TR-I-0215

単語の意味カテゴリーを用いた係り受け整合度の平滑化

Smoothing Technique of Matching Score of Dependencies using Semantic Word Categorization

江原暉将
Terumasa EHARA

1991 5 24

内容梗概

構文的曖昧性を解消するための制約として、単語間の係り受け整合度が有効であり、従来から利用されている。この整合度を求める手法の１つに、係り受けデータを収集し、その度数に基づいて、整合度を計算するものがある。しかし、この場合、度数が小さいデータが大量に存在し、整合度の推定精度が悪い欠点がある。この欠点を改良するために、単語の集合をカテゴリー化し、度数の小さいデータについては、カテゴリー間の整合度を単語間の整合度に変えて用いる手法を提案する。本報告では、整合度学習データを解析対象からのランダムな標本としてとらえ、ある条件の下で、本手法を利用することによって、標本から計算された整合度の推定値の誤差が減少することを理論的に示す。次に、ＡＴＲ対話データベースおよび新聞データベースから抽出された係り受けデータを用いて、整合度計算実験を行い、上記手法の有効性を実証する。

TR-I-0215

単語の意味カテゴリーを用いた係り受け整合度の平滑化

Smoothing Technique of Matching Score of Dependencies using Semantic Word Categorization

江原暉将
Terumasa EHARA

1991 5 24

Abstract

Matching score of word dependencies are effective to disambiguate the syntactic ambiguities. One of the methods in computing the matching score involves using the frequency of the dependencies in a dependency database. This method, however, is unreliable because many items occur at frequencies too low, e.g., 1, to obtain accurate score estimation. To overcome this problem, I propose a new score-computing method in which the scores of low frequency items are computed not by word dependency frequency itself but by the dependency frequency of the semantic categories containng the words. In this report, assuming the learning data for the score-computing is the random samples from the population, it is theoretically shown that the proposed method, under certain conditions, improves accuracy. Experimental results showing the effectiveness of the method are also described. In this experiment, the dependency database which is gathered by ATR and another database made from newspaper articles are used.

目　次

1．はじめに・・・・・・・・・・・・・・・・・・・・・・・・・・・・　1

2．係り受け整合度の計算法・・・・・・・・・・・・・・・・・・・・　2

3．係り受け整合度計算実験・・・・・・・・・・・・・・・・・・・・　5

3．1　実験に使用したデータ・・・・・・・・・・・・・・・・・・・　5

3．2　係り先を固定した実験・・・・・・・・・・・・・・・・・・・　6

3．2．1　実験の方法・・・・・・・・・・・・・・・・・・・・・・・　6

3．2．2　実験の結果・・・・・・・・・・・・・・・・・・・・・・・10

3．2．2．1　係り受けパターンの分布・・・・・・・・・・・・・・10

3．2．2．2　述詞固定係り受け語形ファィルのデータ数・・11

3．2．2．3　カテゴリーの粗さによる推定整合度の変化・・11

3．2．2．4　平滑化パラメータによる
整合度推定精度の変化・・・12

3．2．2．5　国際会議データを新聞データから
予測するときの精度・・・・12

3．2．2．6　カテゴリーを利用するときの有効度の検証・・13

3．2．2．7　誤差推定値と誤差実測値の比較・・・・・・・・14

3．3　係り元を固定したとき・・・・・・・・・・・・・・・・・・14

3．3．1　実験の方法・・・・・・・・・・・・・・・・・・・・・・14

3．3．2　実験の結果・・・・・・・・・・・・・・・・・・・・・・14

3．3．2．1　名詞固定係り受け語形ファィルのデータ数・・14

3．3．2．2　平滑化パラメータによる
整合度推定精度の変化・・・15

3．4　係り元も係り先も固定しない場合・・・・・・・・・・・15

3．4．1　実験の方法・・・・・・・・・・・・・・・・・・・・・・15

3．4．2　実験の結果・・・・・・・・・・・・・・・・・・・・・・15

３．４．２．１　平滑化パラメータによる
整合度推定精度の変化・・・１６

３．４．２．２　平滑化による整合度の変化・・・・・・・・・１６

４．　曖昧性解消への利用・・・・・・・・・・・・・・・・・・５１

５．　関連研究との比較・・・・・・・・・・・・・・・・・・・５３

６．　今後の課題・・・・・・・・・・・・・・・・・・・・・・５３

７．　おわりに・・・・・・・・・・・・・・・・・・・・・・・５６

参考文献・・・・・・・・・・・・・・・・・・・・・・・・５６

付録１　係り受け整合度の計算法と誤差の推定・・・・・・・・５８

付録２　係り受け元データ　（文例ファィル）・・・・・・・・６５

付録３　意味カテゴリー付与済みデータ・・・・・・・・・・・６７

付録４　係り受け語形ファィル・・・・・・・・・・・・・・・６９

付録５　有効度の計算に用いたファィル・・・・・・・・・・・７１

図・表　目次

図３．１　係り受け語形ファィルの度数分布　（語形と分類番号４桁）
図３．２　係り受け語形ファィルの度数分布　（語形）
図３．４　係り受け語形ファィルの度数分布　（分類番号４桁）
図３．５　係り受け語形ファィルの度数分布　（分類番号３桁）
図３．６　係り受け語形ファィルの度数分布　（分類番号２桁）
図３．３　高頻度の係り受け語形
図３．７　実測整合度（実線）と推定整合度（点線）
ｗ＝参加、ｋ＝６，ｊ＝１
図３．８　実測整合度（実線）と推定整合度（点線）
ｗ＝参加、ｋ＝６，ｊ＝２
図３．９　実測整合度（実線）と推定整合度（点線）
ｗ＝参加、ｋ＝６，ｊ＝３
図３．１０　実測整合度（実線）と推定整合度（点線）
ｗ＝参加、ｋ＝６，ｊ＝４
図３．１１　実測整合度と推定整合度の相関　ｗ＝参加、ｋ＝６，ｊ＝１
図３．１２　実測整合度と推定整合度の相関　ｗ＝参加、ｋ＝６，ｊ＝２
図３．１３　実測整合度と推定整合度の相関　ｗ＝参加、ｋ＝６，ｊ＝３
図３．１４　実測整合度と推定整合度の相関　ｗ＝参加、ｋ＝６，ｊ＝４
図３．１５　平滑化パラメータ番号による総誤差の変化　１
述詞固定　実線：参加　点線：送る　破線：出る
図３．１６　平滑化パラメータ番号による総誤差の変化　２
述詞固定　実線：参加　点線：送る　破線：出る
図３．１７　実測整合度（実線）と推定整合度（点線）
ｗ＝参加、ｋ＝６，ｊ＝７
図３．１８　実測整合度と推定整合度の相関　ｗ＝参加、ｋ＝６，ｊ＝７
図３．１９　実測整合度（実線）と推定整合度（点線）
ｗ＝参加、ｋ＝６，ｊ＝１
電話、キーに出現する係り元に限定
図３．２０　実測整合度（実線）と推定整合度（点線）
ｗ＝参加、ｋ＝６，ｊ＝７
電話、キーに出現する係り元に限定
図３．２１　有効度の分布　ｗ＝参加、ｋ＝６，ｊ＝２　分類番号４桁
図３．２２　有効度の分布　ｗ＝参加、ｋ＝６，ｊ＝３　分類番号３桁
図３．２３　有効度の分布　ｗ＝参加、ｋ＝６，ｊ＝４　分類番号２桁
図３．２４　有効度の分布　ｗ＝参加、ｋ＝６，ｊ＝７　分類番号３桁
図３．２５　有効度の分布　ｗ＝参加、ｋ＝６，ｊ＝７　分類番号２桁
図３．２６　誤差推定値と誤差実測値の関係
ｗ＝参加、ｋ＝６，ｊ＝７
実線：実測値　点線：推定値
図３．２７　誤差推定値と誤差実測値の相関
ｗ＝参加、ｋ＝６，ｊ＝７
図３．２８　平滑化パラメータ番号による総誤差の変化　１
名詞固定　実線：会議　点線：円　破線：私
図３．２９　平滑化パラメータ番号による総誤差の変化　２
名詞固定　実線：会議　点線：円　破線：私
図３．３０　平滑化パラメータ番号による総誤差の変化　１
図３．３１　平滑化パラメータ番号による総誤差の変化　２
図３．３２　実測整合度（実線）と推定整合度（点線）

k = 6, j = 1
図３．３３　実測整合度（実線）と推定整合度（点線）
k = 6, j = 7

表３．１　学習データと試験データの組合せ
表３．２　平滑化パラメーターｍ０，ｍ１，ｍ２の値
表３．３　カテゴリーの出現割合
表３．４　述詞固定係り受け語形ファィルのデータ数
表３．５　名詞固定係り受け語形ファィルのデータ数
表３．６　カテゴリーを用いないときと用いるときの総誤差の比較

１．はじめに

日本語を文節間の係り受け解析によって構文解析する場合、構文的曖昧性が問題になることがある。例えば、

名詞$_1$が　動詞$_1$している　名詞$_2$を　動詞$_2$した。　（１．１）

という文では、「名詞$_1$が」は「動詞$_1$している」にも「動詞$_2$した」にも係り得るため、構文的に曖昧である。前者を構文１と呼び、後者を構文２と呼ぶと、例えば、

鳥が　飛んでいる　空を　見た。　（１．２）

では、構文１が妥当であり、

鳥が　曇っている　空を　見た。　（１．３）

では、構文２が妥当である。このような構文的曖昧性を解消するために文節間の係り受け整合度が利用できる。上の例では、「名詞$_1$が」が「動詞$_1$している」に係る整合度と「名詞$_1$が」が「動詞$_2$した」に係る整合度を比較して、その大きい方（整合性の良い方）を選択する方法である。名詞ｎが格ｋで動詞ｖに係る整合度を

Ｓ（ｎ，ｋ，ｖ）　（１．４）

とすると、

Ｓ（鳥，が，飛ぶ）　＞　Ｓ（鳥，が，見る）　（１．５）

と考えられるから、（１．２）では、構文１が選択され

Ｓ（鳥，が，曇る）　＜　Ｓ（鳥，が，見る）　（１．６）

と考えられるから、（１．３）では、構文２が選択される。こうして、構文的曖昧性が解消できる。

本文では、ここに示した様に構文的曖昧性を解消するのに有効な整合度の計算方法について述べる。整合度を求める手法として、従来から、

（１）　母国語話者の内省により整合度を求める
（２）　言語データベースの頻度から整合度を求める

が考えられてきた。

（１）は、ｎ，ｋ，ｖの組合せについて、Ｓ（ｎ，ｋ，ｖ）を内省によって求めるものであり、必要な任意の組合せについて、整合度を求めることが出来る反面、大量の組合せについて、内省実験を行うことは困難である。例えば、１，０００個の名詞と１００個の動詞、５個の格助詞について、この方法で整合度を求めようとすると、５０万回内省を行わなければならず、１回１０秒で実行したとしても、１，４００時間位の時間がかかる。そのため、少数の組合せに関する内

省データから他の組合せの整合度を推定する方法も提案されているが［田中（英）］、現状では（１）の方法は小規模なレベルに留まっている。

（２）の方法は、多量のデータさえ得られれば、単純に頻度の統計を取るだけで、整合度が求められる。しかし、この方法には、言語データベースを用いた、他の言語学的研究と同様に［Atkinson］次の問題点がある。それは、多くの整合性が良いと思われる係り受け関係がデータベースに出現しないか出現しても少数であることである。この様な係り受け関係に関しては整合度が０になるか、あるいは推定誤差が大きくなってしまう。そこで、単語をカテゴリー化して、カテゴリー間の係り受け整合度を求め、この値で単語間の係り受け整合度とする方法が考えられる。従来から行われている意味カテゴリーや意味素性の利用はこの方法の一種であるが、この方法は整合度計算法の理論的根拠があまり明確でなかった。一方、クラスタリングによってカデゴライズする方法［井ノ上］があり、これは、理論的根拠は明確であるが、カテゴリーの性質が直感的に分かりにくく成ることがある。そのため、新しい単語が現れたとき、それをどのカテゴリーに含めたらば良いかが人間に判断できず、再度クラスタリングをやり直さなくては成らない欠点がある。最近、クラスタリングと意味カテゴリーを併用した方法が提案されており［井ノ上２］、上記欠点は克服できるかも知れない。

