形態素解析結果から過分割を検出する統計的尺度

内山　将夫[†]

本稿では，形態素解析の結果から過分割 (正解が分割していないところを形態素解析システムが分割している個所) を検出するための統計的尺度を提案する．もし，形態素解析の結果から過分割を検出できれば，それを利用して形態素解析結果の過分割を訂正する規則を作成できるし，人手修正済みのコーパスで除去しきれていない過分割を発見し取り除くこともできるため，そのような尺度は有用である．本稿で提案する尺度は文字列に関する尺度であり，文字列が分割される確率と分割されない確率との比に基づいていて，分割されにくい文字列ほど大きな値となる．したがって，この値が大きい文字列は過分割されている可能性が高い．本稿の実験では，この尺度を使うことにより，規則に基づく形態素解析システムの解析結果から，高精度で過分割を検出できた．また，人手で修正されたコーパスに残る過分割も検出できた．これらのことは，提案尺度が，形態素解析システムの高精度化に役立つこと，及び，コーパス作成・整備の際の補助ツールとして役立つことを示している．

キーワード：　コーパス, 日本語形態素解析, 分割誤り

Statistical Measure for Detecting Over-Segmentations in Results of Japanese Morphological Analysis

UTIYAMA Masao [†]

This paper proposes a statistical measure for detecting over-segmentations, which are errors in segmentation where a morphological analyzer segments places which should not be segmented, in results of Japanese morphological analysis. Such a measure is useful because we can use detected over-segmentations for creating error correction rules or for removing remaining errors in manually debugged corpora. The measure proposed in this paper is based on the ratio of the probability of a whole string to that of the string being segmented into two parts. Therefore, the value of the measure is high when a given string is rarely segmented into two parts. Consequently, a string rated high by the measure is likely to contain over-segmentations. In the experiments, the measure detected over-segmentations in the results of rule-based morphological analyzers very precisely and it also detected remaining over-segmentations in manually debugged corpora. These results show that the proposed measure is useful for developing high quality Japanese morphological analyzers and for developing/debugging corpora.

KeyWords:　*corpus, Japanese morphological analysis, segmentation error*

† 信州大学工学部電気電子工学科, Department of Electrical and Electronic Engineering, Faculty of Engineering, Shinshu University(現在，郵政省通信総合研究所，Communications Research Laboratory, Ministry of Posts and Telecommunications)

1　はじめに

日本語の形態素解析は，日本語の自然言語処理にとって基本的なものであるので，多くの研究・開発が行われている．形態素解析システム[1]には，主に，人手で作成された規則に基づくシステム (黒橋 長尾 1997; 松本, 北内, 山下, 平野, 今一, 今村 1997; 鷲坂, 山崎, 廣津, 尾内 1997; Fuchi and Takagi 1998, など) と確率に基づくシステム (Nagata 1994; 森 長尾 1998; 山本 増山 1997, など) がある．

本稿では，人手で作成された規則に基づく形態素解析システムを対象として，形態素解析の結果から半自動的に誤りを検出することを試みる．

形態素解析結果から誤りが検出できた場合には，次のような利点がある．

(1)　形態素解析の誤りは，形態素解析システムの弱点を示していると考えられるので，誤りを分析することにより，システムの性能を向上できる可能性がある．

(2)　形態素解析が誤るような表現を連語として登録することで，そのような誤りが再び起きないようにできる (山地, 黒橋, 長尾 1996; Fuchi and Takagi 1998)．

(3)　形態素解析の誤りから誤り訂正規則を作成できるので，その規則を利用して形態素解析の精度を向上できる (横尾, 白井, 奥山, 河村, 池原 1997; 久光 丹羽 1998)．

(4)　形態素解析の誤りに基づいて，形態素解析の規則に割当てるコストを調整したり (小松 1998)，品詞分類を変更する (北内, 宇津呂, 松本 1998) ことができる．

これらのことから，形態素解析結果から誤りを検出することは，形態素解析システムの高精度化に役立つことがわかる．

しかし，形態素解析の結果から誤りを見付けるのは，形態素解析の精度が 97〜99 % (Fuchi and Takagi 1998) に達している現在では，困難になっている．ところが，従来の研究で，形態素解析結果の誤りを利用して形態素解析の精度を向上させようとしている研究では，それらの誤りを人手で発見すること，あるいは，人手で作成されたコーパスと形態素解析結果とを比較することにより発見することが前提になっている．そのため，形態素解析の誤りを発見することはコストが高い作業となっている．

一方，本稿では，従来の研究で人手で発見されることが前提となっていた解析誤り (特に過分割) を，生のコーパスを形態素解析した結果から半自動的に抽出することを目指し，そのための統計的尺度を提案する．更に，本稿では，人手により誤りが修正済みのコーパスに対しても提案尺度を適用し，人手で除去しきれていない誤りを検出することも試みる．もし，人手修正されたコーパスから誤りを検出できたら，提案尺度はコーパス作成・整備の際の補助ツールとして役立つことになる．

以下，2章では，本稿が検出対象とする誤り (過分割) の定義を述べ，それを検出するための統

1 以下，システムとは，形態素解析システムのことであり，解析結果あるいは形態素解析結果とは，形態素解析システムの解析結果のことである

計的尺度について述べる．3章では，提案尺度を，公開されている形態素解析システム (黒橋・長尾 1997; 松本他 1997; 鷲坂他 1997)，および，人手で修正されたコーパス (EDR 電子化辞書研究所 1995; 黒橋, 斎藤, 坂口 1998) に適用した結果について述べると共に，提案尺度を各種統計的尺度と定量的に比較する．4章では，提案尺度の有効性などを論じる．5章は結論である．

2　過分割を検出する統計的尺度

まず，形態素解析システムの解析結果における分割誤りを分類し，検出対象である過分割を定義づける．次に，過分割を検出する統計的尺度について述べる．

2.1　形態素解析結果における分割誤り

分割点という用語を導入し，それを用いて，形態素解析結果における分割誤りを分類する．

まず，長さ n の文字列 Sを，$S = c_1, c_2, ..., c_n$とする．このとき，Sの i 番目の分割候補点とは，文字 c_iと文字 c_{i+1}の間をいう．また，分割候補点が分割点であるとは，その分割候補点が形態素境界である場合をいい，その分割候補点は分割されているという．

たとえば,「$S =$ 休憩室」とすると，1 番目の分割候補点は「休」と「憩」の間であり，2 番目の分割候補点は「憩」と「室」の間である．更に，Sが「休憩/室」のように分割されているとすると，2 番目の分割候補点は分割点である．

次に，形態素解析結果の分割誤りとは，正解と形態素解析結果とで，分割点が異なる場合をいう．そして，分割誤りのなかで，過分割とは，正解で分割されていない分割候補点をシステムが分割している場合をいう．また，分割不足とは，正解で分割されている分割候補点をシステムが分割していない場合をいう[2]．

たとえば,「休憩室」の分割の正解が「休憩/室」であるとき，システムが「休/憩室」と分割したとすると，1 番目の分割候補点 (「休」と「憩」の間) は過分割であり，2 番目の分割候補点 (「憩」と「室」の間) は分割不足である．

なお，形態素解析結果の誤りには，他には品詞付けの誤りがある．これは，形態素への分割

2 ここで定義した過分割と分割不足とは，形態素境界 (分割点) に注目したものであるので，形態素自体に注目した過分割/分割不足の定義とは異なる．たとえば，(久光・丹羽 1998) では，分割の誤りを以下の 3 種類に分類している．なお,「正」で示される分割は，当該文字列の正しい分割を示し,「誤」で示される分割は，形態素解析システムによる誤った分割を示す．

(形態素自体に注目した) 過分割　　正:　今日/の/金/相場/は,...
誤:　今日/の/金/相/場/は,...

(形態素自体に注目した) 分割不足　　正:　ユニックス/ワークステーション
誤:　ユニックスワークステーション

その他の誤り (語境界交差型)　　正:　病気/が/まん延
誤:　病気/がまん/延

この定義では，形態素境界を直接取り扱えないので，本稿での目的には不適切である．なお，語境界交差型の分割誤りは，本稿の定義では，過分割と分割不足が複合したものとなる．たとえば，上の例では,「が」と「ま」の間の分割候補点が分割不足であり,「ん」と「延」の間の分割候補点が過分割である．

自体は正しいが，品詞が間違った場合である．この誤りの検出については，分割不足と同様に，本稿では考察しない．

2.2　過分割の検出尺度の定義

ここで定義する尺度は，文字列に関する尺度であり，与えられた文字列が分割される場合と分割されない場合とで確率を比較し，分割されない確率が高いほど大きな値をとる尺度である．そのため，この尺度の値が大きいような分割をされている文字列は，誤った分割 (過分割) をされている可能性が高い．

より厳密には，与えられた文字列を $S = a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_l$とし，$S$の二つの部分文字列 (形態素[3]) を $A = a_1, \ldots, a_k$と $B = b_1, \ldots, b_l$とするとき，

$$L(A,B) = \log \frac{\Pr(\langle w\rangle, a_1, \ldots, a_k, b_1, \ldots, b_l, \langle /w\rangle)}{\Pr(\langle w\rangle, a_1, \ldots, a_k, \langle /w\rangle)\Pr(\langle w\rangle, b_1, \ldots, b_l, \langle /w\rangle)} \tag{1}$$

は，以下で述べるように，文字列 Sの形態素 A と Bへの切れにくさを表現している．ただし，$\langle w\rangle$ と $\langle /w\rangle$ は，それぞれ，形態素の前後に付ける区切り記号であり，$\Pr(\cdots)$ は，文字列の生起確率である．すなわち，$\Pr(\langle w\rangle, c_1, \ldots, c_k, \langle /w\rangle)$ については，$c_0 = \langle w\rangle, c_{k+1} = \langle /w\rangle$ とすれば，

$$\begin{aligned}
&\Pr(\langle w\rangle, c_1, \ldots, c_k, \langle /w\rangle) \\
&= \Pr(\langle w\rangle) \prod_{i=1}^{k+1} \Pr(c_i | c_0, \ldots, c_{i-1}) \\
&\simeq \Pr(\langle w\rangle) \prod_{i=1}^{k+1} \Pr(c_i | c_{i-n+1}, \ldots, c_{i-1})
\end{aligned} \tag{2}$$

