあること，また，抽出される階層が一般的な階層であるため，正解データとして，人手によって人間の直感に基づいて構築された既存の一般的なシソーラスである EDR 電子化辞書が利用できることによる．なお，分野を医療分野に特化しての実験については，文献 24) に示されている．また，対象を特化した語彙関係の抽出の有効性は，文献 11) において，同一文書の翻訳を元に語の分布を獲得し，対訳のアラインメントを行った例として示されている．

実験では，新聞記事から修飾関係に基づく抽象名詞と形容詞・形容動詞の共起データを取り上げ，抽象名詞の階層構造の構築を試みる．抽象名詞は概念を抽象的に表す語彙であるため，概念を表す際に用いられる語彙，つまり，概念を代表する語彙である．これまでに，英語を対象とした修飾語と被修飾語との共起情報を利用した同義語抽出が行われている[5),17)]．これに対して，我々は抽象名詞と，それを修飾する形容詞・形容動詞との共起情報を利用して，抽象名詞の階層構造を抽出する．実験の内容は，抽象名詞ごとに形容詞・形容動詞との共起状況をベクトルで表現し，そのベクトルの包含関係を測ることによって，2 つの抽象名詞間の階層関係を推定し，その関係を連結することによって，語彙の階層構造を構築する．そこで，本論文では，2 つのベクトル間の包含関係を測ることのできる指標として，補完類似度とオーバラップ相関係数をそれぞれ適用することにした．そして，得られた階層構造を EDR 電子化辞書と比較評価する．補完類似度には二値画像のための補完類似度と多値画像のための補完類似度がある．二値画像のための補完類似度は二値ベクトルしか扱えないが，多値画像のための補完類似度は，多値ベクトルが扱える．そこで，そのベクトル要素として，形容詞・形容動詞と共起関係の強さを表す共起頻度に基づく重みを用いることによって，語彙の階層化への頻度情報の影響を考察する．

本論文の構成は以下のとおりである．2 章では，実験に使用した抽象名詞に関する言語データについて説明し，3 章では，2 語間の階層関係を推定するために検討した手法を順に示す．4 章では，階層構造を構築する手法を示し，5 章に，実験の概要と結果，EDR 電子化辞書との比較方法を示す．6 章では，オーバラップ相関係数と補完類似度との実験結果を比較し，7 章で，共起頻度を考慮することによる階層構造の構築に与える影響を考察する．最後にまとめる．

2. 抽象名詞に関する言語データ

本論文では，提案する手法が階層構造を構築する問題に適用可能であるかどうかを示すため，抽象名詞の分類を目的として作成さた言語コーパスから，抽象名詞の階層構造を構築することを試みる．対象とする抽象名詞は，抽象名詞の分類を目的として，文献 16), 19), 21) に基づき，形容詞・形容動詞の上位語として定義された抽象名詞[8),9)]である．言語データは，2 年分の毎日新聞に含まれる抽象名詞を対象とする抽象名詞として，100 の小説，100 のエッセイ，11 年分の毎日新聞，10 年分の日本経済新聞，7 年分の産業金融流通新聞，14 年分の読売新聞を KNP によって構文解析し，構文解析できた文から，その抽象名詞に前接する形容詞・形容動詞を収集することで作成された．この言語データには，抽象名詞 354 種類，形容詞・形容動詞 6,407 種類が含まれる．この言語データの一部を**図 1** に示す．形容詞・形容動詞に続く数字は前行の抽

思い

つらい 1,269 悔しい 1,236 苦い 704 寂しい 612 狭い 526 嫌な 466 悲しい 396 切実な 388 苦しい 354 不愉快な 327 恥ずかしい 305 やりきれない 283 歯がゆい 252 不快な 245 … … 楽しい 56 不自由な 56 寒い 52 不便な 47 せつない 45 … …

観点

技術的な 38 大局的な 37 経済的な 33 広域的な 31 多角的な 28 戦略的な 21 社会的な 19 歴史的な 1 … … 医学的な 9 法的な 9 基本的な 8 経営的な 8 … …

気持ち

素直な 406 不安な 402 優しい 325 前向きな 232 悲しい 210 つらい 208 不思議な 205 うれしい 204 幸せな 195 寂しい 190 やさしい 145 すがすがしい 134 残念な 120 … … 95 悔しい 93 気楽な 88 気軽な 88 楽しい 87 情けない 82 … …

図 1 実験データの一部
Fig. 1 Samples of experimental data.

象名詞との共起頻度を表している．

この抽象名詞を分類するために作成された言語データは，語彙が 354 個の抽象名詞に限定されているため，扱いやすく，提案手法の適用可能性を示すには妥当であると考えた．実験では，この言語データを抽象名詞の形容詞・形容動詞との共起に関する出現状況として扱う．そして，その出現状況から抽象名詞間の関係を推定し，その関係から階層構造を構築する．

3. 検 討 手 法

提案する手法は，語彙の階層構造をコーパス中の語彙の出現状況に基づいて構築する手法である．具体的には，2 語間の階層関係を決定し，その関係を連結していくことによって，階層構造を構築する．

階層構造の中で上位語は下位語より抽象的な，すなわち，より広い意味を持つ．このため，下位語を修飾しうる語は一般に上位語を修飾することができる．いい換えると，上位語を修飾する語の集合は，下位語を修飾する語の集合を包含する傾向にある．そこで，我々は対象となる語彙についてコーパス中での出現状況をベクトル化し，そのベクトルの包含関係を測ることによって，2 語間の階層関係を決定できると仮定した．

提案手法は大きく分けて，2 つの工程からなる．第 1 の工程では，コーパスにおける 2 語間の階層関係を統計的指標によって推定する．そして，第 2 の工程で，その 2 語間の関係のリストから語彙の階層構造を構築する．本論文では，2 つのベクトル間の重なり度合いを測ることができる補完類似度とオーバラップ相関係数を第 1 の工程にそれぞれ適用することを考えた．また，補完類似度には二値画像のための補完類似度と多値画像のための補完類似度の 2 つがある．そこで，オーバラップ相関係数とこれらの 2 種類の補完類似度

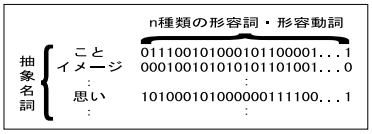


図 2 出現パターンの二値ベクトルでの表現
Fig. 2 Expression of appearance pattern by binary vector.