本報告で述べる方法は、（２）に属するものであり、意味カテゴリーを利用して、少数の度数の係り受けデータに対する整合度を平滑化し、推定精度を向上させようとする方法である。本方法は統計的な根拠に基づいて整合度を計算するため、整合度の意味が明確であり、また、意味カテゴリーを用いているため、新たな単語に対してもそれの属するカテゴリーを人間が判断しやすいという利点がある。

２章では、具体的な整合度計算方法を述べ、３章では、整合度計算実験について述べ、４章では曖昧性解消への利用法を述べ、５章では関連研究と比較し、６章では今後の課題を述べる。

２．係り受け整合度の計算法

本章では（１．４）に示す係り受け整合度Ｓ（ｎ，ｋ，ｖ）の計算法を述べる。ただし、ここでは、一般的にｎを係り元、ｋを関係、ｖを係り先と呼ぶ。また、ｋとｖは固定されており、ｎのみ変化し得る場合について説明する。これ以外の場合もここと同様の議論が成立する。ある言語データベースから係り先ｖに関係ｋで係る係り元の単位語としての集合Ｄを抽出し、これを学習用データとする。また、Ｄ中の異なり語の集合を

$$E = \{n_1, n_2, \cdots, n_M\} \qquad (2.1)$$

とする。また、Ｄの要素数をＮとする。このとき、係り元n_i（ｉ＝１，・・，Ｍ）のＤ中での度数をf_iとする。係り元n_iが関係ｋで係り先ｖに係る整合度Ｓ（n_i，ｋ，ｖ）をｓ$_i$と書くと、

$$s_i = -\frac{f_i}{N} \qquad (2.2)$$

と定めたい。しかし、ｆの値が小さいときは、ｓの誤差推定値が大きくなって

i　　i

しまう。言語データベースを解析対象である母集団からの標本と捉えると、s_iは標本空間上の確率変数となる。母集団の大きさをＮＰ、母集団でのn_iの度数をfp_iとすると、整合度の真値は

$$sp_i = \frac{fp_i}{NP}$$

である。そこで、整合度の誤差推定値e_iはＥＳを標本空間での平均値をとる作用素として

$$e_i = ES[(s_i - sp_i)^2]^{1/2}$$

と定義するのが妥当である。付録１によると、この値は近似的に

$$e_i = f_i^{1/2}/N \qquad (2.5)$$

となる。例えば、$f_i = 1$のときは、e_iの大きさが１／Ｎになってしまう。そこで、平滑化パラメータと呼ばれる実数ｍを定め、

$$f_i < m$$

の場合には、単語をカテゴリーにまとめて、e_iの値を小さくすることを企る。（２．１）のＥをカテゴリー

$$F = \{c_1, c_2, \cdots, c_K\} \qquad (2.7)$$

に直和分割する。そして、c_kをn_iが含まれるカテゴリーとするとき、

$$s'_i = \frac{1}{|c_k|}\Sigma[n_j \in c_k]\{\frac{f_j}{N}\} \qquad (2.8)$$

で整合度を定義する。ここで、$|c_k|$はc_kの要素数である。Σ［述語］式　は「述語」の範囲にわたる「式」の和を意味する。このとき、付録１から誤差推定値e_i'は近似的に次の様になる。

$$e_i' = \{\frac{1}{N^2 * |c_k|^2}\Sigma[n_j \in c_k] f_j + \frac{1}{N^2}(\frac{1}{|c_k|}\Sigma[n_j \in c_k] f_j - f_i)^2\}^{1/2} \qquad (2.9)$$

さらに、カテゴリー内の誤差推定値の和を

$$ec_k^2 = \Sigma\,[n_i \in c_k]\; e_i^2 \qquad (2.10)$$

$$ec_k'^2 = \Sigma\,[n_i \in c_k]\; e_i'^2 \qquad (2.11)$$

と置き、さらに、$E[c_k]$、$V[c_k]$を

$$E[c_k] = \frac{1}{|c_k|}\Sigma\,[n_i \in c_k]\;\frac{f_i}{N^2} \qquad (2.12)$$

$$V[c_k] = \frac{1}{|c_k|}\Sigma\,[n_i \in c_k]\;\frac{1}{N^2}\left(\frac{1}{|c_k|}\Sigma\,[n_j \in c_k]\; f_j - f_i\right)^2 \qquad (2.13)$$

と定義すると、付録1から

$$ec_k'^2 - ec_k^2 = E[c_k] + |c_k| * V[c_k] - |c_k| * E[c_k] \qquad (2.14)$$

となる。カテゴライズによって誤差が減少する条件は（２．１４）≦０であり、これは

$$\frac{(|c_k| - 1)}{|c_k|} \geq \frac{V[c_k]}{E[c_k]} \qquad (2.15)$$

となる。こうして、誤差が減少する条件が（近似的に）求まった。

$$Q[c_k] = \frac{(|c_k| - 1)}{|c_k|} - \frac{V[c_k]}{E[c_k]} \qquad (2.16)$$

のことをカテゴリーc_kを利用することによる有効度と呼ぶ。有効度が正であればカテゴリー化をした方が精度が高く、負であれば、低い。

以上の考察に基づいて、標本での（n_i, k, v）の出現度数f_iが小さい場合は、s_iをn_iを含むカテゴリーc_k上での平均整合度

$$s_i' = \frac{1}{|c_k|}\Sigma\,[n_j \in c_k]\; s_j \qquad (2.17)$$

で置き換える方法を提案する。

3．整合度計算実験

本章では、2章に示した手法の有効性を検証するための整合度計算実験の方法と結果について示す。

3．1　実験に使用したデータ

実験に使用したデータは3種類ある。ＡＴＲ対話データベース［江原］の係り受け関係データから抽出された国際会議に関する電話会話データ（電話）とキーボード会話データ（キー）および、朝日新聞（1年＋84日分）から抽出された係り受けデータ（新聞）［田中（康）］である。ここでは、関係kとして、「が」格関係に限って実験した。係り元、係り先とも文節を単位にして元データ（文例ファィル）を作成した。電話データとキーデータには標準表現と品詞が付与されているのでそれを利用した。新聞データはこれらが付与されていないので、形態素解析を行って標準表現や品詞を求めた。元データの大きさは次に示す通りである。元データの形式と例を付録2に示す。

電話	1，336件
キー	794件
新聞	194，494件

つぎに、元データの係り先語と係り元語にカテゴリー番号を付与した。ここでは、カテゴリーとして、角川の類語国語辞典の意味カテゴリー（分類番号）を利用した［浜西］。分類番号は、例えば

用紙	902b
送る	280,386b,777d,784c

の様に語形に対して、3桁ないし4桁の番号から成り、番号の上位の桁を取ることによってカテゴリーが粗く成って行く。番号が複数あるものは、語義の違いによる分類の違いである。本実験では3桁の分類番号は、4桁目にブランクがあると見なして、全て4桁の分類番号として扱った。

係り元や係り先の語形全体に意味カテゴリーが付与できないときは、次のアルゴリズムに従って付与した。それでも付与できない場合はデータから除外した。除外されたデータには表記法の不一致などにより、意味カテゴリーが付与できなかったものも含まれている。

係り元語形に対する意味カテゴリー付与アルゴリズム

（1）語形全体で付与可能ならば付与する。
（2）機能語を削除して付与可能ならば付与する。
（3）接頭語を削除して付与可能ならば付与する。
（4）接尾語を削除して付与可能ならば付与する。
（5）内容語列を前部分より削除して付与可能なら付与する。
（6）それでも付与不能であれば、データを削除する。

係り先語形に対する意味カテゴリー付与アルゴリズム

（１）語形全体で付与可能ならば付与する。
（２）機能語を削除して付与可能ならば付与する。
（３）接頭語を削除して付与可能ならば付与する。
（４）接尾語を削除して付与可能ならば付与する。
（５）内容語列を後部分より削除して付与可能なら付与する。
（６）それでも付与不能であれば、データを削除する。

このアルゴリズムによって意味カテゴリーが付与されたデータ量は次の通りである。

電話　　８３０件
キー　　５１１件
新聞　１３７，６３７件

データの形式と例を付録３に示す。このデータには類語国語辞典の番号の他に、分類語彙表（増補版）［国研］の番号も付与されている。また、意味カテゴリーが複数付与可能なデータにはそれらをすべて付与した。

つぎに、学習データと試験データを作成するため、新聞データをランダムに５等分し、電話データ、キーデータと共に係り受け語形ファィルを作成した。データ番号とデータ種別の対応は次の通りである。

データ１：電話
データ２：キー
データ３～７：新聞

ここで、意味カテゴリーが複数付与されたデータは、そのカテゴリーが均等に出現したと仮定して、度数を案分してデータを作成した。係り受け語形ファィルは、３１９，２６６レコードから成る。内容例を付録４に示す。

３．２　係り先を固定した実験

３．２．１　実験の方法

まず、係り先ｖを固定して実験を行った。実験の方法は次の通りである。
（１）固定された係り先を持つデータを抽出する。
（２）全データを学習データと試験データに分け、学習データから整合度を求める。この時、度数の少ないデータについては、そのデータの属する意味カテゴリーにわたる平均整合度を用いる。意味カテゴリーとして、分類番号の４、３、２桁までの３種類を用いる。また、カテゴリーの要素数としては、標本でのそのカテゴリーに属するデータの異なり数とする。
（３）試験データを用いて、実測整合度と（２）から得られる推定整合度を比較し、推定誤差を評価する。
以下に実験方法の詳細を述べる。

（１）述詞固定係り受け語形ファィルの作成

係り受け語形ファィルの係り先語形（今の場合、動詞、形容詞、形容動詞、述語名詞などが係り先として考えられるので、これらを総称して述詞と呼ぶ）が固

定された語に等しいレコードのみを取り出し、述詞固定係り受け語形ファィルを作成する。このとき係り先述詞の分類番号は無視して、係り元語形と係り元分類番号が同一のレコードはマージする。

係り元語形｜係り元分類番号｜データ１度数｜～｜データ７度数

固定した述詞としては以下に示す３個を用いた。これらの語は国際会議申込のタスクに多く出現する語である。

固定された述詞（ｗ）：
参加
送る
出る

（２）学習データ述詞固定係り受け語形元ファィル（ＤＬ）と試験データ述詞固定係り受け語形元ファィル（ＤＴ）の作成

データ１からデータ７の中から学習データと試験データを定め、それらの度数を求めて、２つのファィルを作成する。

ＤＬ　係り元語形｜係り元分類番号｜学習データ度数
ＤＴ　係り元語形｜係り元分類番号｜試験データ度数

学習データと試験データの組合せは、表３．１に示す通りである。ｋ＝１，．．．，５は新聞データのみを用いた実験である。ｋ＝６は電話とキーデータを試験データに含めて、新聞データから国際会議のタスクのデータをどれだけ精度よく推定できるかを見るものである。

表３．１　学習データと試験データの組合せ
（データの番号と出典種別の関係は３．１を参照）

番号ｋ	学　習	試　験
1	3　4　5	6　7
2	4　5　6	7　3
3	5　6　7	3　4
4	6　7　3	4　5
5	7　3　4	5　6
6	3　4　5	6　7 1　2

（３）学習データ述詞固定係り受け語形ファィル（ＤＷ）と学習データ述詞固定係り受け分類ファィルｎ桁（Ｄｎ）の作成

ＤＬから整合度計算に必要なＤＷとＤｎ　（ｎ＝２，３，４）を（３．１）以

下のようにして作成する。平滑化パラメータｍ０，ｍ１，ｍ２は表３．２に示す実数である。この結果、学習データのカテゴリー内の平均度数ｆによって、整合度計算法が次の様に異なる。

ｆ≧ｍ０	係り元語形と分類番号４桁に対する整合度
ｍ０＞ｆ≧ｍ１	分類番号４桁に対する平均整合度
ｍ１＞ｆ≧ｍ２	分類番号上位３桁に対する平均整合度
ｍ２＞ｆ	分類番号上位２桁に対する平均整合度

表３．２　平滑化パラメーター
ｍ０，ｍ１，ｍ２の値

番号ｊ	ｍ０	ｍ１	ｍ２
1	0	0	0
2	１０万	0	0
3	１０万	１０万	0
4	１０万	１０万	１０万
5	8	6	4
6	7	5	3
7	6	4	2
8	4	2	0