である．ただし，適当な n により，$\Pr(c_i|c_0, \ldots, c_{i-1})$ を $\Pr(c_i|c_{i-n+1}, \ldots, c_{i-1})$ で近似する．

(1) 式で定義した尺度 $L(A,B)$ は，文字列 $S = AB$の形態素 A と Bへの切れにくさを表している．すなわち，$L(A,B)$ が大きいときには，Sは A と Bには分割されがたい．なぜなら，L の分子は，Sが一つの形態素として (区切り記号が途中に入らずに) 生起する確率であるのに対して，分母は，Sが二つの形態素に分れて独立に生起する確率であるから，L が大きいほど，S が一つの形態素として生起する比が大きいからである．そのため，L が大きいほど分割が誤っている可能性が高い．すなわち，a_kと b_1の間の分割は過分割である可能性が高い．

たとえば，3章の実験によると，「S = 休憩室」のとき，「A = 休」，「B = 憩室」とすると，$\log\Pr(\langle w\rangle, 休, 憩, 室, \langle /w\rangle) = -24.3$，$\log\Pr(\langle w\rangle, 休, \langle /w\rangle) = -12.9$，$\log\Pr(\langle w\rangle, 憩, 室, \langle /w\rangle) = -29.4$ であるので，$L(休, 憩室) = -24.3 + 12.9 + 29.4 = 18.0$ となる．一方，「A = 休憩」，「B = 室」とすると，$\log\Pr(\langle w\rangle, 休, 憩, \langle /w\rangle) = -13.7$，$\log\Pr(\langle w\rangle, 室, \langle /w\rangle) = -10.4$ であるので，$L(休憩, 室) = -24.3 + 13.7 + 10.4 = -0.20$ となる．二つを比較すると，$L(休, 憩室$

3 (1) 式では，形態素の文字列のみを考慮し，品詞は考慮しない．なぜなら，そうすることで，計算が単純になり，かつ，確率推定におけるスパースネスを避けることができるからである．

$) > L$(休憩, 室) であるので，尺度 L によると,「休憩室」については,「休/憩室」という分割の方が「休憩/室」という分割よりも起り難い．すなわち，過分割である可能性が高い．これは我々の言語感覚と一致する．

尺度 L の適用範囲

尺度 L が効果的に検出できるような過分割は，脚注 2 に示した分割誤りのうちで，形態素自体に注目した過分割である．それに対して，語境界交差型における過分割の検出への尺度 L の有効性は，形態素自体に注目した過分割に対するものよりも低いと予想される[4]．なぜなら，問題にしている分割点が語境界交差型の分割誤りの場合，たとえば,「が/まん延」を「がまん/延」と間違えている場合には，L(がまん, 延) を計算するのだが，このとき，$\Pr(\langle w\rangle$, が, ま, ん, 延, $\langle/w\rangle)$ が高い確率値を示し，$\Pr(\langle w\rangle$，が, ま, ん, $\langle/w\rangle)\Pr(\langle w\rangle$, 延 $\langle/w\rangle)$ が低い確率値を示すことは必ずしも期待できないからである．なぜなら，コーパス中で「$\langle w\rangle$がまん延 $\langle/w\rangle$」が頻出し,「$\langle w\rangle$がまん $\langle/w\rangle$」や「$\langle w\rangle$延 $\langle/w\rangle$」が出現しないということは保証できないからである．

ただし，実際上は，(2) 式の確率を推定するときには，$n = 2$ や $n = 3$ により近似するので，(1) 式の計算に関係するのは，分割点の近傍の文字だけになり，その結果，語境界交差型の分割誤りと形態素自体に注目した過分割とで尺度 L の有効性の違いは小さくなると考えられる．

2.3　尺度 L と相互情報量との違い

本節では，尺度 L と，よく知られた統計量である相互情報量 (北, 中村, 永田 1996) との違いを述べる．

形態素 A と形態素 Bの相互情報量 $MI(A, B)$ は，$\Pr(A)$ と $\Pr(B)$ を形態素 A と Bの生起確率とし，$\Pr(A, B)$ を，形態素 A と形態素 Bが，この順番で隣接して生起する確率とすると，

$$MI(A, B) = \log \frac{\Pr(A, B)}{\Pr(A)\Pr(B)} \tag{3}$$

である．次に，(3) 式を，(1) 式と同様に，文字の連鎖として表すと，

$$MI(A, B) = \log \frac{\Pr(\langle w\rangle, a_1, \ldots, a_k, \langle/w\rangle, \langle w\rangle, b_1, \ldots, b_l, \langle/w\rangle)}{\Pr(\langle w\rangle, a_1, \ldots, a_k, \langle/w\rangle)\Pr(\langle w\rangle, b_1, \ldots, b_l, \langle/w\rangle)} \tag{4}$$

となる．$MI(A, B)$ が大きいときには，形態素 A と Bが共起する確率は，それぞれが独立に生起する確率よりも大きいといえるので，$MI(A, B)$ は，形態素 A と形態素 Bとの共起関係の強さ (共起強度) を表す尺度として利用できる．

(1) 式で表される尺度 L と，(4) 式で表される相互情報量とが異なることは，$\Pr(\langle w\rangle, a_1, \ldots, a_k, b_1, \ldots, b_l, \langle/w\rangle) \neq \Pr(\langle w\rangle, a_1, \ldots, a_k, \langle/w\rangle, \langle w\rangle, b_1, \ldots, b_l, \langle/w\rangle)$ であることから分る．また，定性的にいっても，$\Pr(\langle w\rangle, a_1, \ldots, a_k, b_1, \ldots, b_l, \langle/w\rangle)$ は，2.2節で

4 このことは査読者に指摘していただいた．

述べたように，$S = AB$が一つの形態素として生起する確率であるのに対して，$\Pr(\langle w\rangle, a_1, \ldots, a_k, \langle /w\rangle, \langle w\rangle, b_1, \ldots, b_l, \langle /w\rangle)$ は，形態素 A と形態素 Bとが二つの形態素として隣接して生起する確率であるので，これら二つの確率は異なる．

定性的な違いを要約すると，尺度 L は，形態素 $S = AB$が形態素 A と形態素 Bに分割されるときの分割の困難さを表すが，相互情報量は，形態素 A と形態素 Bが隣接して生起するときの生起の容易さを表すと言える．これらは関連していることは確かであるが，基本的には異なる．

なお，3章では，実験の一つとして，尺度 L と相互情報量を含む五つの尺度について，過分割の検出精度を比較する．

3　実験

3.1　実験概要

実験事項

三つの実験を行った．実験 1 では，定性的な評価として，種々の形態素解析システムの解析結果，および，人手修正されたコーパスについて尺度 L を適用し，目視により適用結果を評価した．実験 2 では，訓練コーパスのサイズを変えたときの，尺度 L の過分割検出精度を定量的に評価した．実験 3 では，五つの尺度 (尺度 L/相互情報量/尤度比/改良 Dice 係数 (北村 松本 1997)/Yates 補正されたχ^2) について過分割の検出精度を定量的に比較した．

確率推定の際の設定

教師なし学習/教師あり学習　尺度 L を求めるためには，(2) 式の確率を求める必要があるので，形態素に分割された訓練コーパスが必要である．そのようなコーパスとしては，形態素解析システムにより分割されたコーパスをそのまま用いる場合 (教師なし学習) と，形態素解析結果の誤りを人手で修正したコーパスを用いる場合 (教師あり学習) の二通りが考えられる．そのため，実験 1,2,3 では，この二つの場合について，尺度 L の過分割検出精度などを調べた．

パラメータ推定法　(2) 式の確率を求めるには，n を設定し，かつ，確率推定法も適当に決める必要がある．そのために，本稿では，実験 1 と実験 2 においては，n-gram 確率推定のためのツールとして広く使われている CMU-Cambridge Toolkit(Clarkson and Rosenfeld 1997) を用いて，$n = 3$ の場合について，バックオフスムージングにより推定した．このときのディスカウント法は Witten-Bell discounting(Placeway, Schwartz, Fung, and Nguyen 1993) を用い，カットオフは，文字バイグラムと文字トライグラムの双方で 1 とした[5]．一方，実験 3 においては，

5 CUM-Cambridge Toolkit は，与えられた n-gram である $c_{i-n+1}, \ldots, c_i$の頻度が 0 のとき，その n-gram 確率 $\Pr(c_i|c_{i-n+1}, \ldots, c_{i-1})$ を (n-1)-gram 確率 $\Pr(c_i|c_{i-n+2}, \ldots, c_{i-1})$ から推定する．これをバックオフスムージングという (北他 1996)．バックオフスムージングには，種々の方法があるが，それらは，ディスカウントといって，頻度

最尤推定により求めた確率により尺度 L を計算した．その理由は，尺度 L 以外の尺度においては，通常，最尤推定を用いて，確率を計算しているので，それに合せるためである．また，比較を簡単にするために，$n = 2$ の場合について各種の尺度を比較した．

コーパス

実験 1,2,3 で共通に用いるコーパスは京都大学テキストコーパス version2.0(黒橋他 1998) である．京都大学テキストコーパスは，CD-毎日新聞 95 年度版から約 2 万文を抽出したものであり，形態素・構文解析されている．このコーパスを均等に 2 分割し，実験に用いた．以下では，その一方を京大コーパス A と呼び，他方を京大コーパス B と呼ぶ．京大コーパス A は主に確率推定のための訓練コーパスとして用い，京大コーパス B は主に過分割の検出精度を評価するためのテストコーパスとして用いた．