を単語間の関係の推定にそれぞれ適用し，比較評価する．本実験では，関係推定のために用いる情報はコーパス中の抽象名詞と形容詞・形容動詞とが共起するかどうかという情報のみを用いる．具体的には，抽象名詞ごとに形容詞・形容動詞との共起状況のパターンをベクトルで表現し，そのベクトルの重なり度合いを測ることによって，2 つの抽象名詞間の階層関係を推定する．**図 2** に共起状況を二値ベクトル化したイメージを示す．本研究で用いるベクトルの次元数 n は形容詞・形容動詞の種類数に相当する．そして，二値ベクトルでその出現パターンを表した場合，各要素は，抽象名詞が i 番目の形容詞・形容動詞と共起するならば 1，共起しなければ 0 に相当する．

3.1 オーバラップ相関係数

オーバラップ相関係数（overlap coefficient: OVLP）は二値ベクトル間の類似度を測る尺度の 1 つであり，情報検索に使われる尺度である[12]．この尺度は包含関係を測ることのできる特徴を持っている．つまり，2 つのベクトル間で共通して 1 の要素を持つ次元について，一方のベクトルがその共通する次元以外すべて 0 の要素であれば，値は 1.0 となる．これをいい換えると，他方のベクトルがそのベクトルを完全に包含するということを表す．ベクトル $\vec{F} = (f_1, f_2, ..., f_i, ..., f_n)$ と $\vec{T} = (t_1, t_2, ..., t_i, ..., t_n)$ （$f_i, t_i = 0$ または 1）におけるオーバラップ相関係数は次のように定義される．

$$OVLP(\vec{F},\vec{T}) = \frac{\left|\vec{F}\cap\vec{T}\right|}{MIN(\left|\vec{F}\right|,\left|\vec{T}\right|)} = \frac{a}{MIN(a+b,a+c)}$$

$$a=\sum_{i=1}^{n} f_i\cdot t_i,\quad b=\sum_{i=1}^{n} f_i\cdot(1-t_i),$$

$$c=\sum_{i=1}^{n}(1-f_i)\cdot t_i.$$

ここで，ベクトル $\vec{F}$，$\vec{T}$ は図 2 に示す，抽象名詞がどの形容詞・形容動詞と共起出現し，どれと共起出現しないのかを表す出現パターンに相当し，次元数 n は形容詞・形容動詞の種類数に相当する．したがって，パラメータ a はベクトル $\vec{F}$ を持つ抽象名詞とベクトル $\vec{T}$ を持つ抽象名詞の双方と共起する形容詞・形容動詞の種類数，b はベクトル $\vec{F}$ を持つ抽象名詞とは共起するが，他方とは共起しない形容詞・形容動詞の種類数，c はベクトル $\vec{T}$ を持つ抽象名詞とは共起するが，他方とは共起しない形容詞・形容動詞の種類数に相当する．

3.2 二値画像のための補完類似度

二値画像のための補完類似度（Complementary Similarity Measure for binary images: CSM-b）は劣化印刷文字を認識するために提案された類似尺度である[6]．この補完類似度はテンプレート文字と印刷文字を二値ベクトルで表し，印刷文字のベクトルがテンプレート文字のベクトルをどの程度包含するか，包含関係を測る尺度である．これまでに，この尺度の特徴を生かし，各語彙の出現状況をベクトル化し，コーパス中の 1 対多関係を推定する問題に適用されている[23]．ここでは，階層関係にある語彙対において，上位語である語彙は広義語であるため，下位語である狭義語よりも頻繁に用いられる傾向にあることに着目した．実際に，語彙対を出現状況で比較すると，完全ではないかもしれないが，重なる（包含する）状況が観察できる．ベクトル $\vec{F}=(f_1,f_2,...,f_i,...,f_n)$ と $\vec{T}=(t_1,t_2,...,t_i,...,t_n)$（$f_i,t_i=0$ または 1）における補完類似度は次のように定義される．

$$CSM(\vec{F},\vec{T}) = \frac{ad-bc}{\sqrt{(a+c)(b+d)}}$$

$$a=\sum_{i=1}^{n} f_i\cdot t_i,\quad b=\sum_{i=1}^{n} f_i\cdot(1-t_i),$$

$$c=\sum_{i=1}^{n}(1-f_i)\cdot t_i,\quad d=\sum_{i=1}^{n}(1-f_i)\cdot(1-t_i),$$

この定義式において，a，b，c はオーバラップ相関係数の定義式に含まれるパラメータと同じである．d はどちらの抽象名詞とも共起しない形容詞・形容動詞の種類数に相当する．したがって，次元数 n は a+b+c+d である．$CSM(\vec{F},\vec{T})$ が 1.0 の場合，OVLP と同様に，$\vec{F}$ は $\vec{T}$ を完全に包含することを表す．そして，分子である $ad-bc$ は対称であるが，分母 $(a+c)(b+d)$ は非対称である．したがって，$CSM(\vec{F},\vec{T})$ は $CSM(\vec{T},\vec{F})$ とは a と d が同じである特殊なケースを除いては，異なる．つまり，この定義式は非対称性を持っている．多くの場合，$a << d$ であるため，$CSM(\vec{F},\vec{T})$ と $CSM(\vec{T},\vec{F})$ が等しいことは稀である．実際，実験データにおいては，それぞれの方向について計算した CSM の値が等しくなる単語対はなかった．