ｊ＝１，２，３，４はカテゴリーの粗さによる推定整合度を比較するための予備実験である。ｊ＝１はカテゴライズを用いていない。ｊ＝５，６，７，８は本手法の有効性を示すと共にｍ０，ｍ１，ｍ２の最適値を求める実験である。

（３．１）　学習データ度数　＞＝　ｍ０　ならば　ＤＷへ出力
そうでなければｔｍｐ１へ出力
ＤＷ　　係り元語形｜係り元分類番号（４桁）｜学習データ度数
ｔｍｐ１　係り元分類番号（４桁）｜学習データ度数

（３．２）　ｔｍｐ１を第１欄でソートし、次の処理を行う。
・同一分類番号の平均学習データ度数の値が　＞＝　ｍ１　ならば　同一分類番号の行をまとめて１行にしてＤ４へ出力する。このとき、分散も出力する。
・そうでなければ行を１つにせずにｔｍｐ２に出力する。
Ｄ４　　係り元分類番号（４桁）｜行数｜平均学習データ度数｜分散
ｔｍｐ２　係り元分類番号（３桁）｜学習データ度数

（３．３）　ｔｍｐ２を第１欄でソートし、次の処理を行う。
・同一分類番号の平均学習データ度数の値が　＞＝　ｍ２　ならば　同一分類番号の行をまとめて１行にしてＤ３へ出力する。このとき、分散も出力する。
・そうでなければ行を１つにせずにｔｍｐ３に出力する。
Ｄ３　　係り元分類番号（３桁）｜行数｜平均学習データ度数｜分散
ｔｍｐ３　係り元分類番号（２桁）｜学習データ度数

（３．４）　ｔｍｐ３を第１欄でソートし、次の処理を行う。
・同一分類番号の行をまとめて１行にしてＤ２へ出力する。このとき、分散も出力する。
Ｄ２　　係り元分類番号（２桁）｜行数｜平均学習データ度数｜分散

（３．５）　ｔｍｐ１，ｔｍｐ２，ｔｍｐ３を削除する。

（４）学習整合度ファィルの作成

ＤＷ，Ｄ４，Ｄ３，Ｄ２にわたる全学習データ度数の合計をＮ’とする。次に、（平均）学習データ度数を全学習データ度数の合計Ｎ’で割って、整合度とし、学習整合度ファィルＳＷ，Ｓ４，Ｓ３，Ｓ２に出力する。

ＳＷ　係り元語形｜係り元分類番号（４桁）｜整合度
Ｓ４　係り元分類番号（４桁）｜整合度
Ｓ３　係り元分類番号（３桁）｜整合度
Ｓ２　係り元分類番号（２桁）｜整合度

（５）試験整合度実測ファィルの作成

ＤＴから試験データ度数の総計Ｎ’’を求める。実測整合度＝試験データ度数／Ｎ’’を求め、試験整合度実測ファィルＳＴＡに出力する。

ＳＴＡ　係り元語形｜係り元分類番号（４桁）｜実測整合度

（６）試験整合度ファィルの作成

ＳＴＡの係り元語形と係り元分類番号を用いて、次の方法で推定整合度を求める。

係り元語形と係り元分類番号（４桁）がＳＷにある　ならば　ＳＷの整合度
係り元分類番号の４桁までがＳ４にある　ならば　Ｓ４の整合度
係り元分類番号の３桁までがＳ３にある　ならば　Ｓ３の整合度
係り元分類番号の２桁までがＳ２にある　ならば　Ｓ２の整合度
それ以外　ならば　推定整合度＝０．０

求まった推定整合度をＳＴＡに付加して、試験整合度ファィルＳＴとする。

ＳＴ　係り元語形｜係り元分類番号｜実測整合度｜推定整合度

（７）推定精度の計算

ＳＴの各レコードｉに付いて実測整合度と推定整合度の差である誤差をｅ［ｉ］とする。また、ｅ［ｉ］の絶対値｜ｅ［ｉ］｜をＳＴの全レコードにわたって足し上げたものを総誤差Ｅとする。

以上の方法において、固定された述詞w、学習データ試験データの組合せ番号k、平滑化パラメータの値の番号jの組合せに対して実験を行った時ＳＴのレコードiに対する実測整合度、推定整合度をそれぞれ

$$s[w, k, j, i] \qquad (3.1)$$

$$\bar{s}[w, k, j, i] \qquad (3.2)$$

と書き、レコードiに対する誤差eを

$$e[w, k, j, i] = s[w, k, j, i] - \bar{s}[w, k, j, i] \qquad (3.3)$$

と書く。また、総誤差Eを

$$E[w, k, j] = \Sigma[i \in ST] \mid e[w, k, j, i] \mid \qquad (3.4)$$

で定義する。eやEによって、推定整合度の精度が評価できる。

3．2．2　実験の結果

3．2．2．1　係り受けパターンの分布

まず、予備的な結果として、係り受け語形ファィルの度数分布について述べる。語形と分類番号4桁の係り受けパターンは3．1で述べた通り、異なりデータ数３１９，２６６である。また、データ1から7までを合計した延べデータ数は２１２，３２３である。延べ数が異なり数より小さいのは、分類番号が複数付与されることにより度数1以下のパターンが存在するためである。図3．1に順位順の度数分布を示す。また、図3．2には分類番号を無視した語形のみでの分布を示す。この場合は図3．1とは異なり、最低の度数は1である。最大度数の係り受け語形パターンとその度数は「可能性が強い」の１４，３８８．５である。高頻度の語形パターンを図3．3に示す。語形パターンの異なり数は８６，０３２であり、その内の約７７％に当たる６５，８５７は度数1のパターンである。この様に係り受けパターンは度数の小さいものが非常に多く、このままでは自然言語処理用の知識ベースとして問題があることは前述した通りである。図3．4から図3．6には、分類番号の上位4、3、2桁までの係り受けパターンの分布を示す。カテゴライズが粗くなるに従って、当然のことながら異なりパターン数は減少する。論理的に可能な分類番号の全ての組合せ数と度数がゼロでない異なりパターン数の比較を表3．3に示す。

表3．3　カテゴリーの出現割合

カテゴリー	論理的に可能な組合せ数	出現したパターン数	割　合（％）
語形と4桁	－－－	319,266	--
上位4桁	－－－	205,252	--
上位3桁	1,000,000	122,034	12
上位2桁	10,000	8,796	88

３．２．２．２　述詞固定係り受け語形ファィルのデータ数

固定された述詞に対するデータ量をまず示す。これは表３．４の通りである。この値は、データ１から７の合計である。

出る＞参加＞送る

の順に度数が大きい。「送る」で異なり数が延べ数より大きいのは多義性による。

表３．４　述詞固定係り受け語形
ファィルのデータ数

述詞	異なり数	延べ数	延べ数／異なり数
参加	377	766	2.03
送る	88	72	0.81
出る	1888	6739	3.56

３．２．２．３　カテゴリーの粗さによる推定整合度の変化

本節では、度数によらず一様にカテゴライズした場合、カテゴリーの粗さと推定整合度の関係を調べる。
例として、語形ｗとデータ組合せ番号ｋを

ｗ＝参加
ｋ＝６

に取り、平滑化パラメーター番号

ｊ＝１、２、３、４

に対して、試験データに対する実測整合度と推定整合度を比較する。これらのｊの値によって、次のような基準によってカテゴライズしたときの推定整合度の値が求まる。

ｊ＝１　係り先語形と分類番号の４桁
ｊ＝２　分類番号の４桁
ｊ＝３　分類番号の上位３桁
ｊ＝４　分類番号の上位２桁

この実験により、カテゴライズの粗さによる推定整合度の変化を見ることができる。図３．７から図３．１０にその結果を示す。ｊ＝１の場合、図３．７に示すように、推定整合度がゼロに成る部分がかなりある。つまり、語形と分類番号４桁の組で整合度を計算した場合、試験データには存在するが学習データには全く出現しないものがかなりあることが分かる。ｊ＝２、３、４と進むに従って、カテゴリーが粗くなって行き、推定整合度がゼロになる場合が減少する。その代わ

り、度数の大きい部分での推定整合度が小さい値となり、それに伴って推定誤差が増大する。カテゴリーが粗ければ粗いほど、一様分布に近づいて行くことが見える。また、図３．１１から図３．１４には、実測整合度と推定整合度の相関を示す。

３．２．２．４　平滑化パラメーターによる整合度推定精度の変化

次に全てのｗ，ｊの組合せに対して

ｋ＝１，２，３，４，５

に対する総誤差Ｅ（３．４）式の算術平均

$$\bar{E}[w, j] = (1/5) * \Sigma [k \in \{1, \ldots, 5\}] E[w, k, j] \quad (3.5)$$

を比較する。これによって、平滑化パラメータにより誤差がどの様に変動するかを見ることができる。その結果を図３．１５と図３．１６に示す。図３．１５より次のことが言える。述詞「出る」に対しては、カテゴライズによって誤差が増加しているが、「送る」に対しては、逆に減少している。これは、表３．４に示す延べ数／異なり数に依存するためである。つまり、この値が大きい場合はカテゴリー化しない方が良く、小さい場合はカテゴリー化した方が良い。この値が中程度である「参加」に対しては、カテゴライズによる著しい誤差の変化はないが、やや増加傾向にある。次に、図３．１６を図３．１５と比較すると、「送る」のｊ＝８を例外として、

ｊ＝５，６，７，８

の全ての場合について、図３．１５の最小誤差より、図３．１６の誤差の方が小さいことが分かる。このことから、度数の小さいデータに対して、カテゴリーを利用して整合度を平滑化する２章に述べた手法が有効であることが分かる。図３．１６より、本実験の範囲では、

ｊ＝７

付近が最適なカテゴリー利用法であることが分かる。図３．１７にｗ＝参加、ｊ＝７、ｋ＝６の場合の実測整合度と推定整合度を示す。図３．１８には、それらの相関を示す。図３．１７で度数が大きい部分の推定整合度はカテゴリーを利用しない図３．７に類似しており、一方、度数が小さい部分でもカテゴライズの効果で推定整合度がゼロになる割合が少ない。つまり、度数の大きい部分も小さい部分も、推定整合度の誤差が小さいことが分かる。

なお、５組の試験データに対して、$\bar{E}$を求めた時の変動係数（標準偏差／平均値）は１０％以下であった。

３．２．２．５　国際会議データを新聞データから予測するときの精度

次に、ＡＴＲでの研究課題である国際会議の申込に関するタスクに出現する係り受けは、新聞に出現するものとは異なると予想される。そこで、表題に示す様

に、国際会議データを新聞データから予測する場合、本手法の有効性が期待できる。この点を調べるために試験データに電話とキーが加わっている

k = 6

の場合を考察する。この場合は、学習データとしては新聞データのみが用いられている。再び、

w = 参加

の場合に見てみよう。試験データの中で電話またはキーに出現しているデータのみを取り出して、実測整合度と推定整合度を比較する。このようなデータは４０例あった。j = 1の場合の結果を図３．１９にj = 7の場合の結果を図３．２０に示す。カテゴリーを利用しないj = 1では、多くの推定整合度がゼロになっており、推定精度が低いことが分かる。一方、カテゴリーを利用するj = 7の時は、平滑化の効果によって、推定精度が向上している。

３．２．２．６　カテゴリーを利用するときの有効度の検証

カテゴリーを利用した場合の有効度Q（式２．１６）の値を

w = 参加
k = 6
j = 2, 3, 4, 7

の場合について調べる。j = 2, 3, 4は度数によらず一様にカテゴライズした場合であり、j = 7は３．２．２．４で最適と考えられた場合である。

図３．２１から図３．２５に結果を示す。一様にカテゴライズした場合は、いずれも有効度が負の大きな値を取る部分があり、度数によらない一様なカテゴライズは精度が低いことが分かる。

一方、度数を考慮したk = 7の方法では、その様なことは生じていない。３桁の場合（図３．２４）、負の部分が存在はするが、その大きさは一様な場合より１桁小さい。２桁の場合（図３．２５）は、全て正かゼロである。４桁の場合はデータが２例しかなく、両者ともQ = 0である。

一様にカテゴライズするとき有効度が負の大きな値を取る部分が存在する理由は、カテゴリーの中に度数の大きなデータが存在するためである。例えば、j = 3のD３におけるカテゴリー 713 の要素数は６であり、それらの語形と度数は