3.2 実験 1：目視による尺度 L の評価

実験 1 では，定性的な評価として，種々の形態素解析システムの解析結果，および，人手修正されたコーパスについて尺度 L を適用し，目視により適用結果を評価した．

実験材料：コーパスと形態素解析システム

教師なし学習の場合 教師なし学習では，確率推定用の訓練コーパスと過分割検出用のテストコーパスとが同一である．つまり，確率を推定したコーパス中における過分割を検出する．

このときのコーパスとしては，京大コーパス B と EDR 日本語コーパス version 1.5(EDR 電子化辞書研究所 1995) の全文を用いた．なお，EDR 日本語コーパスは，新聞・雑誌・辞典などの流通文書から 1 文単位でとられた約 21 万文からなるコーパスであり，各文は，形態素・構文・意味解析されている．

これらのコーパスにおける生の文を分割する形態素解析システムとしては，公開されている形態素解析システムのうちから，JUMAN version 3.5(黒橋・長尾 1997)，茶筅 version 1.51(松本他 1997), すもも version 1.3(鷲坂他 1997) を用いた．これらの形態素解析システムは，全て，規則に基づいて形態素解析をするものである．なお，これらの形態素解析システムを用いるときには，ただ一つの (ベストの) 解析結果を出力させた．

これらのコーパスと形態素解析システムとの組み合わせは，EDR コーパスに対しては，三つの形態素解析システム全てを適用したが，京大コーパス B については，JUMAN のみを適用し

が 0 より大きい n-gram の頻度から幾らか割引いて，割引いた分を頻度が 0 の n-gram に分け与える方法により特徴付けられる (これにより頻度が 0 の n-gram の確率が 0 より大きくなる)．そのディスカウントの一手法が Witten-Bell discounting である．また，カットオフとは，ある値 x 以下の頻度で生起した n-gram の頻度を 0 として確率を計算する場合の x のことである．カットオフ以下の頻度の n-gram は，頻度が 0 として扱われるが，バックオフスムージングにより 0 より大きい確率が付与される．

た[6]．また，二つのコーパスの元々の分割 (人手修正済みの分割) についても試した．すなわち，全部で 6 種の形態素分割に対して尺度 L を適用した．

教師あり学習の場合　教師あり学習では，確率推定用の訓練コーパスと過分割検出用のテストコーパスとが異なる．

実験 1 では，京大コーパス A の元々の分割 (JUMAN の解析結果を人手修正したもの) を訓練データとして (2) 式の確率を推定した．そして，その推定値を利用して，JUMAN により形態素解析された京大コーパス B に対して尺度 L を適用した．

実験方法

7 種 (=教師なし 6 種+教師あり 1 種) の形態素分割のそれぞれに対して，その全ての分割点について，前後の形態素から尺度 L を計算した．たとえば，「休/憩室/は/広い/。」のように分割されている文については，四つの分割点において，それぞれ，L(休,憩室)，L(憩室,は)，L(は,広い)，L(広い,。) を計算した．このとき，(2) 式の確率は，3.1節で述べたように，$n = 3$ としてバックオフスムージングにより計算した．

実験結果

7 種の形態素分割のそれぞれに対して，全ての分割点を尺度 L により降順に (同一尺度値の場合はランダムに tie を解消して) ソートし，その上位から異なり 150 個を選んだ[7]．そして，それぞれの異なりについて，一個の分割点を無作為に抽出し，それが過分割であるかを判定した[8]．なお，判定は筆者による．

判定した 150 個の分割点について，それが実際に過分割であった数を表 1に示す．表から分かるように，これら 150 個の中に過分割が占める割合は非常に高い．たとえば，表 1では，茶

6 こうした理由は，京大コーパスは主に実験 2,3 における定量的な評価に使用することを意図したものであり，実験 1 では，EDR コーパスを主な対象としたからである．

7 たとえば，「休/憩室/は/広い/。」と「休/憩室/ は/狭い/。」という文があるとき，それぞれの文について，L(休,憩室) が求まるが，この二つの L は同じ文字列を同じように分割しているので，異なりとしては一つである．ただし，二つの分割点を比べたとき，分割点の前後の形態素の字面は同じであっても，品詞が異なる場合には異なる分割点として扱った．すなわち，たとえば，$L(A, B)$ と $L(a, b)$ という二つの分割点があり，字面上は，$A = a$,$B = b$ であったとしても，A と a の品詞が異なるか，Bと b の品詞が異なる場合には，それらの分割点は，異なるものとして扱った．そうした理由は，尺度 L が検出できるのは過分割だけであっても，我々が実際に興味があるのは品詞付けの誤りを含めた形態素解析結果の誤りだからである．

8 過分割かどうかの判定，すなわち，形態素解析システムによる分割点が実際に切って良いかどうかの判定は，複合名詞について困難であるが，もしも，形態素への分割の結果として生じる括弧付けが，筆者の内省に基づいた括弧付けと交差する (cross bracketing) なら，その分割点による分割は誤り (過分割) とする．たとえば，「東南アジアツアー」は，筆者の内省によれば (((東南) アジア) ツアー) という構造をしているので，「東南/アジアツアー」という分割は過分割とする．なぜなら，この分割では，((東南) (アジアツアー)) という構造になるので，括弧が交差するからである．一方，「東南アジア/ツアー」は ((東南アジア) ツアー) という構造なので正解とする．なお，分割の正誤の判定が困難なものについては，品詞を参照し，もし品詞が誤っていたら分割も誤りとした．ただし，上位異なり 150 個については，付録の表 4から表 6にある例と同様に，形態素の途中で分割されているものがほとんどであるので，分割の正誤の判定に迷うような例は少ない．

表 1　上位異なり 150 分割点における過分割の数

学習方法	コーパス	解析システム	過分割の数
教師なし	EDR	人手 (元々の分割)	43
		JUMAN	126
		茶筌	128
		すもも	125
	京大コーパス B	人手 (元々の分割)	49
		JUMAN	98
教師あり	京大コーパス B	JUMAN	125

筅には 128 個の過分割がある．一方，平均的には，茶筌の分割が過分割であるのは，1.5 %以下であると言ってよい[9]．つまり，茶筅の 150 個の分割点のうちで，過分割は，平均的には，$150 \times 0.015 = 2.25$ 個以下である．よって，茶筅の解析結果から 128 個の過分割を検出するためには，平均的には，$(128/2.25) \times 150 \simeq 8533$ 個以上の分割点を調べなければならないことになる．同様なことが，他の形態素解析システムによる分割結果，あるいは，人手で修正された分割結果についても言える．これより，尺度 L を用いることにより，形態素解析結果から過分割を効率的に抽出できるといえる．なお，表 1において，JUMAN で解析された京大コーパス B からの過分割検出結果について，教師なし学習の場合と教師あり学習の場合とを比べると，教師あり学習の方が検出個数が多い．これは，教師あり学習の方が，(2) 式の値を正確に推定できるからであると解釈できる．

さらに，教師なし学習の場合の 6 種の形態素分割のそれぞれについて，上位異なり 150 個中の過分割から上位 12 個の過分割を付録の表 4と表 5に示す．表で「数」とある欄には，そのような過分割を含む文の数がある．また,「形態素/品詞」とある二つの欄は，尺度 L を計算した分割点の前後の形態素と品詞を示す．なお，品詞は，それぞれの形態素解析システムの品詞である．また，表の解析結果は，各解析システムが一つだけ解析結果を出力した場合のものである．もし，複数の解析結果も出力するようにすれば，表中の文について，当該の形態素解析システムが正解を含む解を出すことはある．

これらの表に示されている過分割の中には，何らかの規則性があるとすぐに分るものもある．たとえば，EDR コーパスの元々の分割に含まれる過分割 (表 4) においては,「引き下げ/よう」が「引き下/げ/よう」と分割されていたり,「掲げ/、」が「掲/げ/、」のように分割されるなど，動詞の語幹が分割される例が大半である[10]．一方，EDR コーパスに対する茶筌の過分割では，

9 (Fuchi and Takagi 1998) によると，茶筌の EDR コーパスにおける形態素解析結果の適合率 (=100×(茶筌の形態素解析結果の形態素で正解と一致したものの数/茶筌の形態素解析結果の形態素の総数)) は，字面が一致していた場合を一致とすると，98.5 %である．ここで，形態素の適合率は，実験 2 でも示すように，分割点の適合率よりも低くなる．なぜなら，形態素が一致するためには，その前後の分割点も一致しなくてはならないため，形態素が一致するというのは，分割点の一致よりも厳しい条件であるからである．そのため，分割点が過分割であるのは 1.5 %以下と言って良い．

10 EDR コーパスの元々の分割においては,「掲げ、」という文字列を含む文が 14 例あるが，そのうち表 4の 1 例のみが,「掲/げ/、」という分割であり，その他の 13 例は,「掲げ/、」という分割である．このことは「掲/げ/、」が過分割である

「結果」が「結 (普通名詞) /果 (普通名詞)」と分割されていたり,「考えて」が「考 (普通名詞)/えて (普通名詞)」と分割されているが，このような例に含まれる規則性は，もしあったとしても，容易には分らない．

いずれにしろ，尺度 L を使うことにより，ある程度の量の，形態素解析結果の過分割が，教師なし学習により容易に抽出できることが分かる．このような例を集めるのは人手では手間が掛る．また，尺度 L は人手修正後のコーパスに残る過分割も検出できるため，コーパス作成・整備の際の補助ツールとしても役立つと考える．

また，付録の表 6には，教師あり学習の場合について，尺度 L の値が上位 12 個の過分割を示す．ここで，教師あり学習の結果である表 6における JUMAN の過分割と，教師なし学習の結果である表 5における JUMAN の過分割とを比べると，表 6においては「護/熙」など固有名詞が占める割合が多いが，表 5では固有名詞は一つ (「若/乃/花」) しか存在しないことがわかる．表 5に固有名詞が少ないのは，固有名詞は未知語である場合が他の品詞と比べて多いため，常に過分割される場合も多くなり，その結果として尺度 L の値が小さくなる場合が多いためである．このように，形態素解析システムが常に過分割してしまうような場合を検出するためには，人手修正済みコーパスが必要であると言える．