3.3 多値画像のための補完類似度

多値画像のための補完類似度（Complementary Similarity Measure for gray-scale images: CSM-g）は二値画像のための補完類似度 CSM-b を拡張した尺度である[18]．CSM-b はグラフィカルデザインなどの汚れに強いが，二値化状態やスキャンの条件が強く影響する．CSM-b は 2×2 分割表の特殊な例であり，その一般形として，CSM-g が定義された．この尺度は直接グレースケールで表される多値画像を扱うため，二値化状態やスキャンの条件に影響されにくいという特徴を持つ．ベクトル $\vec{F_g}=(f_{g_1},...,f_{g_i},...,f_{g_n})$ と $\vec{T_g}=(t_{g_1},...,t_{g_i},...,t_{g_n})$（$f_{g_i},t_{g_i}=0$ から 1）における補完類似度は次のように定義される．

$$CSM_g(\vec{F_g},\vec{T_g}) = \frac{a_g d_g - b_g c_g}{\sqrt{nT_{g^2}-T_g^2}}$$

$$a_g=\sum_{i=1}^{n} f_{g_i}\cdot t_{g_i},\quad b_g=\sum_{i=1}^{n} f_{g_i}\cdot(1-t_{g_i}),$$

$$c_g=\sum_{i=1}^{n}(1-f_{g_i})\cdot t_{g_i},\ d_g=\sum_{i=1}^{n}(1-f_{g_i})\cdot(1-t_{g_i}),$$

$$T_g=\sum_{i=1}^{n} t_{g_i},\quad T_{g^2}=\sum_{i=1}^{n} t_{g_i}^2$$

この定義式において，本研究では，二値画像のための補完類似度と同じように，次元数 n は形容詞・形容動詞の種類数に相当し，非対称性を持つ．しかしながら，各要素 f_{g_i}，t_{g_i} は，抽象名詞が i 番目の形容詞・形容動詞と頻繁に共起するかどうかの状況を表す，共起頻度に基づく重みを用いる．実験では，下記のような重みを用いた．式中の $Freq(noun,adj)$ は抽象名詞 $noun$ が形容詞・形容動詞 adj と共起する頻度で

ある．

$$Weight(noun, adj) = \frac{Freq(noun, adj)}{Freq(noun, adj) + 1}$$

多値画像のための補完類似度を適用する際に，我々は，抽象名詞が形容詞・形容動詞と共起する頻度に注目した．もし，その状況がコーパスに頻繁に現れるのであれば，その抽象名詞と形容詞・形容動詞は親密な関係にあると推測できる．逆に，まったく共起しない抽象名詞と形容詞・形容動詞は疎遠な関係にある推測できる．そこで，この補完類似度が多値ベクトル間の重なり度合いを測ることができることを利用し，それぞれの共起頻度に基づく重みをベクトルの要素とした．この関数では，*noun* と *adj* が共起しないのであれば 0 を，1 回共起するのであれば 0.5 を得る．そして，共起頻度が 2 回以上であれば，緩やかに 1.0 に近づくまで増加する関数である．我々は，1 回でも共起することは何回も共起することよりも重要な情報であることを重みに含ませることを考えた．この意図を代弁する関数は多く考えられるが，本論文では，この簡素な関数を使用して，共起頻度の考慮がもたらす影響を調べることにした．

4. 階層構造の構築方法

本章では，コーパスからの階層構造の構築工程を示す．階層の構築には，閾値以上の単語対を使う．この閾値の設定については，次章に示す．

(1) 各尺度を用いて，2 語間ごとに出現パターン間の包含関係を測り，2 語間の階層関係を推定する．もし単語 X の出現パターンが単語 Y の出現パターンを包含する度合い（包含度）が，単語 Y のパターンが単語 X のパターンを包含する度合いより高いなら，2 語間には，X が上位語，Y が下位語である階層関係があるとし，単語対を (X,Y) と表す．逆に，Y が上位語，X が下位語であるなら，単語対は (Y,X) と表される．

(2) 包含度を正規化し，閾値 TH 未満の単語対を削除する．

(3) 各単語 C_0 について，階層構造を構築する．

(a) 単語 C_0 が上位語である単語対のうち最も高い包含度を持つ対 (C_0, C_{-1}) を階層の初期値 $C_0 - C_{-1}$ とする．

(b) 階層の最後尾に位置する単語 C_{-1} を上位語に持つ単語対のうち最も高い包含度を持つ対 (C_{-1}, C_{-2}) を見つけ，下位語である単語 C_{-2} を階層 $C_0 - C_{-1}$ の最後尾に連結する．ただし，単語 C_{-2} は現行の階層に含まれていないものに限る．

(c) 工程 (3b) に沿った対 $(C_{-i}, C_{-(i+1)})$ $(i > 1)$ が選択できる間，工程 (3b) を繰り返す．

(d) 階層の先頭に位置する単語 C_0 を下位語に持つ単語対のうち最も高い包含度を持つ対 (C_1, C_0) を見つけ，上位語である単語 C_1 を階層 $C_0 - C_{-1} - C_{-2} - ... - C_{-n}$ の先頭に連結する．ただし，単語 C_1 は現行の階層に含まれていないものに限る．

(e) 工程 (3d) に沿った対 (C_{j+1}, C_j) $(j > 0)$ が選択できる間，工程 (3d) を繰り返す．

(4) 構築した階層について，もし短い階層が単語の順序が保持された状態でより長い階層に完全に含まれる（包含される）なら，短い階層を階層の集合から削除する．たとえば，階層 B – D – E – F とより長い階層 A – B – C – D – E – F があるとする．このとき，短い階層を構築する単語がすべて長い階層 A – B – C – D – E – F に順序が保持された状態で存在する，つまり完全に包含されるるので，B – D – E – F を階層の集合から削除する．

5. 実　　験

実験において，各尺度を用いて構築された階層を比較するために，抽象名詞「こと」を最上位語に持つ階層をできるだけ多く構築できる閾値を設定し，構築される階層の条件をそろえた．「こと」は意味的に広く使える抽象名詞であり，実験データにおいて，最も共起する形容詞・形容動詞が多い抽象名詞である．得られた階層を**図 3** にいくつか示す．実験的に設定した閾値 TH はそれぞれ以下のような値である．