教団	２．０
組合	１．０
団体	１１．０
労組	１．０
会社	０．５
社	８．５

である。「団体」と「社」（実際は「会社名＋社」のパターンと思われる）の度数が大きく、他は小さい。このため、Q = －３．４となってしまう。一方、k = 7の場合は、「団体」と「社」はＤＷに入り、他はＤ２に入るため、このようなことは生じない。

なお、この実験で利用した以下のファィルの内容を付録５に示す。

ＤＬ
ＤＴ
ｊ＝２に対するＤ４
ｊ＝３に対するＤ３
ｊ＝４に対するＤ２
ｊ＝７に対するＤＷ，Ｄ４，Ｄ３，Ｄ２

３．２．２．７　誤差推定値と誤差実測値の比較

誤差推定値である（２．９）式と誤差実測値である（３．３）式の絶対値を

ｗ＝参加
ｋ＝６
ｊ＝７

の場合に付いて図３．２６に示す。また、それらの相関を図３．２７に示す。かなり、ばらつきが認められる。

３．３　係り元を固定したとき

３．３．１　実験の方法

係り元の名詞（ｗ）として、

固定された名詞（ｗ）：
会議
円
私

の３種類を選択した。これらも、述詞の場合と同様、国際会議のタスクで頻出する名詞である。その他の条件は３．２．１と同一である。ただし、語「述詞」は「名詞」と読み変える。表３．１のデータ組合せ、表３．２の平滑化パラメータの値も同一である。

３．３．２　実験の結果

３．３．２．１　名詞固定係り受け語形ファィルのデータ数

表題に示すデータ数は表３．５の通りである。これは、データ１から７の合計である。

私＞円＞会議

の順に度数が大きい。

表３．５　名詞固定係り受け語形
ファィルのデータ数

述詞	異なり数	延べ数	延べ数／異なり数
会議	193	310.9	1.61
円	206	521.3	2.53
私	804	798.9	0.99

３．３．２．２　平滑化パラメーターによる整合度推定精度の変化

名詞固定の実験結果として、表題の結果を述べる。全てのｗ，ｊの組合せに対して

ｋ＝１，２，３，４，５

に対する総誤差Ｅの平均を比較する。これによって、平滑化パラメータにより誤差がどの様に変動するかを見ることができる。その結果を図３．２８と図３．２９に示す。図３．２８より次のことが言える。名詞「円」に対しては、カテゴライズによって誤差が増加しているが、「私」に対しては、逆に減少している。これは、表３．５に示す延べ数／異なり数に依存するためである。つまり、この値が大きい場合はカテゴリー化しない方が良く、小さい場合はカテゴリー化した方が良い。この値が中程度である「会議」に対しては、カテゴライズによる著しい誤差の変化はないが、やや増加傾向にある。次に、図３．２９を図３．２８と比較すると、

ｊ＝７

の場合に誤差が最低に成っており、特に、図３．２８の最小誤差より、小さいことが分かる。このことから、度数の小さいデータは、カテゴリーを利用して整合度を平滑化する２章に述べた手法が有効であることが分かる。

なお、５組の試験データに対して、$\bar{E}$を求めた時の変動係数（標準偏差／平均値）は１８％以下であった。

３．４　係り元も係り先も固定しない場合

３．４．１　実験の方法

係り元も係り先も固定しない点が、３．２、３．３と異なるのみで、その他の条件は同一である。特に、表３．１のデータ組合せ、表３．２の平滑化パラメータの値は同一である。ただし、実験方法において、「述詞固定」という語は削除して読む。また、パラメータｗは存在しない。

３．４，２　実験の結果

３．４．２．１　平滑化パラメーターによる整合度推定精度の変化

実験結果として、表題の結果を述べる。全てのｊの組合せに対して

ｋ＝１，２，３，４，５

に対する総誤差Ｅの平均を比較する。これによって、平滑化パラメータにより誤差がどの様に変動するかを見ることができる。その結果を図３．３０と図３．３１に示す。図３．３０より、カテゴライズによって誤差が増加することが分かる。次に、図３．３１を図３．３０と比較すると、

ｊ＝８

の場合に誤差が最低に成っている。これは、述詞固定や名詞固定の場合と若干異なっているが、

ｊ＝７

でも、誤差はかなり小さくなっている。いずれにしても、度数の小さいデータに対して、カテゴリーを利用して整合度を平滑化する２章に述べた手法が有効であることが分かる。

なお、５組の試験データに対して、$\bar{E}$を求めた時の変動係数（標準偏差／平均値）は１８％以下であった。

表３．６にｊ＝１とｊ＝７の場合の総誤差をまとめて示す。

表３．６　カテゴリーを用いないときと用いるときの総誤差の比較

実験種別		E	E'	E'／E	述べ数 異なり数
係り先固定	参加	0.56	0.44	0.80	2.03
	送る	0.90	0.61	0.67	0.81
	出る	0.39	0.37	0.95	3.56
係り元固定	会議	0.49	0.45	0.92	1.61
	円	0.28	0.26	0.93	2.53
	私	0.72	0.55	0.77	0.99
全体		0.54	0.49	0.91	0.69