3.3　実験 2：過分割検出精度の定量的評価

教師なし学習により何か統計的に興味のある言語現象を発見するような応用 (新納 井佐原 1995; 池原, 白井, 河岡 1995; 下畑, 杉尾, 永田 1995; 久光 丹羽 1997, など) においては，新聞記事などの大規模なコーパスが比較的用意に入手できるので，訓練コーパスのサイズは深刻な問題ではない．これは本稿における過分割検出の場合でも同様である．しかし，教師あり学習の場合には，訓練コーパスを構築するのはコストが掛るため，なるべく小さな訓練コーパスであることが望ましい．そこで，実験 2 では，主に教師あり学習の場合を対象として，訓練コーパスのサイズと過分割検出精度との関係を調べた．ただし，教師なし学習の場合についても，教師あり学習と比較するために，訓練コーパスのサイズと過分割検出精度との関係を同様に調べた．

実験材料：コーパス

確率推定用の訓練コーパスとしては京大コーパス A を用い，過分割検出の精度を調べるテストコーパスとしては JUMAN により分割された京大コーパス B を用いた．このことは，教師あり学習と教師なし学習とで共通である．ただし，教師あり学習では京大コーパス A の元々の分割から (2) 式の確率を推定し，教師なし学習では，京大コーパス A を JUMAN により形態素解

ことを傍証している．これと同様なことが，表 4のその他の例についても言える．なお,「掲げ/、」という分割を含む例には,「虹/を/描/い/た/旗/を/掲げ/、/高らか/に/歌/う/。」や「独特/の/理想/を/掲げ/、/実行/し/た/人/だっ/た/。」のような例がある．

析した結果から (2) 式の確率を推定した[11].

テストコーパスの各種統計　テストコーパスである京大コーパス B について，その元々の分割を正解と看做して[12]，分割の正誤を判定したときの統計を表 2に示す.

表 2　京大コーパス B における分割点についての統計

正解における分割点の数	232572
JUMAN による分割点の数	233048
一致した分割点の数	231816
分割点の再現率	99.7 %
分割点の適合率	99.5 %
過分割の数	1232
分割不足の数	756
分割の間違いの数 (過分割の数+分割不足の数)	1988
100×(過分割の数/分割の間違いの数)	62.0 %
100×(過分割の数/JUMAN による分割点の数)	0.5 %

表 2より，分割の間違いに占める過分割は 62.0 %である．加えて，過分割の周辺には分割不足も起りやすいと言えるので，過分割が検出できれば，その周囲も調べることにより，分割誤りの多くが検出できると言える.

しかし，分割点の再現率 (=100×(一致した分割点の数/正解における分割点の数)) と適合率 (=100×(一致した分割点の数/JUMAN による分割点の数)) は，それぞれ，99.7 %，99.5 %と非常に高い[13]．また，JUMAN の分割点全体の中で過分割である分割点は 0.5 % (=100 %−適合率) であるので，過分割を見付けるのは人手では困難であると考える.

実験方法

約 1 万文からなる訓練コーパスから，約 1000，2000, ..., 10000 文を選び，それぞれの場合について，$n = 3$ としてバックオフスムージングにより (2) 式の確率を推定し，それを利用して約 1 万文からなるテストコーパスにおける全分割点の尺度 L の値を計算した．そして，全ての分割点を尺度 L により降順にソートし，上位の分割点から，過分割かどうかを，テストコーパスの元々の分割を正解として調べた.

11 教師なし学習において，京大コーパス A を訓練コーパスにした場合と，京大コーパス B を JUMAN により形態素解析した結果を訓練コーパスとした (訓練コーパスとテストコーパスが同一の) 場合とでは，(後述する (5) 式で定義する分割点調査率の意味における) 過分割検出精度は，ほぼ等しい.

12 実験 1 で見付けた過分割についても修正はしていない.

13 参考のため，(Nagata 1994) の基準による，形態素の再現率 (=100×(一致した形態素の数/正解における形態素の数)) と適合率 (=100×(一致した形態素の数/JUMAN による形態素の数)) を求めると，それぞれ，99.1 %と 98.9 %になる (ただし，字面が一致していれば形態素が一致したと看做す)．これらからも分るように，形態素の再現率と適合率とは分割点のものに比べて低い.

実験結果

まず，全訓練データを使用した場合についての実験結果を述べ，次に訓練データを 1000 文ずつ増加した場合についての実験結果を述べる．

全訓練データを使用した場合　図 1には，約 1 万文の訓練データ全てを使って確率推定した場合について，教師あり学習と教師なし学習のそれぞれについて，過分割検出の再現率 (percent recall) に対する適合率 (percent precision) および分割点調査率 (percent examination) を示す．ここで，

$$\begin{aligned}
\text{再現率} &= 100 \times \frac{\text{検出された過分割の数}}{\text{テストコーパスにおける過分割の数}}, \\
\text{適合率} &= 100 \times \frac{\text{検出された過分割の数}}{\text{尺度 } L \text{ の上位から順番に調べた分割点の数}}, \\
\text{分割点調査率} &= 100 \times \frac{\text{尺度 } L \text{ の上位から順番に調べた分割点の数}}{\text{テストコーパスにおける全分割点の数}}.
\end{aligned} \tag{5}$$

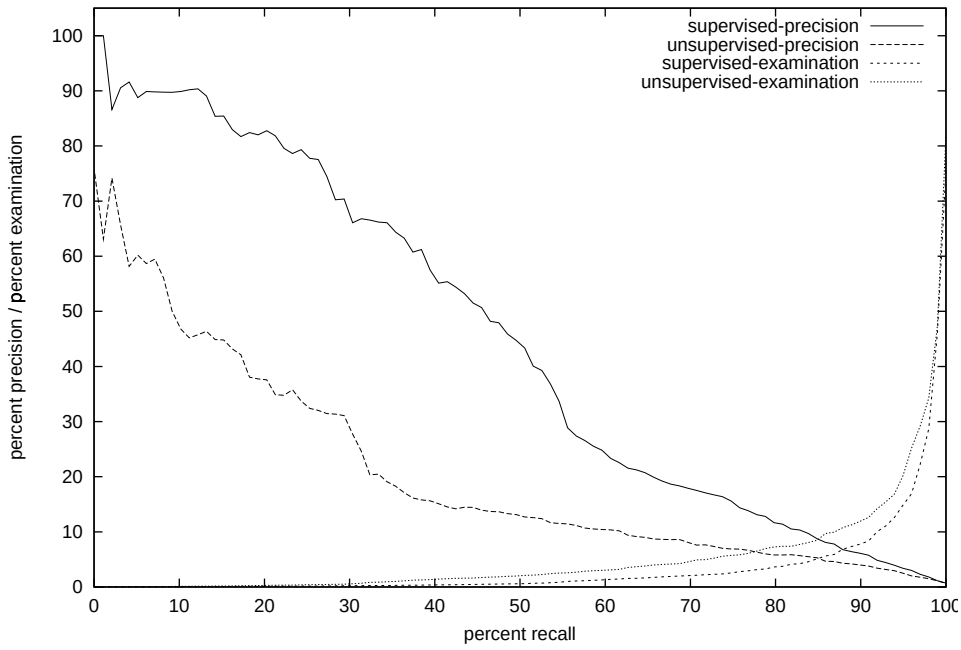

図 1　再現率と適合率/分割点調査率

図 1の 教師あり学習の場合の適合率 (supervised-precision) および教師なし学習の場合の適合率 (unsupervised-precision) のプロットから分かるように，上位における過分割検出の適合率は

非常に高い．たとえば，再現率が 10.0 %のとき，適合率は，教師あり学習の場合に 90.5 %であり，教師なし学習の場合に 46.8 %であるが，これらは，JUMAN の分割点全体の中で過分割が占めるパーセンテージである 0.5 %の，180 倍以上，および，90 倍以上である．この適合率の高さは，実験 1 での結果を裏付けるものである．

また，図 1の教師あり学習の場合の分割点調査率 (supervised-examination) および教師なし学習の場合の分割点調査率 (unsupervised-examination) から分かるように，一部の分割点を調べるだけで多くの過分割を検出できると言える．たとえば，全体の過分割のなかから再現率 50 %で過分割を見付けるためには，教師あり学習の場合には，全分割点の 0.5 %を調べればよく，教師なし学習の場合には，全分割点の 2.0 %を調べればよい．さらに，90 %の過分割を見付けるためには，教師あり学習の場合には，全分割点の 7.8 %を調べればよく，教師なし学習の場合には，全分割点の 12.2 %を調べればよい．一方，もし，無作為に分割点を調べるという方法により，過分割を検出しようとしたならば，50 %の過分割を見付けるためには，平均的には，全分割点の 50 %を調べる必要があり，90 %の過分割を見付けるためには，90 %の分割点を調べる必要がある．

以上より，尺度 L を使うことにより，過分割の検出が効率良くできると言える．

なお，再現率，適合率，分割点調査率の間には

$$\text{適合率} = K \times \frac{\text{再現率}}{\text{分割点調査率}}. \tag{6}$$

という関係が成立する．ただし，Kはテストコーパスに固有の定数であり，

$$K = 100 \times \frac{\text{テストコーパスにおける過分割の数}}{\text{テストコーパスにおける全分割点の数}}.$$

(6) 式から，分割点調査率と再現率が決まれば適合率が決まることが分かる (e.g., 分割点調査率が小さければ適合率は高い)．そのため，以下では，再現率に対する分割点調査率のみに基づいて過分割の検出精度を評価する．そして，同一の再現率に対して分割点調査率が小さいとき過分割の検出精度が高いと言い，その逆のときに過分割の検出精度が低いと言うことにする．