- OVLP の場合，$TH = 0.2$
- CSM-b の場合，$TH = 0.2$
- CSM-g の場合，$TH = 0.12$

5.1 EDR 電子化辞書との比較方法

本論文では，4 章に記述した構築手法によって得られた階層を評価するために，既存の EDR 電子化辞書[4] に含まれる形容詞・形容動詞の概念階層と比較する．EDR 電子化辞書は 1995 年，計算機による自然言語の利便的な処理のために開発された．この辞書は人手で編集された 11 個の辞書からなり，そこには概念辞書，単語辞書，日英辞書などが含まれている．構築した階層構造は形容詞・形容動詞の上位語として定義さ

こと – 状態 – 状況 – 兆候
こと – 状態 – 傾斜 – 勾配
こと – 状態 – 関係 – かかわり – つきあい
こと – 面 – イメージ – 印象 – 顔立ち – 品格 – 血筋
こと – 時 – 温度 – 幸福感
こと – 時 – 様子 – 顔つき – 面持ち – 口ぶり
こと – 規模 – 数 – 量
こと – 方 – 空間 – 面積
こと – 面 – イメージ – 印象 – 風格 – 家柄 – 血統 – 家系 – 血筋
こと – 面 – イメージ – 性格 – 印象 – 一面 – 態度 – 人柄 – 気質 – 気風 – 気性
こと – 面 – イメージ – 美しさ – 若さ – 大胆さ
こと – 面 – イメージ – 体 – 体格 – 背
こと – 面 – 側面 – 意味 – 方向 – 観点 – 目 – 視野 – 角度 – アイディア
こと – 面 – 側面 – 意味 – 色彩 – 意味合い – 観点 – 見地 – 分野 – 領域 – 枠組み – 枠
こと – 面 – 層 – 階級 – 血筋

図 3 構築された階層の一部
Fig. 3 Samples of extracted hierarchies.

れた抽象名詞[8] の階層構造である．このため，EDR 電子化辞書における形容詞・形容動詞に関する概念階層を正解データとして評価に用いた．また，この辞書には形容詞・形容動詞に関する概念階層が 932 階層あり，深さ 3 から 14 の範囲に分布する．実験において得られた階層は深さ 3 から 15 の範囲に分布するため，この概念階層との比較は実験で得られた階層を評価することに適していると考えた．具体的には，EDR 電子化辞書からの階層と一致する度合い（一致度）をそれぞれの尺度によって得られた階層について測定することで，得られた階層を比較評価する．

しかしながら，得られた階層と異なり，EDR 電子化辞書の概念階層を構成する概念は単語ではなく，概念 ID と説明文で記述されているため，抽象名詞で構築された階層と比較するには，変換が必要であった．そこで，各概念記述について内容語である名詞，動詞を取り出し，さらに，それらの単語に類義語を付与し，その列で文を置き換えた．同様に，構築された階層中の抽象名詞にも類義語を付与し，使用単語の違いを軽減した．用いた類義語は EDR 電子化辞書から抽出したものである．このように変換された階層において，構築された階層の各ノードにある抽象名詞は，その抽象名詞とその類義語で，ノード（抽象名詞，類義語 $_1$，類義語 $_2$,...）と表され，EDR 電子化辞書から抽出した階層の各ノードにある概念は，その概念記述にある内容語とそれらの類義語で，ノード（内容語 1，類義語 1_1，類義語 1_2，内容語 2，類義語 2_1，類義語 2_2,...）と表される．このとき，構築された抽象名詞の階層のノードにある抽象名詞または類義語が，EDR 電子化辞書の階層のノードにある内容語または類義語と一致するのであれば，そのノードはその EDR 電子化辞書の階層のノードと一致すると考える．たとえば，x を抽象名詞または内容語，x’，x” を x の類義語としたとき，実験において構築された階層が次のように表現される階層であるとする．

「A(a, a’, a”) – B(b, b’, b”) – C(c, c’, c”)
– D(d, d’, d”)」

同様に，この階層に対応する EDR 電子化辞書の階層が次のように表現される階層であるとする．

「P(p, p’, a) – Q(q, b, b”) – R(r, r’, r”) –
S(s, s’, d, f, f’, f”) – T(t, t’, g, g’)」

下線は階層間で一致する単語とその単語を持つノードを示し，単語 a を持つノード A はノード P，b や b” を持つノード B はノード Q，単語 d を持つノード D はノード S と一致する．したがって，構築された階層は EDR 電子化辞書の階層と 3 つのノードが一致すると数え，その階層の一致度を 3 と定義する．ただし，構築された階層と EDR 電子化辞書の階層とに共通するノードがあってもその上下関係が逆転している場合は数えない．たとえば，上記の例で，EDR 電子化辞書の階層が Q – P – R – S – T となっている場合は，B と Q，D と S は一致するが，A と P は一致しない，あるいは，A と P，D と S は一致するが，B と Q は一致しないと考え，この場合の一致度は 2 となる．

なお，共通するノードの数え方が複数ある場合は，最も値が大きくなる数え方を採用する．

また，この方法で EDR 電子化辞書の階層と構築された階層を比較する中で，構築された階層において階層関係にある単語が EDR 電子化辞書では類義語である場合があることが分かった．たとえば，我々の手法では次のような階層「こと – ところ – イメージ – 雰囲気 – 空気 – 感情 – 心情 – 心境 – 感慨 – 思い出」を構築する．EDR 電子化辞書では，類義語とは「同じ概念にリンクされる単語」と定義されており，我々は EDR 電子化辞書における類義語を集めることができる．実際，この階層において下線で示す語の関係は，EDR 電子化辞書では「感情」と「心情」，「心境」は類義語，「雰囲気」と「空気」も類義語である．この階層の EDR 電子化辞書の階層との一致度を厳密に数えた場合，「こと – ところ – イメージ – 雰囲気（または，空気）– 感情（または，心情，心境）– 思い出」が一致し，一致度は 6 となる．しかしながら，もし EDR 電子化辞書では類義語である単語間の階層関係を許す場合，「こと – ところ – イメージ – 雰囲気 – 空気 – 感情 – 心情 – 心境 – 思い出」が一致し，一致度は 9 となる．本論文では，これらの類義語間の階層関係をコーパスに依存した関係として考え，後者の条件で，EDR 電子化辞書の階層との一致度を測った．

6. オーバラップ相関係数と補完類似度との比較

はじめに，オーバラップ相関係数（OVLP）を用いて抽出した階層と，二値画像のための補完類似度（CSM-b）を用いて抽出した階層とを比較する．**表 1** に EDR 電子化辞書の概念階層と OVLP によって得られた階層との一致度を示す．同様に，**表 2** に EDR 電子化辞書の概念階層と CSM-b によって得られた階層との一致度の深さごとの分布を示す．たとえば，表 2 にお

表 1 深さごとの OVLP による階層の一致度の分布
Table 1 Distribution of OVLP hierarchy for each depth.