３．４．２．２　平滑化による整合度の変化

ｋ＝６

の場合に付いて、平滑化を行わないとき（ｊ＝１）と、行うとき（ｊ＝７）の実測整合度と推定整合度を図３．３２と図３．３３に示す。但し、グラフ作成の都

合上、データを１／１０に間引いて示した。整合度が小さい部分で、平滑化をする場合の方が、推定整合度がゼロ（10^{-8}で表示）になることが少ない。

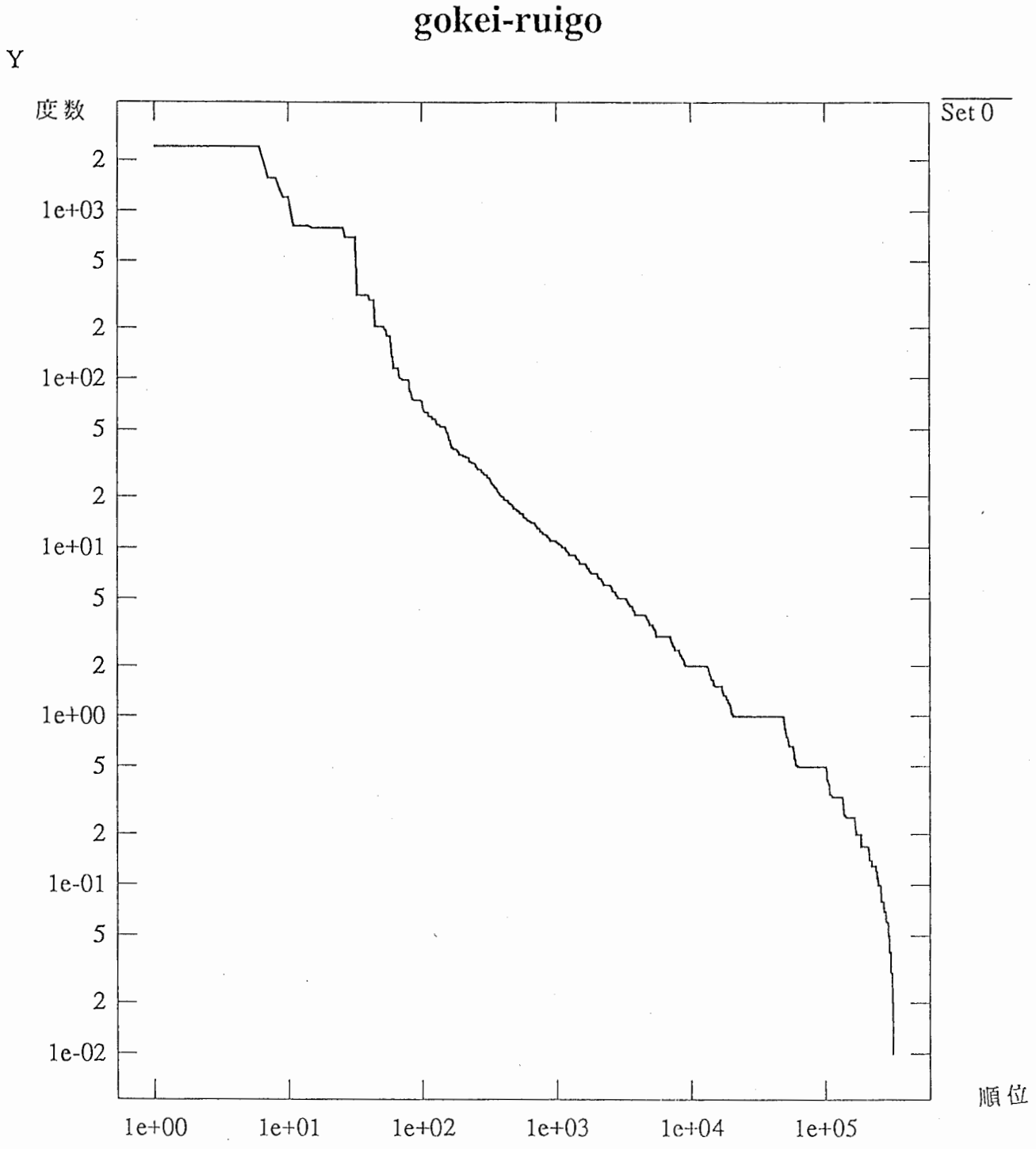


図３．１　係り受け語形ファィルの度数分布
（語形と分類番号４桁）

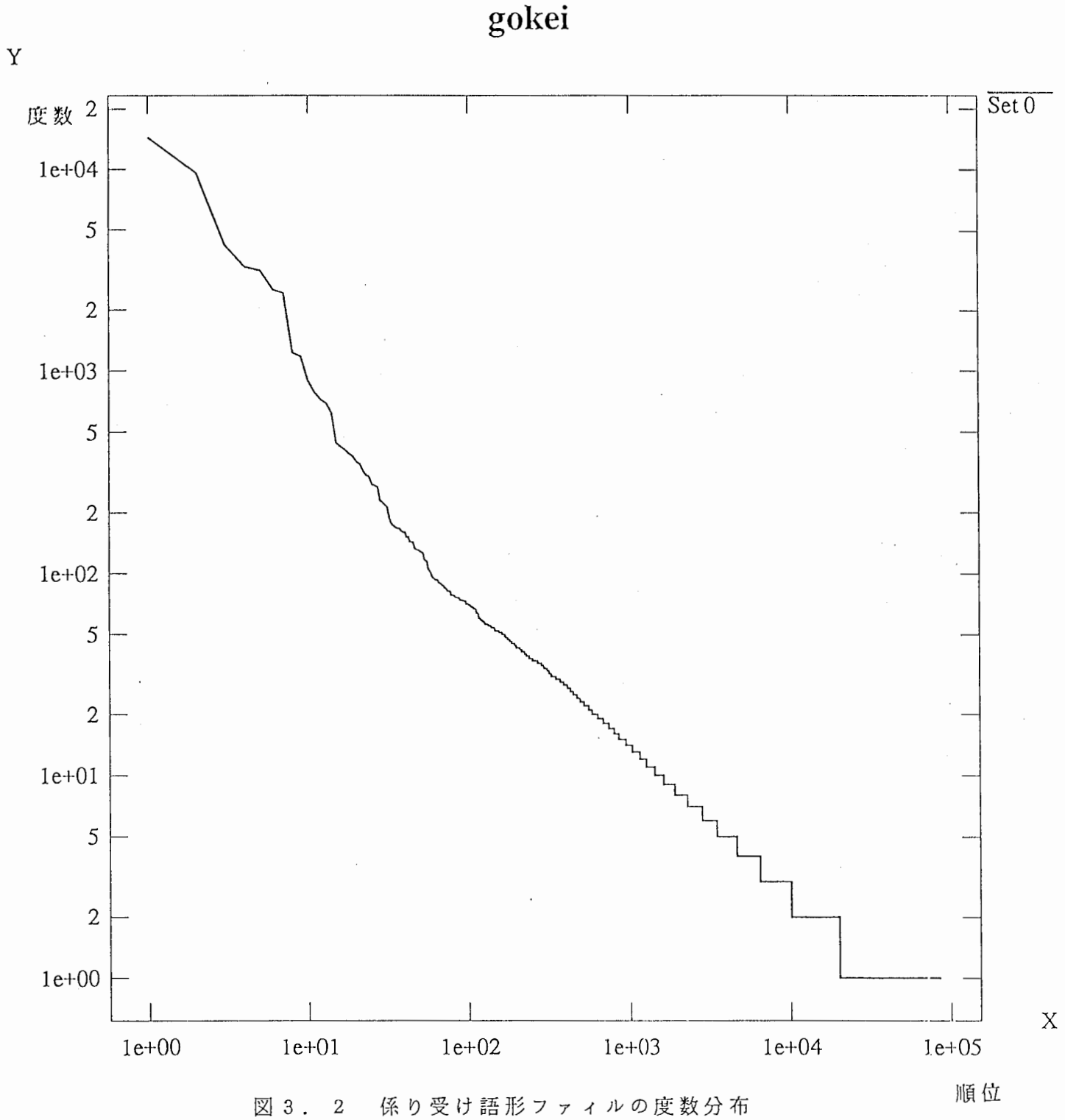


図３．２　係り受け語形ファィルの度数分布
（語形）

可能性|強い|14388.5
見方|強い|9568.32
疑う|強い|4192.2
見方|強まる|3264
声|強まる|3136
動く|強まる|2516.8
可能性|強まる|2417
批判|強い|1224
傾向|強まる|1174
反発|強い|889.92
見通す|強まる|782.08
圧力|強まる|719
不満|強い|689.52
印象|強い|620.16
人|多い|441.98
イメージ|強い|424.32
疑う|強まる|409
批判|強まる|391
抵抗|強い|379.2
要望|強い|357
反対|強い|347.52
色彩|強い|322.56
声|出る|305.76
円|急騰|299.97
結果|出る|276.64
影響|出る|272.48
状態|続く|268
観測|強まる|230
公算|大きい|225.4
感|強い|218.88
感|強まる|213
動く|出る|188.16
風|吹く|176.4
必要|生じる|172.26
要求|強まる|169
率|高い|167.28
意見|出る|167.04
結論|出る|162.24
ケース|多い|160.38
取引|始まる|160
思う|強い|151.68
雨|降る|151
声|高まる|143.55
風|強い|142.8
性格|強い|142.08
高|進む|133
傾向|強い|131.52
意見|一致|131
批判|出る|130
意見|強い|128.64