1000 文ずつ訓練データを増やした場合　図 2には，過分割検出の再現率が 25,50,75 %の場合 (recall25,recall50,recall75) について，教師あり学習の場合と教師なし学習の場合における，訓練文数 (Num. of training sentences) と分割点調査率の関係を示す．

図 2から，教師あり学習の場合 (supervised-recall25,50,75) については，訓練文数が増加すると，再現率が 50 %と 75 %においては，分割点調査率が明確に減少していると言える．また，再現率が 25 %についても緩やかに分割点調査率は減少している．一方，教師なし学習の場合 (unsupervised-recall25,50,75) については，訓練文数が増えていっても，2000 文以上については，分割点調査率は (若干の変動はあるが) ほぼ横ばいである．

このことは，教師あり学習については，訓練データが多くなれば多くなるだけ，(2) 式の確率

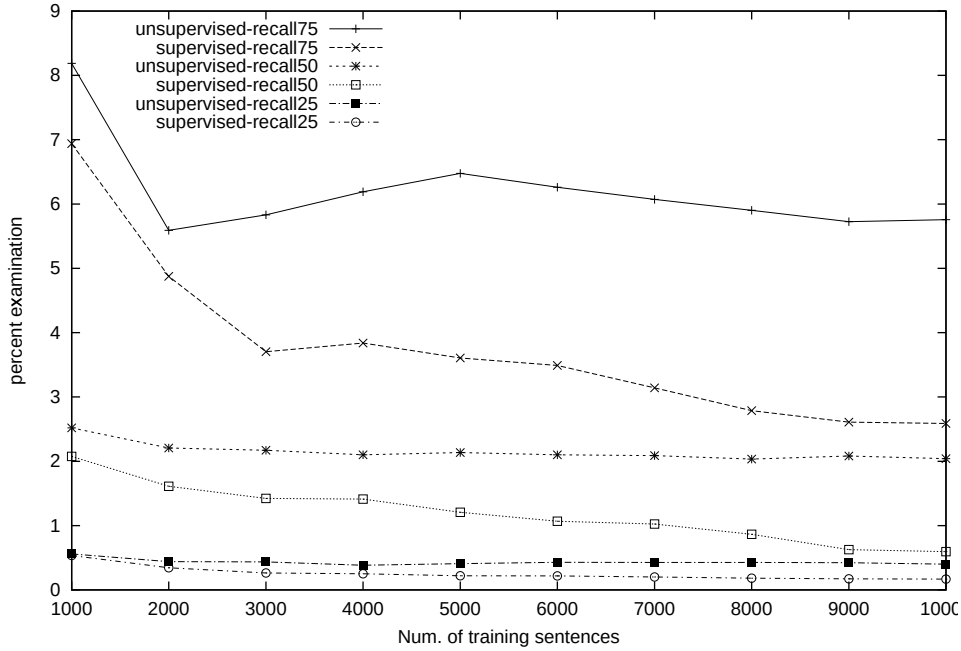

図 2　訓練データを増やした場合の再現率と分割点調査率

を精密に推定できるが，教師なし学習については，訓練データが多くなったとしても，その確率推定に対する効果は，教師あり学習の場合に比べれば，小さいことを示している．

3.4　実験 3：各種尺度の比較

実験 3 では，尺度 L，相互情報量，尤度比，改良 Dice 係数 (北村・松本 1997)，Yates 補正されたχ^2，の五つの尺度について，過分割の検出精度を比較した．ここで，尤度比，改良 Dice 係数，Yates 補正されたχ^2は，(久光・丹羽 1997) において，有用な括弧表現を抽出するために有効であるとされた尺度である．また，尤度比は，(影浦 1997) でも，2 文字間の連関の尺度として，漢字列の分割に有効であることが示されている．

以下では，まず，本実験のテストコーパスとした京大コーパス B について，そこでの分割点の出現頻度の統計について述べる．この出現統計は，あとで，各尺度間の過分割検出精度の違いを説明するときの資料に用いる．次に，各尺度を定義し比較する．

テストコーパスにおける分割点の出現統計

形態素 A の最後の文字を a, 形態素 Bの最初の文字を b とし，a と b に挟まれるような分割点を，前後 1 文字で区別される分割点と呼ぶ．実験 3 では，分割点といえば，前後 1 文字で区別される分割点のこととする．つまり,「ab/cd」と「xb/cy」のような分割点は，分割点の前後 1 文字が同じであるので，区別しないで同一タイプの分割点として扱う．

表 3は，テストコーパスとした京大コーパス B における分割点について，過分割である分割点とそうでない分割点のそれぞれに対して，出現頻度ごとの，分割点の異なり数などを調べたものである．ここで，分割点の総数を F, 頻度 rにおける分割点の異なり数を k_rとすると，頻度 rにおける延べ数は $f_r = r \times k_r$であり，$F = \sum_r f_r$である．表 3では，頻度 rにおける「延べ%」は $100 \times f_r/F$であり,「累積%」は$\sum_{s=1}^{r} 100 \times f_s/F$である．

表 3　京大コーパス B における分割点の出現統計

頻度	過分割である分割点			過分割でない分割点		
	異なり数	延べ%	累積%	異なり数	延べ%	累積%
1	645	52.4	52.4	24180	10.4	10.4
2	104	16.9	69.2	7187	6.2	16.6
3	29	7.1	76.3	3579	4.6	21.3
4	11	3.6	79.9	2133	3.7	24.9
5	9	3.7	83.5	1413	3.0	28.0
6	5	2.4	86.0	1062	2.7	30.7
7	2	1.1	87.1	764	2.3	33.0
8	3	1.9	89.0	644	2.2	35.3
9	1	0.7	89.8	496	1.9	37.2
10	1	0.8	90.6	411	1.8	39.0
11 以上	8	9.4	100.0	3586	61.0	100.0

表 3から，過分割である分割点の出現頻度は，そうでない場合に比べて，低頻度であると言える．これは，過分割である分割点の数自体が少ないことが主な原因である．また，過分割である場合とそうでない場合の分布の様子を比べると，過分割である分割点の場合には，頻度が 1 か 2 であるような場合が全体の 50 %以上を占めていることから分かるように，低頻度の方に分布が偏っている．

各尺度の定義

まず，$n = 2$ として，(2) 式を用いて，(1) 式を変形すると，

$$L(A, B) = \log \frac{\Pr(a_k, b_1)}{\Pr(a_k, \langle / w \rangle) \Pr(\langle w \rangle, b_1)} \tag{7}$$

となる．一方，(3) 式を同様に変形すると

$$MI(A,B) = \log \frac{\Pr(\langle /w\rangle, \langle w\rangle)}{\Pr(\langle /w\rangle)\Pr(\langle w\rangle)}$$

という無意味な値になるので，区切り文字とそれに隣接する文字は特に強く結合すると仮定し，

$$\Pr(\langle w\rangle, c_1, \ldots, c_k, \langle /w\rangle) = \Pr(\langle w\rangle, c_1)\Pr(c_2|\langle w\rangle, c_1)\cdots\Pr(\langle /w\rangle, c_k|c_{k-1})$$

のような変形をすると，

$$MI(A,B) = \log \frac{\Pr(a_k, \langle /w\rangle, \langle w\rangle, b_1)}{\Pr(a_k, \langle /w\rangle)\Pr(\langle w\rangle, b_1)} \tag{8}$$

となる．なお，以下では，$a = a_k$,$b = b_1$とする．

(8) 式から，$n = 2$ においては，相互情報量 $MI(A,B)$ は，$a\langle /w\rangle$ と $\langle w\rangle b$ をそれぞれ一つの項と看做せば，この 2 項に関する通常の相互情報量の式と一致することがわかる．そこで，尤度比，改良 Dice 係数，Yates 補正されたχ^2についても，これら 2 項に基づいて，その値を計算する．

以下では，(久光・丹羽 1997) に基づいて，尤度比，改良 Dice 係数，Yates 補正されたχ^2を定義する．また，尺度 L と相互情報量についても，確率を最尤推定した形で定義する．

各尺度を定義する準備として，まず，$f_{ij}(i,j = 1,2)$ は，分割表で示すと

	後続文字が $\langle w\rangle b$	後続文字が $\langle w\rangle b$ 以外
先行文字が $a\langle /w\rangle$	f_{11}	f_{12}
先行文字が $a\langle /w\rangle$ 以外	f_{21}	f_{22}

である．より厳密には，$f(\cdots)$ を文字列の頻度とし，v, w, x, yを，$\langle w\rangle$ と $\langle /w\rangle$ を含む任意の文字としたとき，

$$\begin{aligned} f_{11} &= f(a, \langle /w\rangle, \langle w\rangle, b) \\ f_{12} &= \textstyle\sum_{xy\neq\langle w\rangle b} f(a, \langle /w\rangle, x, y) \\ f_{21} &= \textstyle\sum_{vw\neq a\langle /w\rangle} f(v, w, \langle w\rangle, b) \\ f_{22} &= \textstyle\sum_{vwxy} f(v,w,x,y) - f_{11} - f_{12} - f_{21} \end{aligned}$$

である．また，

$$\begin{aligned} f_{i.} &= f_{i1} + f_{i2} \\ f_{.j} &= f_{1j} + f_{2j} \\ F &= \textstyle\sum_{i,j} f_{ij} \end{aligned}$$

である．

尤度比　ここでの「尤度比」は，$a\langle /w\rangle$ と $\langle w\rangle b$ の 2 項が従属とした場合と独立とした場合との最尤推定量による尤度比であり，

$$\lambda = 2\sum_{i,j} f_{ij}\left\{\log\frac{f_{ij}}{F} - \log\frac{f_{i.}f_{.j}}{F^2}\right\} \tag{9}$$