深さ	EDR 電子化辞書にある階層との一致度						平均
	1	2	3	4	5	6	
3	4	36	4				2.00
4	0	21	43	17			2.95
5	0	15	37	22	2		3.14
6	0	0	9	7	5	0	3.81
7	0	1	3	4	0	1	3.38
8	0	0	0	0	0	0	0.00
9	0	0	0	0	0	0	0.00
10	0	0	1	0	0	0	3.00
						全体の平均	2.28

表 2 深さごとの CSM-b による階層の一致度の分布
Table 2 Distribution of CSM-b hierarchy for each depth.

深さ	EDR 電子化辞書にある階層との一致度									平均
	1	2	3	4	5	6	7	8	9	
3	1	2	1							2.00
4	0	6	18	9						3.09
5	0	7	23	12	3					3.24
6	0	4	12	9	7	4				3.86
7	0	2	2	10	4	3	1			4.32
8	0	0	1	6	6	3	0	0		4.69
9	0	0	1	4	5	4	5	1	0	5.55
10	0	0	2	0	2	2	0	0	1	5.29
11	0	0	1	0	0	1	0	0	0	4.50
12	0	0	1	1	0	0	0	0	2	6.25
									全体の平均	4.28

いて，深さ 3 の CSM-b の階層は 4 つあり，そのうち EDR 電子化辞書との一致度が 1 のものは 1 つ，一致度が 2 のものは 2 つ，一致度が深さと同じ 3，すなわち双方が完全に一致するものは 1 つある．表中では，一致度と階層の深さが同じである階層の数は下線付きで表される．また，「平均」は深さ 3 の階層の一致度の平均（$(1*1+2*2+1*3)/4 = 2.00$）であり，「全体の平均」は全階層の一致度の総和を階層の総数で割った値である．

実験において，OVLP を用いて得られた階層は 232 個，CSM-b を用いて得られた階層は 189 個であった．表 1，2 に示されるように，OVLP の階層は深さ 3 から 10 の範囲に分布し，CSM-b それらの階層は深さ 3 から 12 の範囲に分布する．この結果から，CSM-b は OVLP より得られる階層の数は少ないが，OVLP よりも長い（深い）階層を得られることが分かる．また，表 1 から，OVLP の階層の多くは 2 から 4 の一致度を持ち，最も高い一致度 6 を持つ階層が 1 つであることが分かる．これは，OVLP によって得られた階層が深さ 3 から 5 に集中しているためである．一方，表 2 から，CSM-b の階層の多くは 2 から 6 の一致度を持っているが，全体的に広く分布し，最も高い一致度 9 を持つ階層が 3 つある．深さごとの平均を見ると，CSM-b の階層はより深い階層がより高い一致度を持つという，深さによる一致度の増加傾向が見られる．

一致度の「全体の平均」を見ると，全体的に CSM-b の階層は OVLP の階層よりも EDR 電子化辞書の階層に一致することが分かる．また，各深さにおける一致度の平均から，同じ深さの階層について，CSM-b は OVLP よりも EDR 電子化辞書の階層に一致する階層を構築していることが分かる．

また，どちらの尺度を用いた階層においても，多くの抽象名詞は EDR 電子化辞書の概念階層において，階層の根に近い上位概念と一致することが分かった．現在のシソーラスでは，語彙は人間の直感に基づいて，アプリオリにカテゴリ化され，分類されている．このことから，少なくとも，どちらの階層も根に近い部分は，人間の直感に近いと考えられる．そして，EDR 電子化辞書の階層との一致度による評価から，CSM-b を用いて構築した階層は OVLP を用いて構築した階層よりも人間の直感に近い階層であると考察する．

7. 共起頻度を考慮することによる階層構造構築への影響

共起頻度は共起する単語間の関係の強さを測ることができる重要な情報である．6 章において，二値画像のための補完類似度（CSM-b）がオーバラップ相関係数（OVLP）よりも人間の直感に近い階層を得られることを示した．しかしながら，CSM-b は，二値ベクトル間の包含関係を測る尺度であるため，出現パターンを表すベクトルには，0，1 で表現できる，共起するかしないかという状況しか考慮できず，その頻度情報を考慮することができない．そこで，多値画像のための補完類似度（CSM-g）を階層構造の構築に適用することを考えた．CSM-g は多値ベクトル間の包含関係を測ることができる．つまり，ベクトルの要素に共起頻度に基づく重みを利用できる．この章では，CSM-g を使って，共起頻度を考慮した場合の階層構築への影響を調べる．具体的には，CSM-g を用いて構築した階層を CSM-b による階層と比較し，共起頻度を考慮することで，より良い階層構造が得られるかを考察する．

7.1 構築された階層間の比較

まず，4 章に示した構築手法に沿って，CSM-b による階層と CSM-g による階層とを照らし合わせる．**表 3** に示すように，CSM-b は 189 個の階層，CSM-g は 178 個の階層を抽出した．そのうち，共通して抽出できた階層は 28 個しかなく，そのほとんどは深さ 3 から 6 と短い階層であった．たとえば，深さ 5 の階層「こと – 状態 – 関係 – つながり – 縁」がそこには含まれる．また，一方の尺度による階層を他方の尺度による階層が完全に包含する階層を比べると，CSM-b の階層を完全に包含する CSM-g の階層は，CSM-g の階層を完全に包含する CSM-b の階層より多い（(D) < (E)）．これは，CSM-g が CSM-b よりも長い（深い）階層を抽出できる特徴を持つことを示唆している．CSM-b の階層を完全に包含する CSM-g の階層の例を**図 4** に示す．ここで，下線は完全に包含された CSM-b の階層を構成する抽象名詞を表す．

実際，階層の深さを比較すると，CSM-b の階層は深さ 3 から 12 の範囲に，CSM-g の階層は深さ 3 から 15 の範囲に分布する．また，階層を構成する抽象名詞の異なり数を比較すると，全 354 種類中，CSM-b は 318，CSM-g は 314 であり，網羅性には差が見られ

表 3 CSM-b と CSM-g の階層の数に関する比較
Table 3 Comparison of hierarchies for the numbers.