図３．３　高頻度の係り受け語形

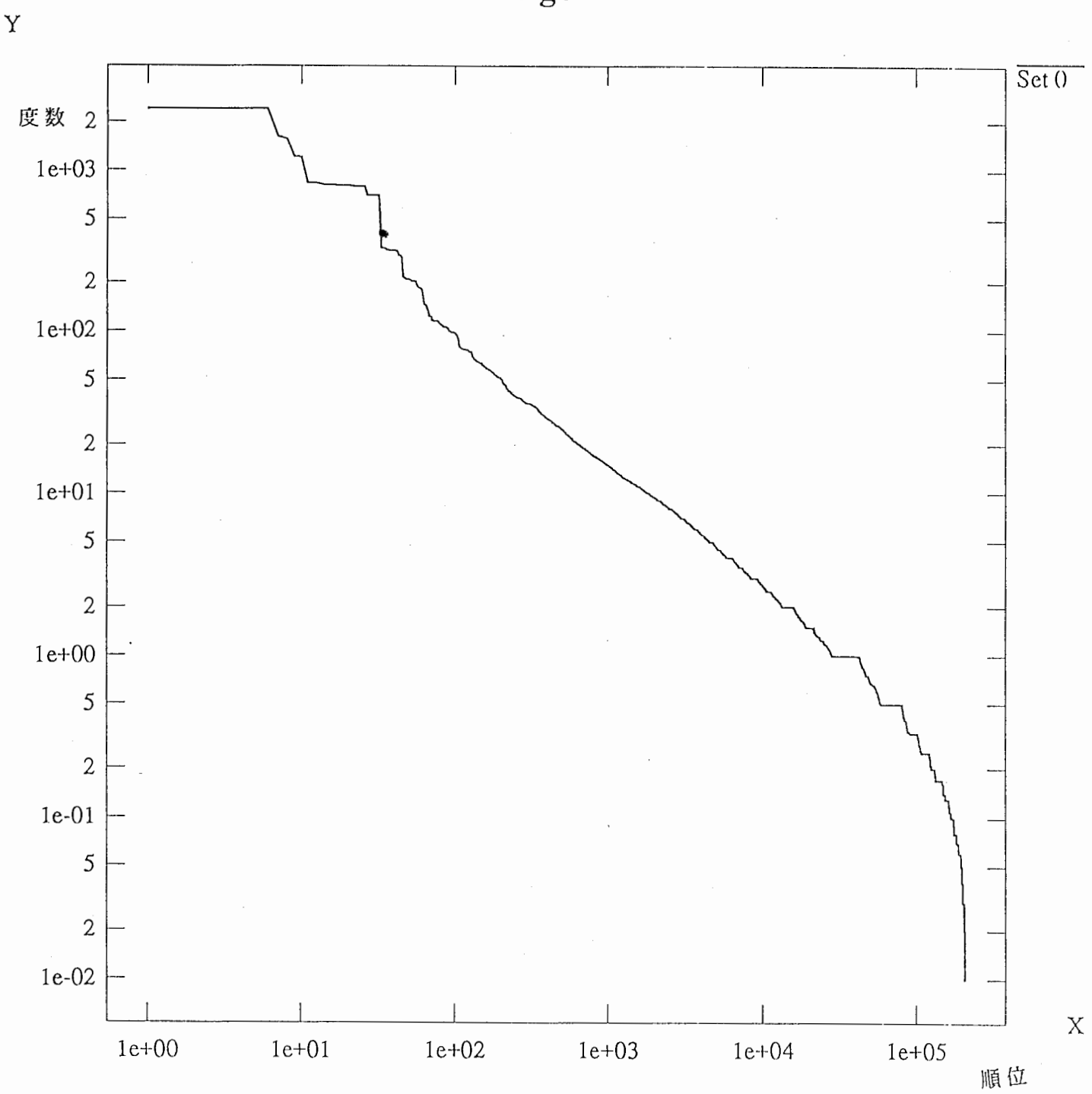


図３．４　係り受け語形ファィルの度数分布
（分類番号４桁）

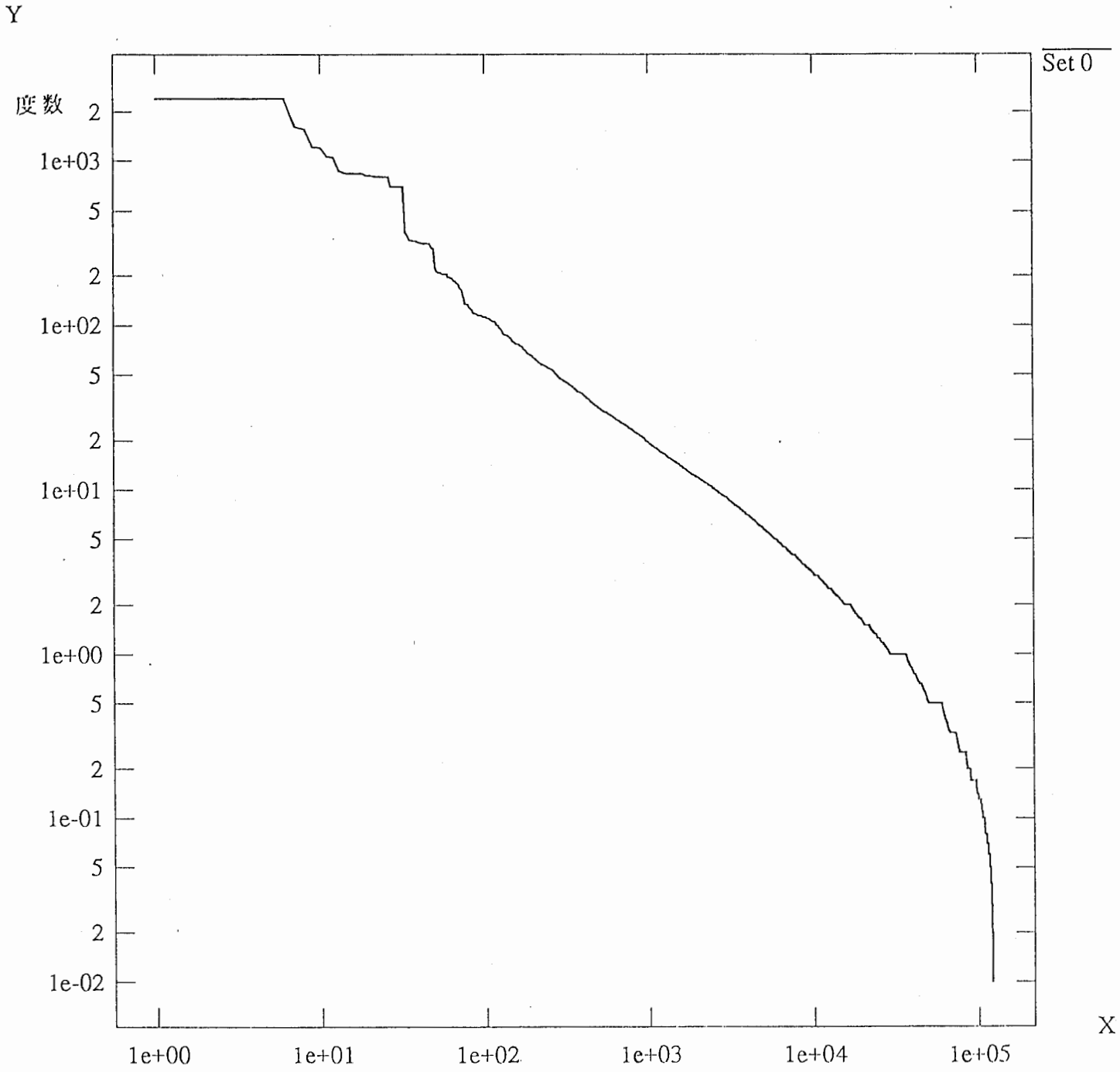


図３．５　係り受け語形ファィルの度数分布
（分類番号３桁）

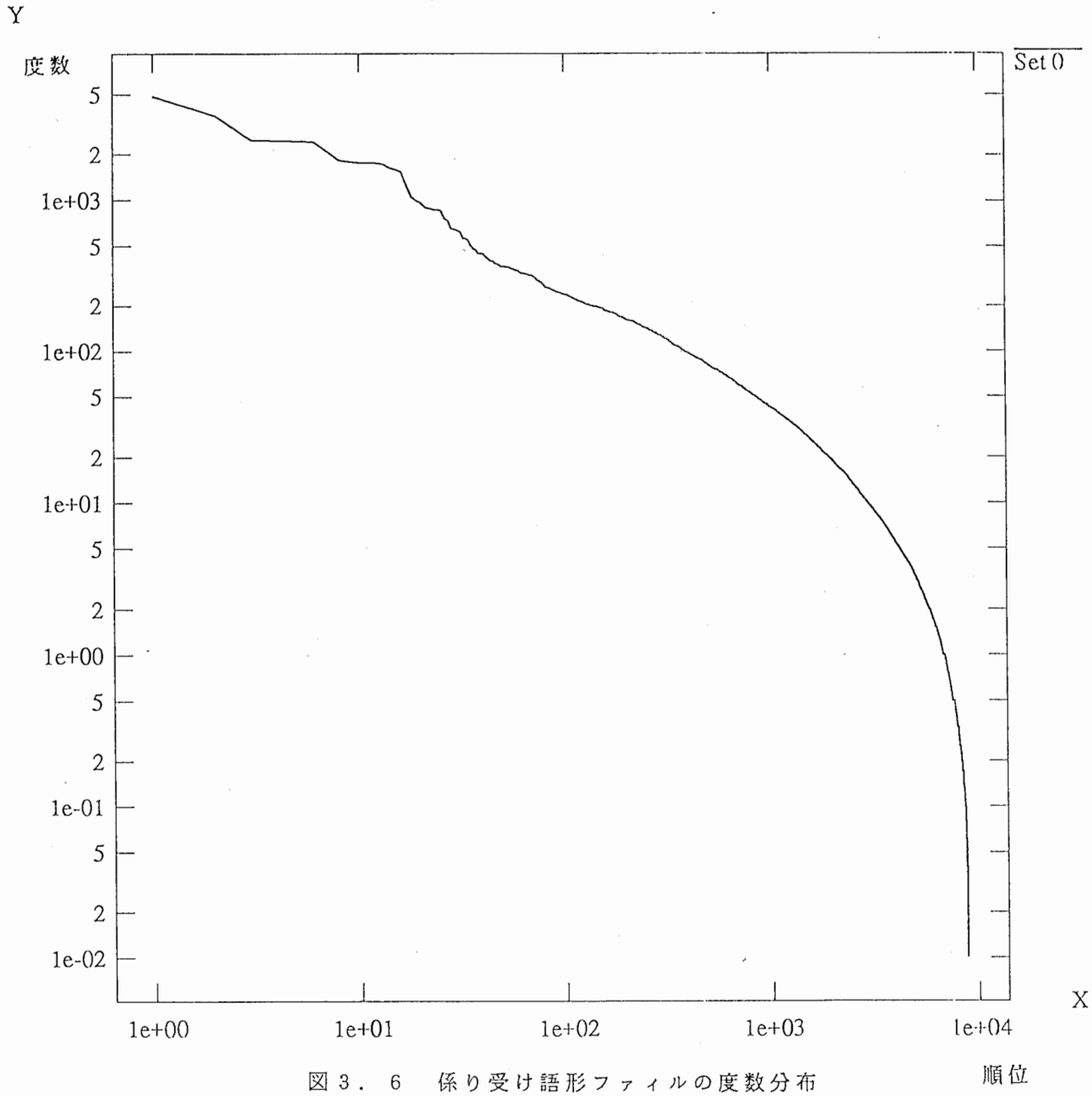


図３．６　係り受け語形ファィルの度数分布
（分類番号２桁）

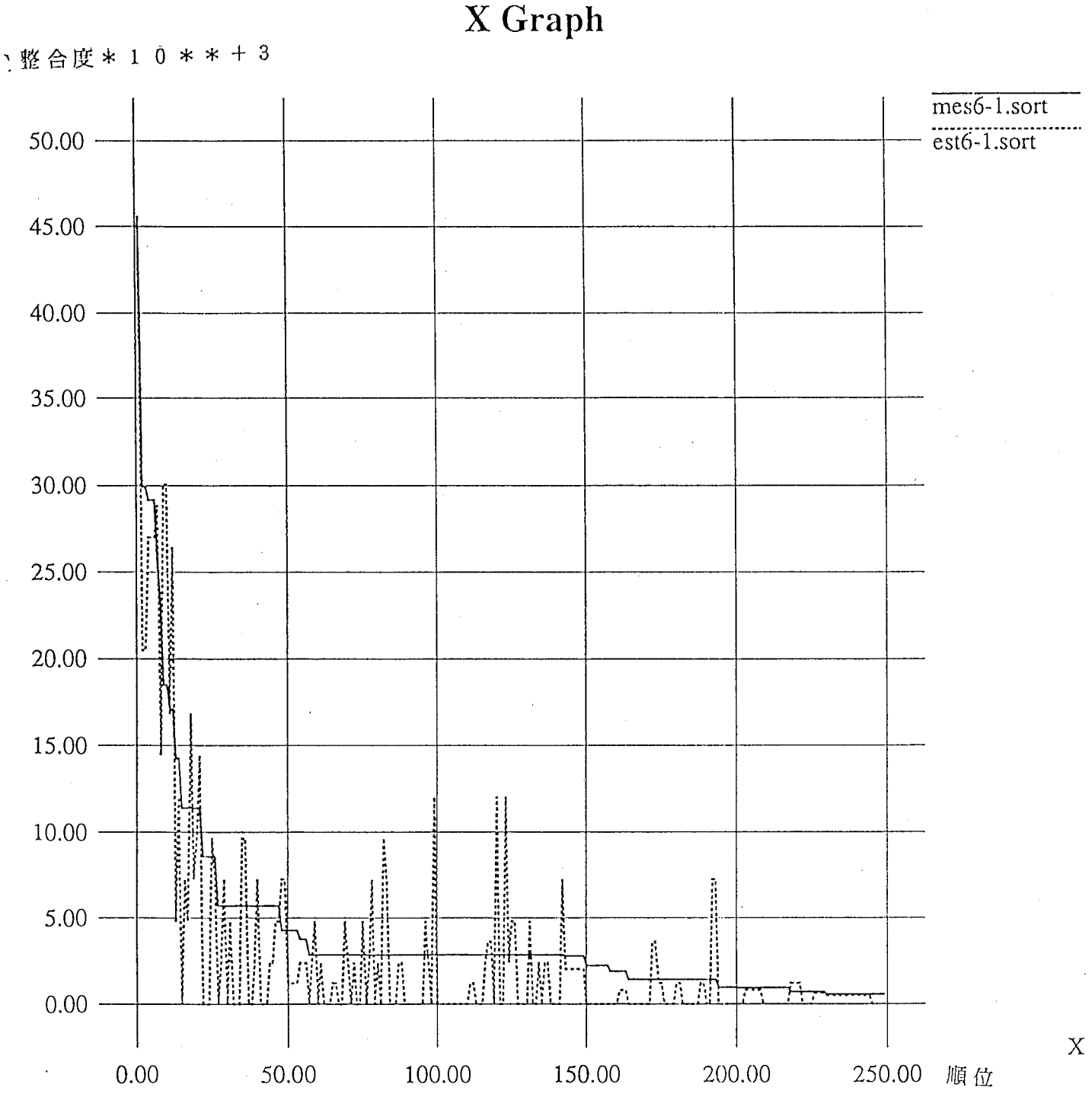


図３．７　実測整合度（実線）と推定整合度（点線）
w＝参加、k＝6，j＝1