である．なお，上式では，分割点のソートに無関係な項は除いてある．

λは，2 項が従属して生起する度合が強いとき，正で大きな値をとる．しかし，これだけでは必ずしも共起強度が強いとは言えない．たとえば，

10	1
1	10

と

1	10
10	1

は同じλとなる．これらのうち前者は共起強度が強いが，後者は弱い (反発している)．このことを考慮して，$\lambda > 0$ のときには，(影浦 1997) と同様に，Yule の $Y(= \frac{\sqrt{f_{11}f_{22}}-\sqrt{f_{12}f_{21}}}{\sqrt{f_{11}f_{22}}+\sqrt{f_{12}f_{21}}})$ の符合を付けることにより，分割点をソートした．

Yates 補正されたχ^2　λと同様に独立性の判定に用いられる尺度である．

$$\chi^2 = \frac{F(|f_{11}f_{22} - f_{12}f_{21}| - F/2)^2}{f_{1.}f_{2.}f_{.1}f_{.2}} \tag{10}$$

なお，χ^2に関しても，λと同様な理由から，Yule の Yの符合を付けて分割点をソートした．

改良 Dice 係数　(北村・松本 1997) で，対訳単語間の類似度として，提案されている尺度である．

$$\text{改良 Dice 係数} = (\log f_{11})\frac{2f_{11}}{f_{1.} + f_{.1}}. \tag{11}$$

相互情報量　(8) 式より，分割点のソートに無関係な項は除くと，

$$MI' = \log \frac{f_{11}}{f_{1.}f_{.1}}. \tag{12}$$

尺度 L　(7) 式より，分割点のソートに無関係な項は除くと，

$$L' = \log \frac{f(a,b)}{f_{1.}f_{.1}} \tag{13}$$

ここで，上記の各尺度について，もし，$f_{ij} = 0$，あるいは，$f(a,b) = 0$ となる場合には，それぞれを 0.1 として計算した．

教師なし学習の場合での各尺度の比較

各尺度について，JUMAN により形態素解析された京大コーパス B を訓練およびテストコーパスとして，過分割の再現率に対する分割点調査率を評価した結果を図 3に示す．

図 3から分るように，改良 Dice 係数 (Dice)，λ(lambda)，χ^2(chi ^ 2) の分割点調査率は，L' や MI'と比べて大きい．すなわち，過分割検出精度は低い．この原因は，これらの尺度が，統計的に有意と言えないような低頻度の共起関係をノイズとして排除するような尺度であるからである．すなわち，頻度が 1 とか 2 とかの共起関係の尺度値は，これらの尺度では大きくならない[14]ため，(表 3に示されるように) 低頻度である過分割が排除されるためである．

14 特に，改良 Dice 係数では，頻度が 1 の共起関係については，値が 0 となる

このような性質は，(久光・丹羽 1997) や (影浦 1997) や (北村・松本 1997) のような，一般的に共起強度が高い共起関係を必要とするような応用に対しては適していたが，低頻度事象である過分割を検出するには適さない．

一方，MI'の過分割検出精度は，再現率 50 %程度のところまでは，L'とほぼ同じである (実際には若干低い)．これは，MI'が低頻度の共起関係を過大評価する (久光・丹羽 1997) からであろう．つまり，再現率が低いところでは，低頻度で，かつ，共起強度の強い表現を選択的に拾ってくるが，そのようなものは過分割であることが多いため，検出精度が高いと解釈できる．しかし，再現率が上ってくると，比較的頻度が高い過分割も増えてくるため，共起強度だけでは，過分割なのか，そうでない分割かが区別できなくなり，検出精度が下がると言える．

これらの尺度に対して，L'は，分割されるか分割されないかを直接モデル化した尺度であるため，再現率が高くなっても検出精度が高いものと考える．

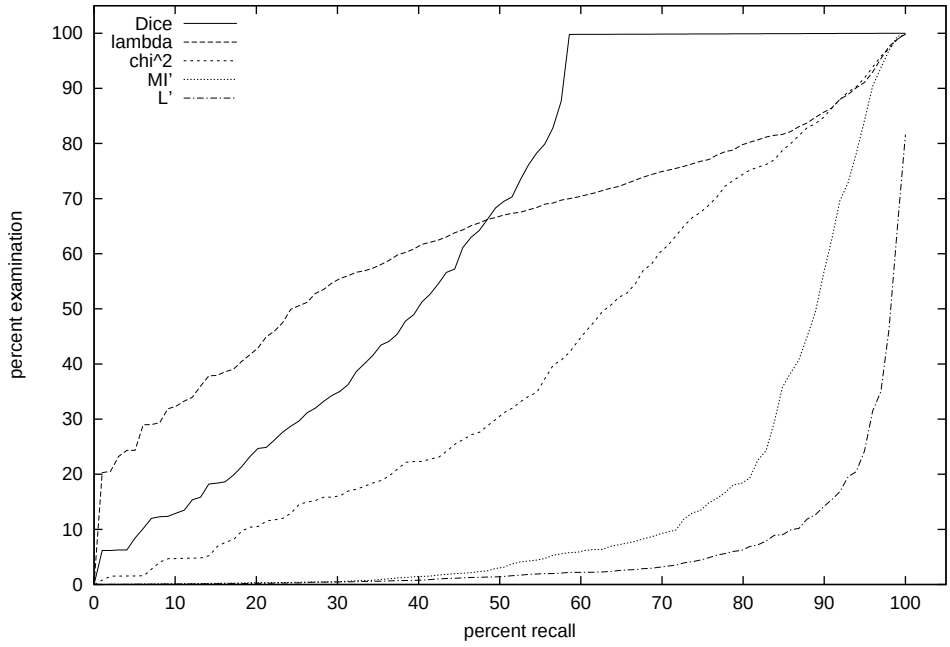

図 3　過分割の再現率と分割点調査率 (教師なし学習)

なお，筆者は，予備実験として，相互情報量と Yate 補正したχ^2を，隣接する形態素間について，(文字ではなく) 形態素を単位とする 2 項関係に基づいて計算してみたが，その性質は尺度 L とは非常に異なっていた．相互情報量の性質と Yate 補正したχ^2の性質とは，互いに若干は異なるが，おおまかには，二つの尺度とも，固有名詞 (「福沢/諭吉」など) や四字熟語 (「不

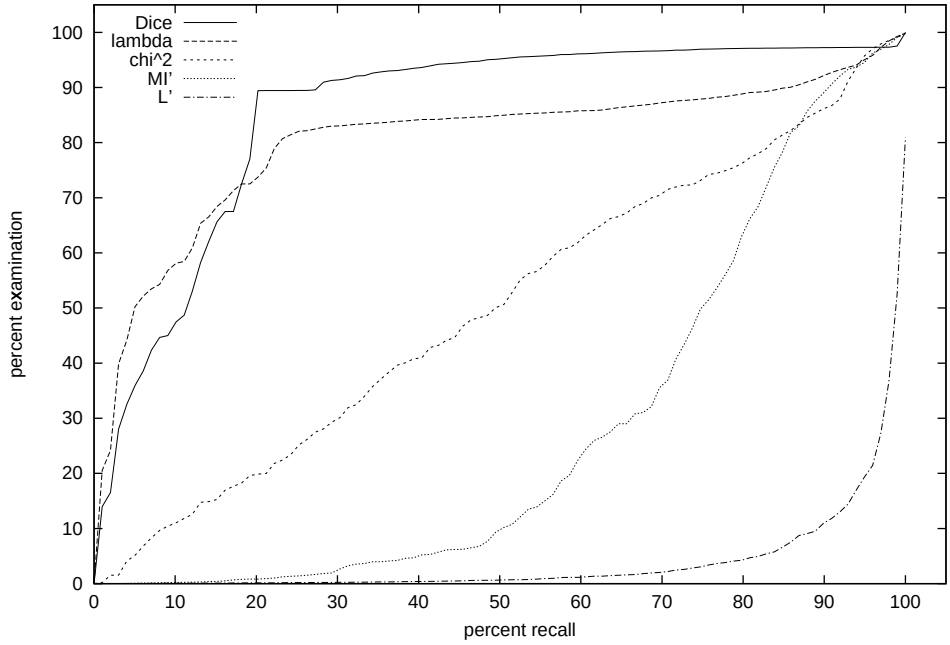

図 4　過分割の再現率と分割点調査率 (教師あり学習)

眠/不休」など) を取ってくる傾向が強かった．これらの隣接形態素は，それ自体は有用な表現ではあるが，これらの隣接形態素間の分割が間違っているわけではないので，本稿での目的である過分割の検出には適さない．

その他，共起や定型表現を抽出する研究として，特に文字列レベルに関係するものでは，(新納・井佐原 1995; 池原他 1995; 下畑他 1995, など) がある．これらの研究では，大量の生テキストコーパスから，統計量を用いることにより,「に関して」や「に対しては」などの定型的な表現を抽出する．これらの表現は有用な表現ではあるが,「に対して」を「に/対/し/て」と分割しても，過分割ではないことからも分るように，これらの手法は，本稿での目的である過分割の検出には適さない．

教師あり学習の場合での各尺度の比較

各尺度に対して，京大コーパス A の元々の分割を訓練コーパス，京大コーパス B をテストコーパスとして，過分割の再現率に対する分割点率調査を評価した結果を図 4に示す．

図 4では，図 3と同様に尺度 L の分割点調査率が一番小さい．そして，図 4と図 3 を比べると，尺度 L については，図 4の教師あり学習の方が図 3の教師なし学習の場合よりも分割点調

査率が小さい．

一方，尺度 L 以外の尺度については，教師あり学習の方が分割点調査率は大きくなっている．これは，教師あり学習の場合の方が，教師なし学習の場合よりも，テストコーパスにおいて，過分割の前後の文字の共起強度が小さいことを示している．この理由は，教師あり学習においては，訓練コーパスで過分割であるような分割点が人手により除かれているため，テストコーパスで過分割であるような分割点は訓練コーパスで出現することが稀となり，その結果，共起強度が小さくなるからである．このことから，尺度 L 以外の尺度については，教師あり学習をしても過分割検出精度が高くならないことが分かる．