階層の種類		数
(A)	CSM-b によって得られた階層	189
(B)	CSM-g によって得られた階層	178
(C)	共通して得られた階層	28
(D)	CSM-g の階層を包含する CSM-b の階層	5
(E)	CSM-b の階層を包含する CSM-g の階層	38

こと – ところ – イメージ – 印象 – 外見 – 物腰 – 気品 – 品格 – 血統 – 家系
こと – 時 – 日和 – 温度 – 幸福感
こと – ところ – しぐさ – 面影 – 可愛さ

図 4 CSM-b の階層を完全に包含する CSM-g の階層の例
Fig. 4 Examples of a CSM-g hierarchy including a CSM-b hierarchy.

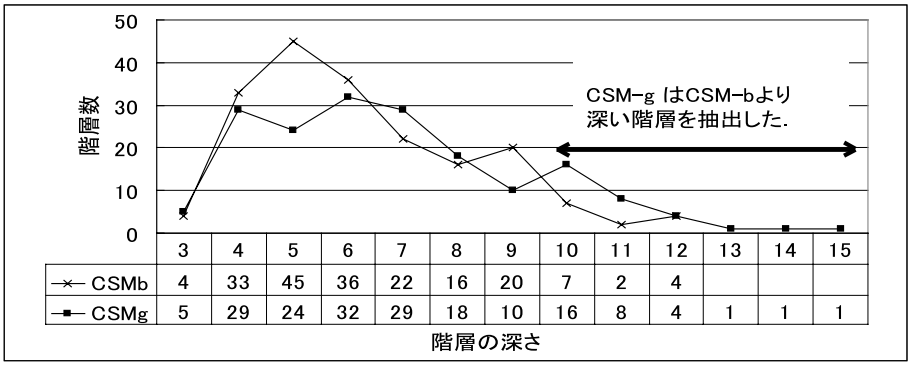


図 5 CSM-b の階層と CSM-g の階層の深さに関する分布
Fig. 5 Distribution of hierarchies for each depth.

表 4 深さごとの CSM-g による階層の一致度の分布
Table 4 Distribution of CSM-g hierarchy for each depth.

深さ	EDR 電子化辞書にある階層との一致度									平均
	1	2	3	4	5	6	7	8	9	
3	1	3	1							2.50
4	0	6	13	10						3.14
5	0	3	9	9	3					3.50
6	0	1	11	12	6	2				3.91
7	0	1	5	10	8	5	0			4.38
8	0	0	4	5	7	2	0	0		4.39
9	0	0	0	6	1	3	0	0	0	4.70
10	0	0	0	0	2	6	4	3	1	6.69
11	0	0	1	2	1	3	1	0	0	5.13
12	0	0	0	0	0	0	0	1	3	8.75
13	0	0	0	0	0	0	1	0	0	7.00
14	0	0	0	0	0	0	0	1	0	8.00
15	0	0	0	0	0	0	0	0	1	9.00
									全体の平均	5.47

なかった．**図 5** に CSM-b の階層と CSM-g の階層の深さに関する分布を示す．これらの結果から，CSM-g は抽出できる階層の数は CSM-b より少ないけれども，より深い階層を抽出できることが分かる．

7.2 EDR 電子化辞書にある階層との一致度

次に，6 章でのオーバラップ相関係数との比較と同様に，それぞれ得られた階層を人手によって構築された EDR 電子化辞書にある形容詞・形容動詞の概念階層と比較し，階層の一致度を測る．表 2 に示す CSM-b の階層に関する一致度の分布と**表 4** に示す CSM-g の階層に関する一致度の分布から，より深い階層はより高い一致度を持つ傾向にあることが分かる．表 4 においては，その傾向を表 2 よりはっきりと見ることができる．また，**図 6** に示すように階層の深さごとに一致度の平均を比べると，深さ 8 と 9 以外の深さにおいて，CSM-g のほうが CSM-b より高い値を持つこ

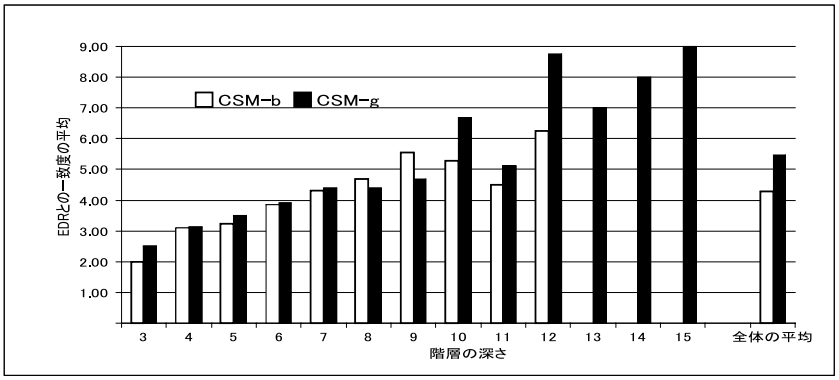


図 6 深さごとの一致度の平均による比較
Fig. 6 Comparison for the average of agreement levels.