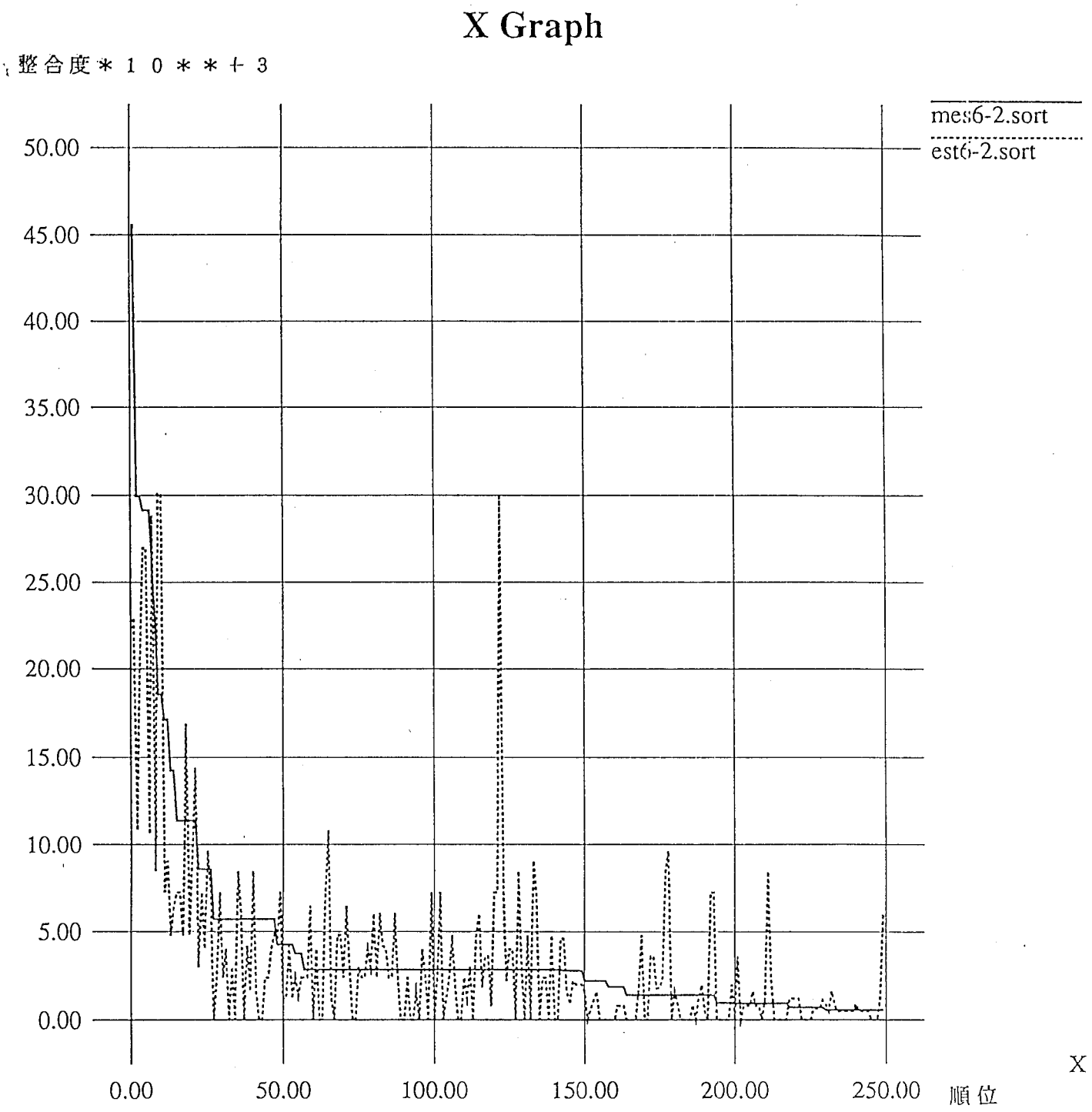


図３．８　実測整合度（実線）と推定整合度（点線）
ｗ＝参加、ｋ＝６，　ｊ＝２

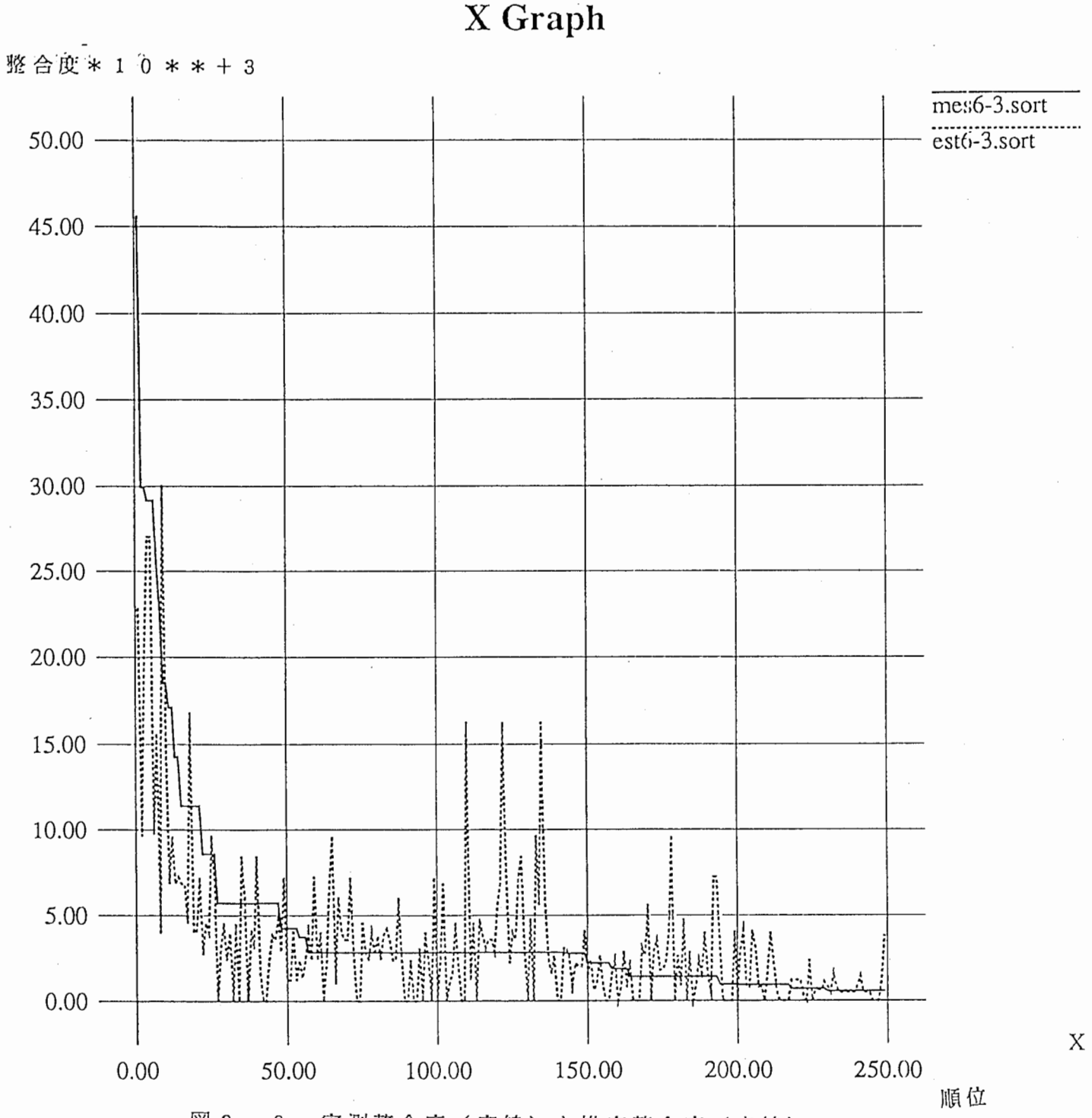


図３．９　実測整合度（実線）と推定整合度（点線）
ｗ＝参加、ｋ＝６，　ｊ＝３

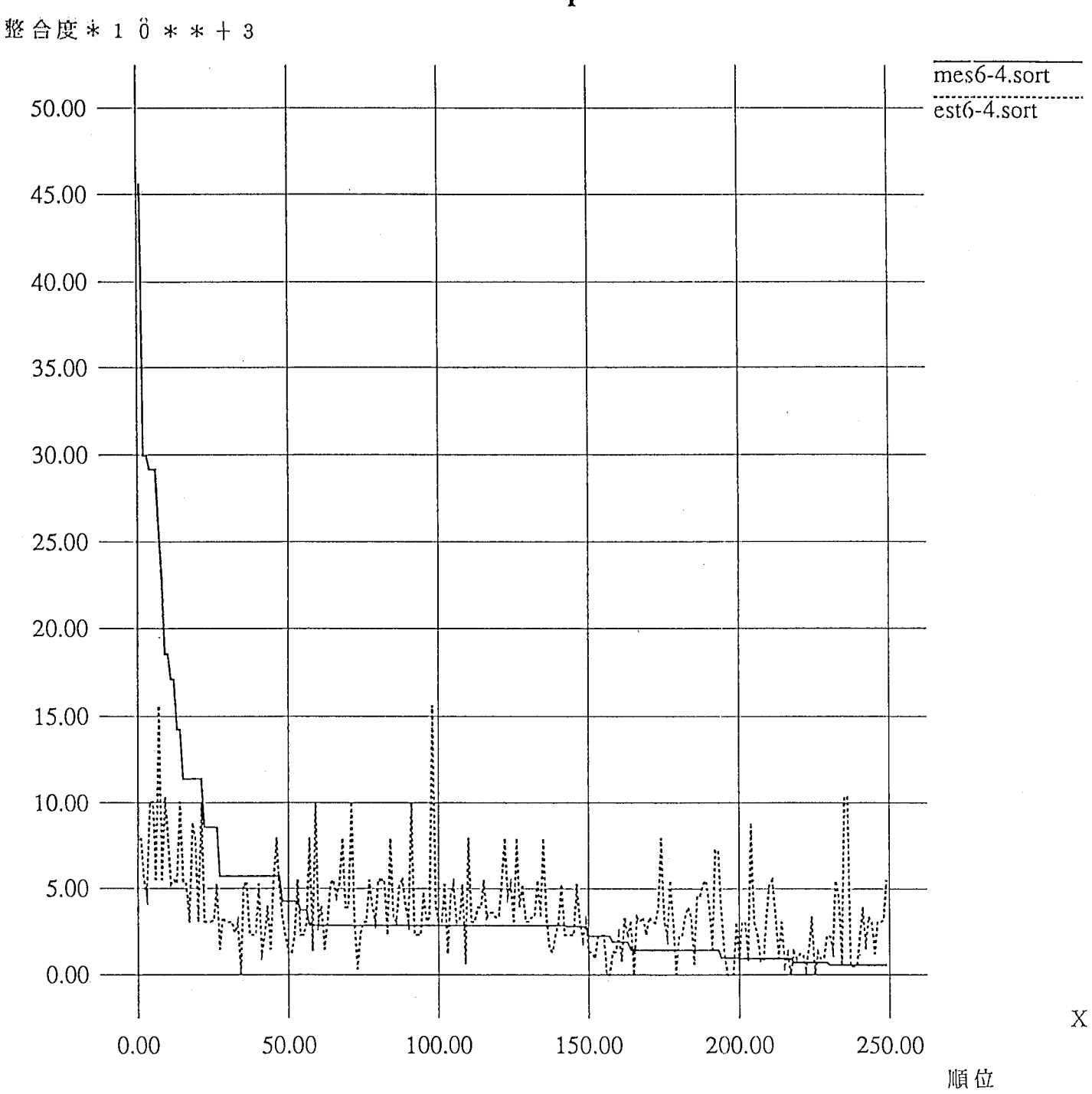


図３．１０　実測整合度（実線）と推定整合度（点線）
w＝参加、k＝６，j＝４

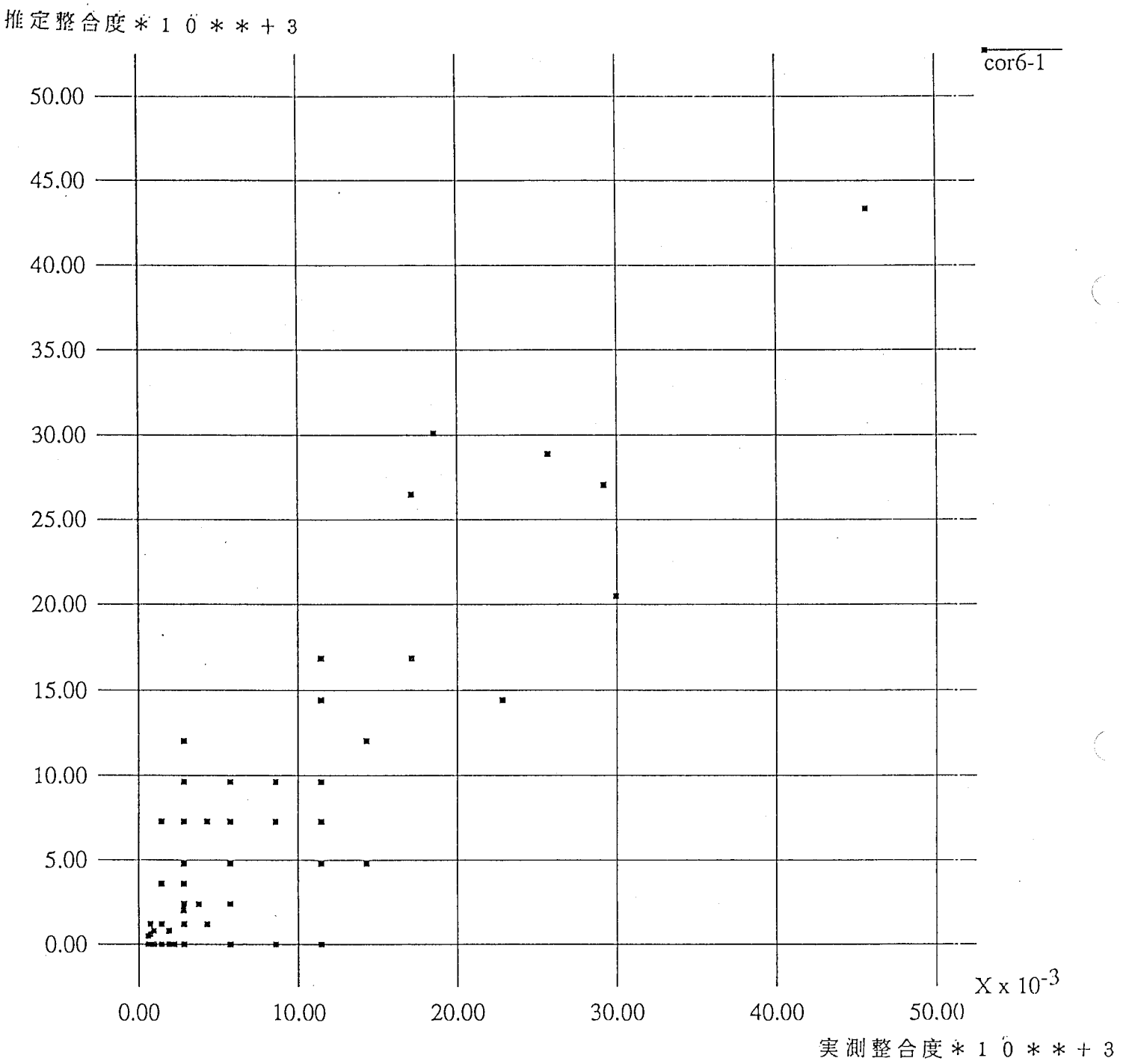


図３．１１　実測整合度と推定整合度の相関
ｗ＝参加、ｋ＝６，　ｊ＝１