4　考察と今後の課題

確率推定法と尺度 L

本稿の実験では，(2) 式の確率を推定するために，3.1節に述べたような n の値と確率推定法を用いたが，確率推定の方法には，最尤推定やバックオフスムージングの他にも様々な方法があり，さらに，バックオフスムージングについても様々な discounting の方法があるので，これらを適用した場合の尺度 L の過分割検出精度について網羅的に調べることを今後の課題としたい．

本稿でこのことを網羅的に調べなかった理由は，本稿での主要な目的は，1章で述べたように，従来の研究で人手で発見されることが前提となっていた過分割を，尺度 L を用いることにより半自動的に抽出できることを示すことにあったので，そのことを示すためには，なんらかの (代表的な) 確率推定法を利用した場合について示すだけで十分であったからである．すなわち，確率推定法の優劣 (尺度 L と併用したときの過分割検出精度の良否) を調べることは副次的な事柄であったからである．

なお，予備実験として，京大コーパスを利用し，

- n=2，または，n=3
- Witten-Bell discounting によるバックオフスムージング，または，最尤推定

から作られる四つの組合せのそれぞれについて (2) 式の確率を推定し，尺度 L による分割点調査率を調べた結果は，教師あり学習と教師なし学習の双方について，$n = 2$ に最尤推定を組み合わせたものと $n = 3$ にバックオフスムージングを組み合わせたものとがほぼ等しく良く (教師なし学習では前者が若干良く，教師あり学習では後者が若干良い)，その他の組み合わせ ($n = 3$ と最尤推定および $n = 2$ とバックオフスムージング) は，この二つよりも劣っていた．このような結果の原因としては，n や確率推定法の違いの他に，訓練データのサイズが約 1 万文と比較的少ないことが影響していると考えられる．

確率に基づく形態素解析システムへの適用

確率に基づいた形態素解析システム (あるいは単語分割システム) には，文字の連鎖確率に基づいたシステム (山本・増山 1997; 小田 北 1998) と単語や品詞の連鎖確率に基づいたシステム (Nagata 1994; 伊藤 西村 1997; 森・長尾 1998, など) がある．

(山本・増山 1997; 小田・北 1998) では，本稿とは実現手法は異なるが，形態素境界の情報を文字に取り込むことにより，文字列により形態素列を表現している．そして，入力文に対して，(形態素境界情報を含む) 文字の連鎖確率が最大になるような解を求めることにより，最適な形態素列を得ている．一方，(Nagata 1994; 伊藤・西村 1997; 森・長尾 1998, など) では，単語や品詞 n-gram に基づいて形態素解析をしており，文字情報を直接用いているのは未知語モデルに限定されている．

これらの形態素解析システムの解析結果からも尺度 L が過分割を検出できるかを調べることは今後の課題であるが，(山本・増山 1997; 小田・北 1998) と (Nagata 1994; 伊藤・西村 1997; 森・長尾 1998, など) を比べた場合，前者は，文字の連鎖確率を直接用いて形態素列への分割を行なっている点が，尺度 L と極めて類似しているため，前者に尺度 L を適用した場合の過分割検出精度は，(文字レベルでの分割の最適化を行なっていない) 後者に適用した場合と比較して劣ることが予想される．しかし，(Nagata 1994; 伊藤・西村 1997; 森・長尾 1998, など) の品詞や単語の n-gram に基づくシステムに尺度 L を適用した場合についても，規則に基づく形態素解析システムに比較すれば，最適な形態素列を求めるときに，可能な分割を相互に比較し最高確率のものを出すという形で，尺度 L に用いた情報が既に用いられているとも言えるため，尺度 L の有効性は低いと予想される．

分割不足の検出

筆者は，予備実験として，実験 1,2 と同様の確率推定法で，JUMAN により解析された京大コーパス B 全体を訓練およびテストコーパスとして，教師なし学習での確率推定値を用い，尺度 L により分割不足の検出を試みた．つまり，尺度 L の値が小さい位置が形態素として結合されている場合について，それが実際に分割不足かを確かめた．

その結果は，分割不足の再現率が 10 %の時点で，既に適合率が 4 %であり，実験 2 における，再現率が 10 %のときの適合率が 47 %と比べて非常に劣っていた．

その理由の一つは，形態素解析システム中の形態素に比較的長い単位が多く，かつ，分割不足として抽出されたものの多くが，その長い単位の形態素を短い単位に分割しようとしているためである．たとえば，分割不足として抽出されたものの上位には，'/' が候補位置とすると，「穴を/あけた」「目を/見張る」「あっという/間」などがある．

このような場合と，明確に間違いである分割不足とを区別することは尺度 L には不可能なので，尺度 L により分割不足を検出するのは，過分割の場合ほどには上手くいかない．

5　おわりに

本稿では，形態素解析結果から過分割を検出する統計的尺度を提案した．その尺度は，文字列に関する尺度であり，文字列が分割される確率と分割されない確率との比に基づいていて，分割されにくい文字列ほど大きな値となる．したがって，この値が大きい文字列は過分割されている可能性が高い．

提案尺度を使うことにより，規則に基づいた形態素解析システムの解析結果から高精度で過分割を検出できたし，人手で修正されたコーパスに残る過分割も検出できた．また，提案尺度の過分割検出精度は，その他の統計的尺度と比べて高かった．これらのことは，提案尺度が，形態素解析システムの高精度化に役立つこと，及び，コーパス作成・整備の補助ツールとして役立つことを示している．

今後は，提案尺度を実際に使い，形態素解析システムの精度向上やコーパスの整備に役立てたい．

謝辞

本稿に対して有益なコメントを下さった筑波大学山本幹雄助教授，および，日頃議論して下さる信州大学音声信号処理研究室の各位に感謝する．本稿では，一般に公開されている，JUMAN，茶筌，すもも，京都大学テキストコーパス，CMU-Cambridge Toolkit を利用させていただいた．このことに対して関係者の方々に感謝する．

参考文献

Clarkson, P. and Rosenfeld, R. (1997). “Statistical Language Modeling using the CMU-Cambridge Toolkit.” In *Eurospeech'97*, pp. 2707–2710.

EDR 電子化辞書研究所 (1995). “EDR 電子化辞書マニュアル.”.

Fuchi, T. and Takagi, S. (1998). “Japanese Morphological Analyzer using Word Co-occurrence – JTAG –.” In *COLING'98*, pp. 409–413.

久光徹 丹羽芳樹 (1997). “統計量とルールを組み合わせて有用な括弧表現を抽出する手法.” 情報処理学会自然言語処理研究会研究報告, 122-17, pp.113–118.

久光徹 丹羽芳樹 (1998). “書き換え規則と文脈情報を用いた形態素解析後処理.” 情報処理学会自然言語処理研究会研究報告, 126-8, pp.55–62.

池原悟, 白井諭, 河岡司 (1995). “大規模日本語コーパスからの連鎖型および離散型の共起表現の自動抽出法.” 情報処理学会論文誌, **36** (11), 2584–2596.

伊藤伸泰 西村雅史 (1997). “N-gram を用いた日本語テキストの単語単位への分割.” 情報処理学会自然言語処理研究会研究報告, 122-9, pp.57–62.

影浦峡 (1997). “文字単位の bigram 尺度に基づく複合漢字列の単位切り手法.” 言語処理学会第 3 回年次大会発表論文集, pp.477–480.

北研二, 中村哲, 永田昌明 (1996). 音声言語処理. 森北出版.

北村実穂子 松本裕治 (1997). “対訳コーパスを利用した対訳表現の自動抽出.” 情報処理学会論文誌, **38** (4), 727–735.

北内啓, 宇津呂武仁, 松本裕治 (1998). “誤り駆動型の確率モデル学習による日本語形態素解析.” 情報処理学会自然言語処理研究会研究報告,124-6,pp.41–48.

小松英二 (1998). “解析誤りデータを用いたコスト最小法形態素解析のコスト関数の構成法.” 情報処理学会自然言語処理研究会研究報告,123-2,pp.9–16.

黒橋禎夫 長尾真 (1997). “日本語形態素解析システム JUMAN version 3.5.” 京都大学大学院工学研究科.

黒橋禎夫, 斎藤由衣子, 坂口昌子 (1998). “コーパス作成の作業基準 version 1.6.” 京都大学.

松本裕治, 北内啓, 山下達雄, 平野善隆, 今一修, 今村友明 (1997). “日本語形態素解析システム『茶筅』 version 1.5 使用説明書.” 奈良先端科学技術大学院大学.

森信介 長尾眞 (1998). “形態素クラスタリングによる形態素解析精度の向上.” 自然言語処理, **5** (2), 75–103.

Nagata, M. (1994). “A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A^* N-Best Search Algorithm.” In *COLING'94*, pp. 201–207.

小田裕樹 北研二 (1998). “PPM*モデルによる日本語単語分割.” 情報処理学会自然言語処理研究会研究報告, 128-2, pp.9–16.

Placeway, P., Schwartz, R., Fung, P., and Nguyen, L. (1993). “The Estimation of Powerful Language Models from Small and Large Corpora.” In *ICASSP-93*, pp. II–33–36.

下畑さより, 杉尾俊之, 永田淳次 t (1995). “隣接文字の分散値を用いた定型表現の自動抽出.” 情報処理学会自然言語処理研究会研究報告, 110-11, pp.71–78.

新納裕幸 井佐原均 (1995). “擬似 N グラムを用いた助詞的定型表現の自動抽出.” 情報処理学会論文誌, **36** (1), 32–40.

鷲坂光一, 山崎憲一, 廣津登志夫, 尾内理紀夫 (1997). “情報検索のための高速日本語形態素解析システム「すもも」.” 情報処理学会第 54 回全国大会，第 2 分冊，pp.59–60.

山地治, 黒橋禎夫, 長尾眞 (1996). “連語登録による形態素解析システム JUMAN の精度向上.” 言語処理学会第 2 回年次大会発表論文集, pp.73–76.