とが分かる．これらのことから，CSM-g は全体的に CSM-b より EDR 電子化辞書の概念階層に近い，つまり，共起頻度を考慮したほうが，人間の直感に近い階層を構築していると考えられる．

また，OVLP の階層と CSM-b の階層との比較と同様に，CSM-g の階層も EDR 電子化辞書の概念階層と一致したノード（抽象名詞）を分析すると，多くの場合，最上位近くの上位概念と一致する傾向にあった．したがって，本論文で構築した抽象名詞の階層は，どの尺度を適用した場合でも，少なくとも形容詞・形容動詞の階層構造の上位の辺りにおいて人間の直感にあった抽象名詞の階層構造を構築できると考察する．

7.3 2 語間の階層関係を推定する能力の比較

本研究では，語彙の階層構造を構築するために，4 章に示す第 1 工程で出現パターンの包含関係を測ることで，2 語間の階層関係を推定する．このため，この関係の推定結果が抽出される階層を大きく左右する．そこで，CSM-b と CSM-g の 2 つの抽象名詞間の階層関係を推定する能力を比較検証する．これらの推定結果を見ると，いくつかの抽象名詞の対において，推定結果が反対であることが分かった．**表 5** にそのような対の例を示す．これらの対は CSM-g によって左の名詞が上位語，右の名詞が下位語として推定されたが，CSM-b は CSM-g とは反対に推定したものである．

実験において，表 5 にあげたような推定結果が異なる対は，計算された 17,201 の対の中で，5%以下の 836 個含まれていたことが分かった．また，このような対の多くは構築された階層の中間層に現れることが分かった．たとえば，「こと – ところ – イメージ – 印象 – 感じ – 気分 – 気持ち – 思い – 願い – 念」は CSM-g によって構築された階層である．このように，CSM-g は「気持ち」を上位語，「思い」を下位語と関係を推定した．一方，CSM-b は「思い」を上位語，「気持ち」を下位語と推定した．**表 6** に，「気持ち」と「思い」が相互に包含する度合いを CSM-g と CSM-b のそれぞれによって推定した値を示すように，「思い」の出現パターンが「気持ち」の出現パターンを包含する度合いと，逆に「気持ち」が「思い」を包含する度合いを比べると，CSM-g も CSM-b もほとんど差がない．補完類似度はベクトル間の包含関係を測る尺度であるため，それぞれの方向で測った CSM の値がほとんど同じであるならば，その 2 つの名詞のパターンは似ていることを示していることになる．し

表 5 推定結果が異なる抽象名詞の対の一部
Table 5 Noun pairs estimated oppositely.

対
(ところ，イメージ)
(ところ，面)
(ところ，印象)
(ところ，性格)
(ところ，性質)
(ところ，感触)
(気持ち，思い)
(輝き，光沢)
(空間，面)
(言葉，意見)
(心，心情)
(日和，温度)

表 6 単語対「気持ち」と「思い」の CSM の値
Table 6 Differences in CSM values for "*omoi*" and "*kimochi*".

$(\vec{F},\vec{T})$	(思い，気持ち)	(気持ち，思い)	差
CSM-g	0.7632	0.7700	+0.0068
CSM-b	0.8094	0.8064	−0.0030

表 7 単語対「ところ」と「イメージ」の CSM の値
Table 7 Differences in CSM values for "*tokoro*" and "*imeeji*".

$(\vec{F},\vec{T})$	(イメージ，ところ)	(ところ，イメージ)	差
CSM-g	0.6468	0.6631	+0.0163
CSM-b	0.7156	0.6767	−0.0389

たがって，この「思い」と「気持ち」は類義語であると考えるのが自然であろう．事実，この 2 つの名詞は EDR 電子化辞書において類義関係にある．しかしながら，我々は CSM の値によって，単純に 2 語間の上位下位関係を決定しており，類義関係にあるかどうかを判定していない．2 語の関係を推定する際に，上位下位関係だけでなく，類義関係も考慮して，補完類似度に基づく定義を導入することは今後の課題である．

一方で，それぞれの方向で測った CSM の値が大きく異なる対を見つけることができた．たとえば，**表 7** に示す「ところ」と「イメージ」の対である．この対は，CSM-g によって「ところ」が上位語，「イメージ」が下位語と推定されたが，CSM-b は「イメージ」を上位語，「ところ」を下位語と推定した．この違いは，共起頻度を考慮することによって得られた違いである．共起する形容詞・形容動詞は一緒なのに，反対に推定されたということは，「ところ」が「イメージ」よりそれぞれの形容詞・形容動詞と共起する頻度が高いことを表している．したがって，この結果は共起頻度を考慮している CSM-g のほうが，2 語間の関係を厳密に推定できることを示唆している．

8. ま と め

本論文では，コーパスにおける出現パターンの包含関係に基づき，自動的に語彙の階層構造を構築する手法を提案した．そして，提案手法の適用可能性を示すために，形容詞・形容動詞と共起する抽象名詞の階層構造を構築することを試みた．本論文では，2 語間の階層関係を見つけるために補完類似度を適用した．その結果，実験において，補完類似度は劣化印刷文字を認識するために開発された類似尺度であるにもかかわらず，コーパスから語彙の階層構想を抽出できることを示した．さらに，各形容詞・形容動詞との共起頻度を考慮し，抽象名詞の階層構築への影響を調査した．具体的には，ベクトルの要素として，共起頻度に基づく重みを用い，多値画像認識のために改良された補完類似度を適用した．そして，共起頻度の情報を考慮しない二値画像のための補完類似度によって構築された階層と比較したところ，得られる階層の数は少ないにもかかわらず，より深い階層構造が得られた．そして，EDR 電子化辞書の概念階層との一致度において比較した結果，共起頻度を考慮したほうがより人間の直感に近い階層を構築しうることを示した．

一方，実験において得られた階層構造には，既存の階層とは異なる人間の直感にあった階層が見つけられる．たとえば，「こと– 面 – イメージ – 印象 – 感じ – 気分 – 気持ち – 感情 – 心情 – 心境– 感慨 – 思い出」という階層が得られた．このような階層構造を分析し，評価することも必要である．また，階層構造に類似関係や同義関係を組み込むことも今後の課題である．