山本幹雄 増山正和 (1997). “品詞・区切り情報を含む拡張文字の連鎖確率を用いた日本語形態素解析.” 言語処理学会第 3 回年次大会発表論文集, pp.421–424.

横尾昭男, 白井諭, 奥山信輔, 河村美佐子, 池原悟 (1997). “日本語形態素解析の誤りの回復について.” 言語処理学会第 3 回年次大会発表論文集, pp.429–432.

付録

EDR の元々の分割　接頭:接頭語

JUMAN あるいは京大コーパス　サ変:サ変名詞，時相:時相名詞，普通:普通名詞，名尾:名詞性名詞接尾辞，名頭:名詞接頭辞，感動:感動詞，副助:副助詞，未定:未定義語

茶筌　サ変:サ変名詞，引用:引用助詞，感動:感動詞，形容:形容詞，固有:固有名詞，終助:終助詞，普通:普通名詞，副名:副詞的名詞，未定: 未定義語，

すもも　形頭:形容詞的接頭語，形動:形容動詞，固有:固有名詞，接尾:接尾語，普通:普通名詞，未定:未定義語

図 5　表 4,5,6における品詞の略語の説明

表 4　EDR コーパスの/元々/JUMAN/茶筌/すもも/の分割における過分割

数	前文脈	形態素/品詞	形態素/品詞	後文脈
	元々の分割における過分割			
1	争の原理が働くからコストを	引き下/動詞	げ/語尾	よう、良質のものを
1	もマルクス・レーニン主義を	掲/動詞	げ/語尾	、労働者党の１党独
4	その赤字資金の調達は海外に	求/動詞	め/語尾	たので、国内の個人
2	以上	述/動詞	べ/語尾	てきたように，トン
1	象観測衛星や海洋観測衛星を	打ち上/動詞	げ/語尾	，そのデータが日常
3	組合せ論などに多くの業績を	挙/動詞	げ/語尾	ている．
8	年前からサンゴの勢いが衰え	始/動詞	め/語尾	、調査したところ７
4	格フレームの	考/動詞	え/語尾	方は，自然言語の意
1	オスミウムとイリジウムには	耐/接頭	食性/名詞	と硬さを利用しての
4	スに響く特有の音響が左党に	受/動詞	け/語尾	人気が高まった。
1	クルート側が、懸命に防戦に	努/動詞	め/語尾	ていた時期と符合し
2	年は株による財テクで赤字を	埋/動詞	め/語尾	ています。
	JUMAN の過分割			
1	、モノを扱う店員のように、	はっ/感動	きりいって/動詞	横流しの役得もない
1	その銘柄の株価が跳ね上がり	始/人名	めた/副詞	じゃないですか。
1	てんびんや棒	ばか/普通	り/普通	以外の，今日使用し
1	き地に山をなしているのは、	あ/普通	る/普通	製パン業者がこのほ
1	，正しい測定のためにはうき	ばか/普通	りや/人名	液体を入れる容器を
1	北京近郊にある	蘆/普通	溝橋/普通	は、８００年もの歴
1	と各国の情報機関はたがいに	嫉/普通	妬心/普通	が強く、また情報を
2	山形　　本日はお	忙/普通	しい/動詞	ところありがとうご
2	当夜は、このビニール袋ごと	持/普通	って/副助	学部長室に忍び込ん
1	アミ育ちで、紅い髪の、背の	ひ/動詞	ょ/未定	ろっとした男だ。
1	産が大打撃を受けるのはまず	間違/普通	いなさ/普通	そうだ。
1	ハト派には、	言/普通	う/普通	をはばかる雰囲気が
	茶筌の過分割			
1	雪の少ない	弥/固有	陀/未定	ケ池のほとりに小さ
3	さまの話だが、うまく行けば	結/普通	果/普通	は同じということだ
1	母は、おかわりの分も	考/普通	えて/普通	と、いつも食事を余
1	こかのひなびた温泉宿にでも	もぐ/動詞	りこんだ/形容	に違いない。
2	じいさんの時代と、たいして	変/形容	わっ/感動	ちゃいないのだ。
1	れませんが、今の気持ちは、	うれ/動詞	しい/動詞	の一言につきます。
1	囲の明るさに応じて見やすい	よ/終助	うに/普通	調光され、前方の視
1	身内に	厳/普通	しく/普通	というか、梶山静六
1	しかし、こと	経/動詞	済/普通	問題に関しては、こ
1	の発射装置が展開され、１カ	所以/普通	上/副名	その基地が置かれて
1	こもったような声でぼそぼそ	としゃ/サ変	べる/普通	男がだれなのか、思
1	くせが始まったと聞き流せば	すべ/普通	て/引用	円満に運ぶのではな
	すももの過分割			
4	、中央対地方という図式は、	そのよ/固有	うに/普通	はしないというタテ
1	北京近郊にある	蘆/普通	溝橋/普通	は、８００年もの歴
6	ビルの計画段階から、	省エネ/普通	ルギー/未定	を考えていた松下興
2	以下，	簡/普通	単/普通	のため最小化問題を
4	小にすることを，論理関数の	簡/普通	単/形頭	化という．
2	式は，ハードウェアの実現の	容/普通	易/普通	性などから２進数で
1	機能的研究の成果は、改めて	評/普通	価/普通	さるべきものと考え
2	とすれば、遠くない時期に再	利上/未定	げ/接尾	の可能性もある。
1	もう１つ大きな	間/動詞	題/未定	は将来のバックアッ
1	ば各種機器の強度をそれほど	頑/形動	丈/普通	にする必要もなく、
1	めて絵筆とし、刑務所の床を	ふくぞう/普通	きん/固有	で表面の汚れをぬぐ
1	信サービスに比べると十分な	厳/固有	密/普通	性をもつわけではな

表 5　京大コーパスの/元々/JUMAN/の分割における過分割 (教師なし学習)

数	前文脈	形態素/品詞	形態素/品詞	後文脈
	元々の分割における過分割			
1	町を歩いとった。今は変に気	取/普通	っ/サ変	とる。ここは芦屋や
1	州で中国とＡＳＥＡＮ諸国が	南/普通	沙/普通	問題で初の非公式外
1	なぞの	浮世/普通	絵師/普通	、東洲斎写楽が独特
1	陳平両氏らが退陣要求の署名	集/普通	め/名尾	を始めた。
1	夕方から夜にかけて、東名	高/普通	速道/普通	では静岡県の日本坂
1	総理府男女共同	参/普通	画室/普通	が三日付で発表した
1	るスポーツを身に着けられる	部活/普通	動/普通	にしてほしいものだ
1	前七時半ごろ、山梨県の甲斐	駒/普通	ケ/普通	岳八合目を登山中の
1	チェコの独立系	有/普通	力紙/普通	「リドベ・ノビニ」
1	日本労働組合	総連/普通	合/普通	会が、欧米五カ国の
1	手腕を、パネッタ大統領首席	補/普通	佐官/普通	に買われ、国務省か
1	自民党	北海/普通	道連/普通	会長の佐藤孝行衆院
	JUMAN の過分割			
1	いっぱいになるが、畳敷きの	休/普通	憩室/普通	は広く、カラオケも
1	、いじめられる方も変に落ち	込/普通	ま/名頭	へんかった。今のい
1	ただし、準決勝の勝ちっ	ぷ/未定	り/普通	は神鋼より鮮やか。
1	大きく、過去、栃錦、初代若	乃/未定	花/普通	、柏戸、北の湖が金
1	借家人の優先権を認める臨時	処/普通	理法/普通	が発動されたことも
1	、時速五〇キロで走行中、橋	北詰/人名	め/名尾	のコンクリート製欄
1	「あら、ホテルでごちそうを	食/普通	べた/副詞	じゃない」といった
1	っと気軽に楽しめるような品	ぞ/未定	ろ/普通	えをし、価格も２０
1	なぞの	浮世/普通	絵師/普通	、東洲斎写楽が独特
1	幡東区役所は「本籍地などを	申/普通	請人/普通	が知っていれば、そ
1	対戦相手の貴乃花	攻/普通	略法/普通	については「四つに
1	陳平両氏らが退陣要求の署名	集/普通	め/名尾	を始めた。

表 6　京大コーパスに対する JUMAN の過分割 (教師あり学習)

数	前文脈	形態素/品詞	形態素/品詞	後文脈
2	に並べてあるが、ここからも	艾/普通	澤/人名	の表現の特異さを味
1	だ、起訴によって代表の麻原	彰/人名	晃/人名	被告らの殺人の罪を
1	ただし、準決勝の勝ち	っ/未定	ぷ/未定	りは神鋼より鮮やか
1	晴、武村正義、羽田孜、細川	護/普通	熙/人名	、横路孝弘、渡辺美
6	発想したといわれているが、	埴/普通	谷/普通	氏の独白はこの原点
1	当主の継承とは万	之/普通	丞/普通	が野村万蔵家の公式
1	選挙ずくめの今年、	小/名頭	渕/未定	は自民党の選挙準備
1	石原氏は首相の	伊勢参/サ変	拝取り/サ変	やめについて「あく
4	サッカーリーグチェアマン、	川/普通	淵/普通	三郎さんは東京五輪
1	四年後に迎える	蓮/普通	如/普通	上人五百回遠忌に合
1	、時速五〇キロで走行中、橋	北詰/人名	め/名尾	のコンクリート製欄
2	捜索の容疑は、二信組の高橋	治/普通	則/未定	、鈴木紳介両理事長

略歴

内山 将夫:　筑波大学第三学群情報学類卒業 (1992). 筑波大学大学院工学研究科

博士課程修了 (1997). 信州大学工学部電気電子工学科助手 (1997). 郵政省通信総合研究所非常勤職員 (1999). 博士 (工学).

(0 年 0 月 0 日 受付)
(0 年 0 月 0 日 再受付)
(0 年 0 月 0 日 採録)