参 考 文 献

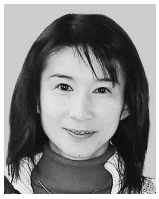
1) 安藤まや，関根　聡：上位語・下位概念を含む連体修飾表現の言語的分析，言語処理学会第 10 回年次大会発表論文集，pp.205–208 (2004).
2) Berland, M. and Charniak, E.: Finding parts in very large corpora, *The 37th Annual Meeting of the Association for Computational Linguistics*, pp.57–64 (1999).
3) Caraballo, S.A.: Automatic construction of a hypernym-labeled noun hierarchy from text, *The 37th Annual Meeting of the Association for Computational Linguistics*, pp.120–126 (1999).
4) EDR 電子化辞書 (1995). http://www2.nict.go.jp/kk/e416/EDR/index.html
5) Grefenstette, G.: *Explorations in Automatic Thesaurus Discovery*, Kluwer Academic Publishers, Boston/Dordrecht/London (1994).
6) Hagita, N. and Sawaki, M.: Robust Recognition of Degraded Machine-Printed Characters using Complimentary Similarity Measure and Error-Correction Learning, *The SPIE–The International Society for Optical Engineering*, No.2442, pp.236–244 (1995).
7) Hearst, M.A.: Automatic acquisition of hyponyms from large text corpora, *The 14th International Conference on Computational Linguistics*, pp.539–545 (1992).
8) Kanzaki, K., Ma, Q., Yamamoto, E., Murata, M. and Isahara, H.: Adjectives and their abstract concepts—Toward an objective thesaurus from a semantic map, *The 2nd International Workshop on Generative Approaches to the Lexicon*, pp.177–184 (2003).
9) Kanzaki, K., Yamamoto, E., Ma, Q. and

Isahara, H.: Construction of an objective hierarchy of abstract concepts via directional similarity, *The 20th International Conference on Computational Linguistics*, Vol.2, pp.1147–1153 (2004).
10) 北 研二：言語と計算（4）確率的言語モデル統計モデル，東京大学出版会 (1999).
11) Ma, Q., Kanzaki, K., Zhang, Y., Murata, M. and Isahara, H.: Self-Organizing Semantic Maps and its Appliaction to Word Alignment in Japanese-Chinese Parallel Corpora, *Neural Networks*, Vol.17, No.8–9, pp.1241–1253 (2004).
12) Manning, C.D. and Schutze, H.: *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge MA (1999).
13) 松本裕治，須藤 茂，中山拓也，平尾 努：複数の言語資源からのシソーラスの構築，情報処理学会研究報告，FI-042, pp.23–28 (1996).
14) Miller, A., Beckwith, R., Fellbaum, C., Gros, D., Millier, K. and Tengi, R.: Five Papers on WordNet, Technical Report CSL Report 43, Cognitive Science Laboratory, Princeton University (1990).
15) 中山拓也，松本裕治：シソーラスへの未登録語の自動登録，情報処理学会研究報告，NL-120, pp.103–108 (1997).
16) 根元今朝男：「が格」の名詞と形容詞とのくみあわせ，「電子計算機のための国語研究 II」，pp.63–73, 国立国語学研究所 (1969).
17) Ruge, G.: Automatic Detection of Thesaurus relations for Information Retrieval Applications, *Lecture Notes in Computer Science*, Vol.1337, pp.499–506 (1997).
18) Sawaki, M., Hagita, N. and Ishii, K.: Robust Character Recognition of Gray-Scaled Images with Graphical Designs and Noise, *The International Conference on Document Analysis and Recognition*, pp.491–494 (1997).
19) Schmid, H.-J.: *English Abstract Nouns as Conceptual Shells*, Mouton de Gruyter (2000).
20) 正津康弘，徳永健伸，田中穂積：国語辞典とシソーラスの統合，情報処理学会研究報告，NL-153, pp.141–146 (2003).
21) 高橋太郎：文中にあらわれる所属関係の種々相，国語学 103，pp.1–16, 国語学会 (1975).
22) 鶴丸弘昭，日高 達，吉田 将：単語間の上位–下位関係の自動抽出，情報処理学会研究報告，FI-003, pp.1–8 (1986).
23) 山本英子，梅村恭司：コーパス中の一対多関係を推定する問題における類似尺度，自然言語処理，Vol.9, No.2, pp.45–75 (2002).
24) Yamamoto, E. and Isahara, H.: Knowledge Acquisition Based on Automatically-Extracted Word Hierarchies form Domain-Specific Texts, *International Conference Recent Advances in Natural Language Processing 2005*, pp.632–636 (2005).

(平成 17 年 10 月 6 日受付)
(平成 18 年 4 月 4 日採録)

山本 英子（正会員）

1998 年豊橋技術科学大学大学院工学研究科情報工学専攻修士課程修了．2002 年同大学院工学研究科電子・情報工学専攻博士後期課程修了．博士（工学）．現在，独立行政法人情報通信研究機構自然言語グループ有期研究員．自然言語処理，情報抽出の研究に従事．言語処理学会，人工知能学会各会員．



神崎 享子

1994 年早稲田大学大学院文学研究科修士課程修了．1998 年同大学院文学研究科博士後期課程単位取得後満期退学．2001 年神戸大学大学院自然科学研究科博士課程修了．博士（学術）．現在，独立行政法人情報通信研究機構自然言語グループ研究員．自然言語処理，言語学の研究に従事．言語処理学会，計量国語学会，日本言語学会，日本語学会各会員．

井佐原 均（正会員）

1980 年京都大学大学院工学研究科修士課程修了．博士（工学）．同年通商産業省電子技術総合研究所入所．1995 年郵政省通信総合研究所．現在，独立行政法人情報通信研究機構知識創成コミュニケーション研究センター自然言語グループリーダー，同アジア研究連携センター自然言語ラボラトリー長．自然言語処理，語彙意味論の研究に従事．言語処理学会，人工知能学会，日本認知科学会各会員．