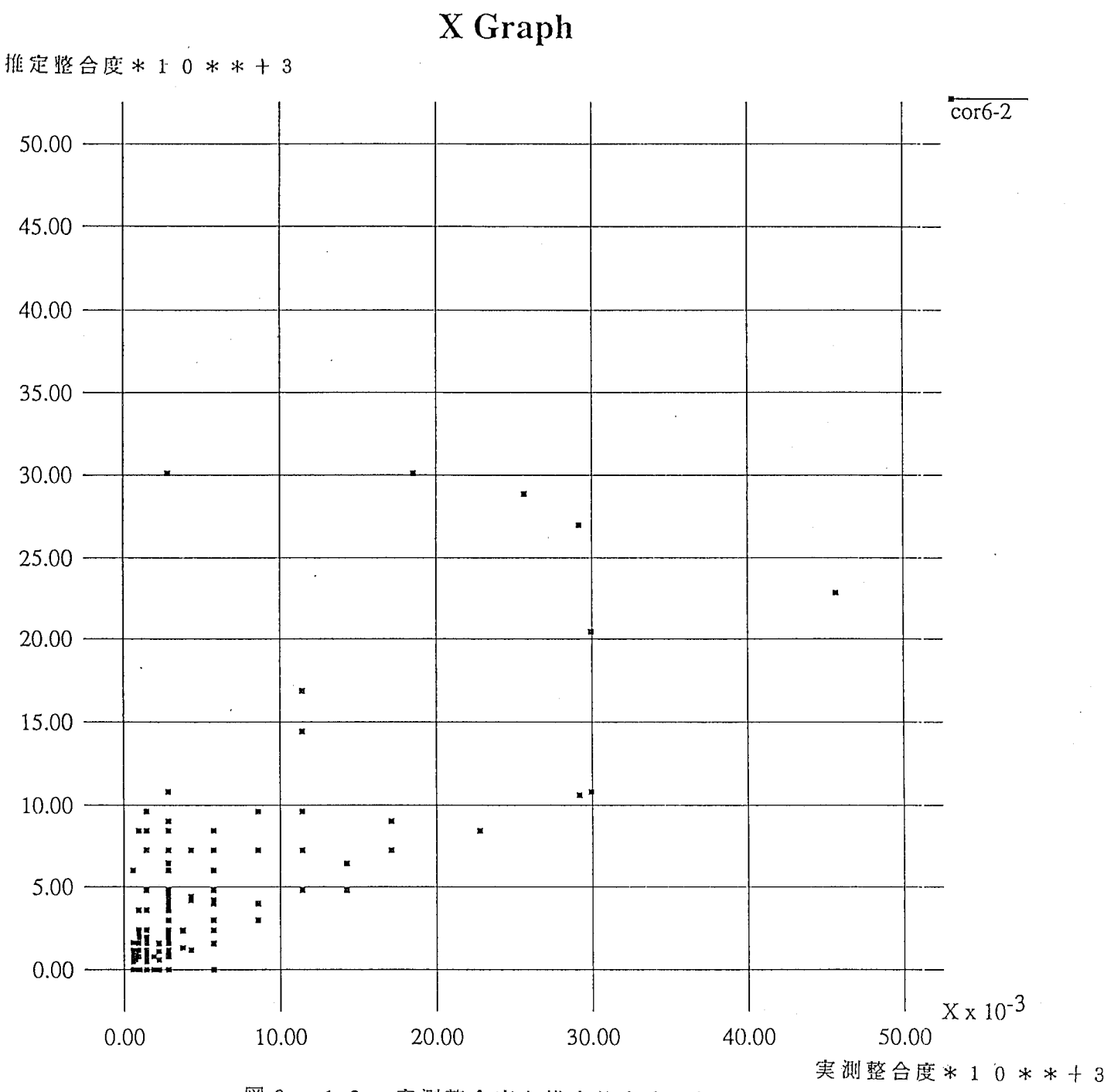


図３．１２　実測整合度と推定整合度の相関
ｗ＝参加、ｋ＝６，　ｊ＝２

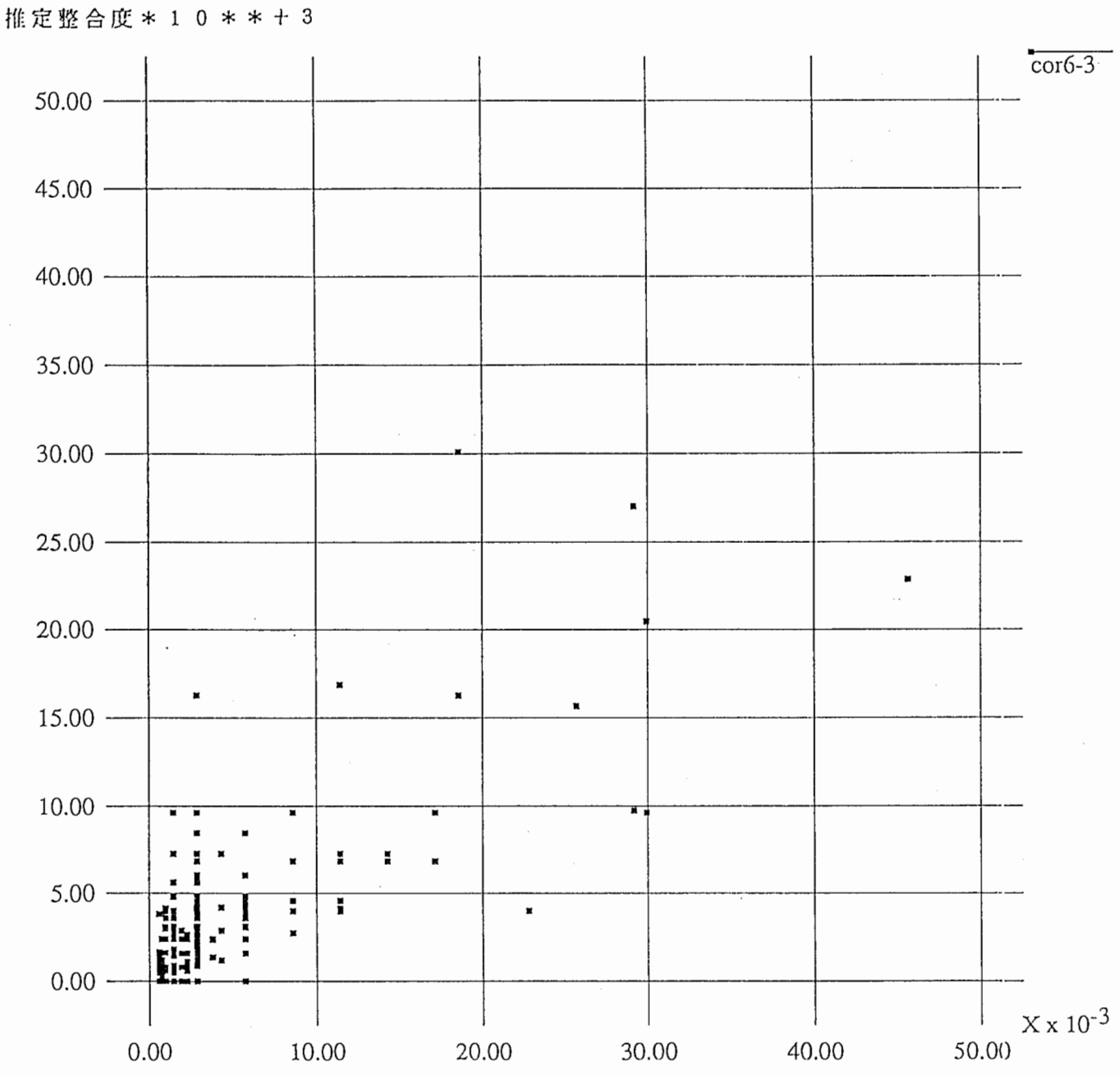


図３．１３　実測整合度と推定整合度の相関
ｗ＝参加、ｋ＝６，　ｊ＝３

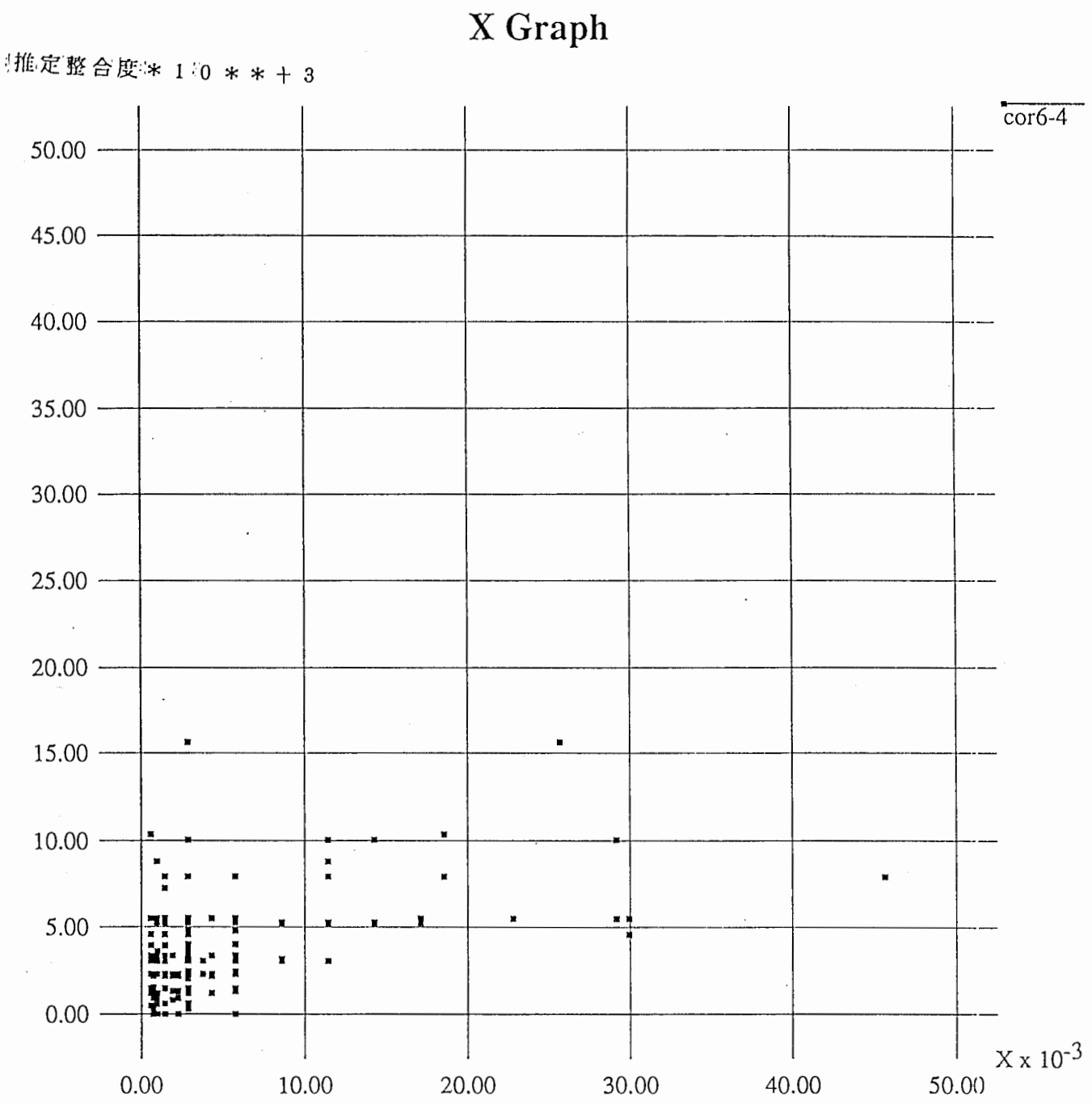


図３．１４　実測整合度と推定整合度の相関
ｗ＝参加、ｋ＝６，　ｊ＝４

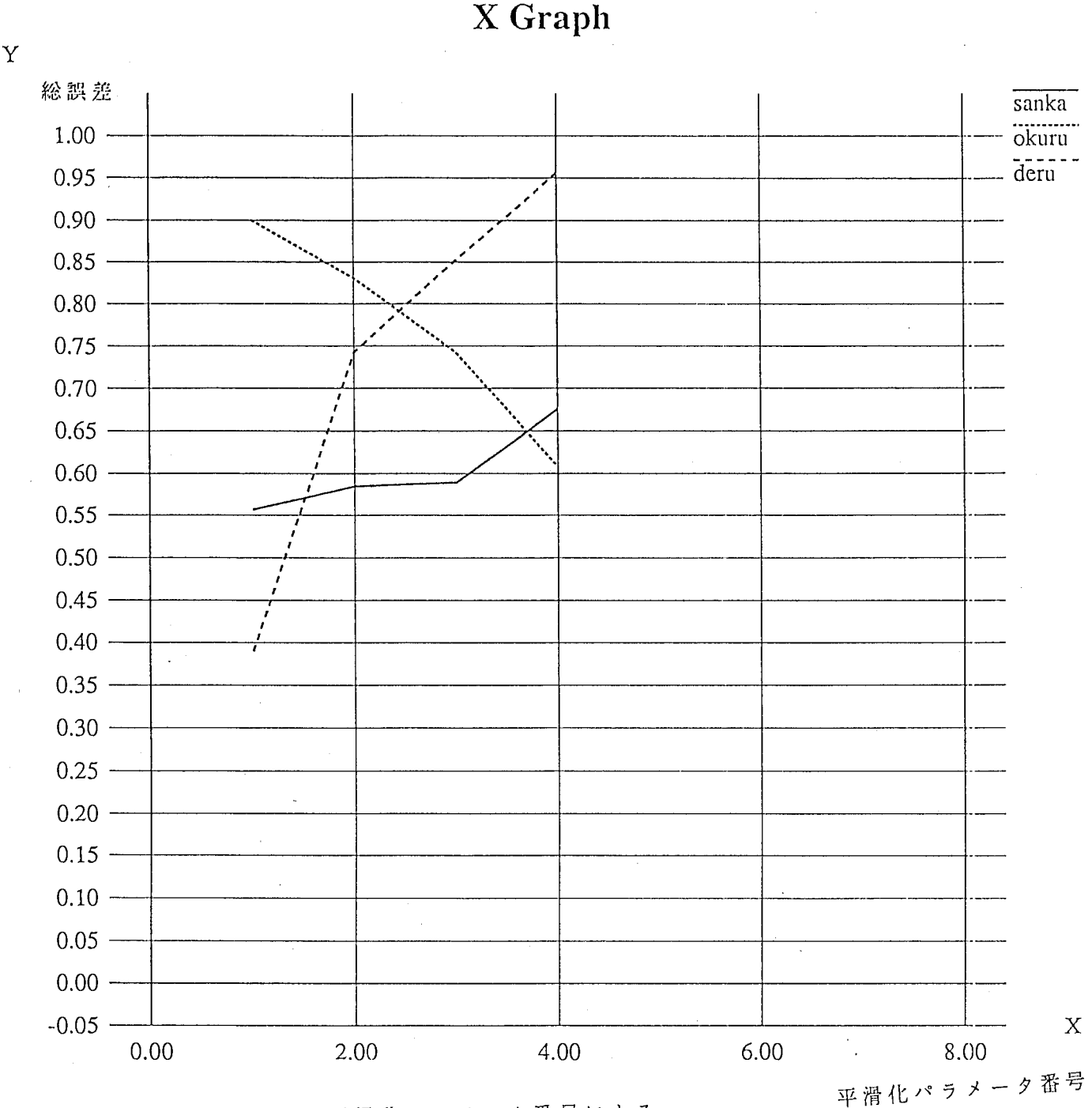


図３．１５　平滑化パラメータ番号による
総誤差の変化　1
述詞固定
実線：参加　点線：送る　破線：出る

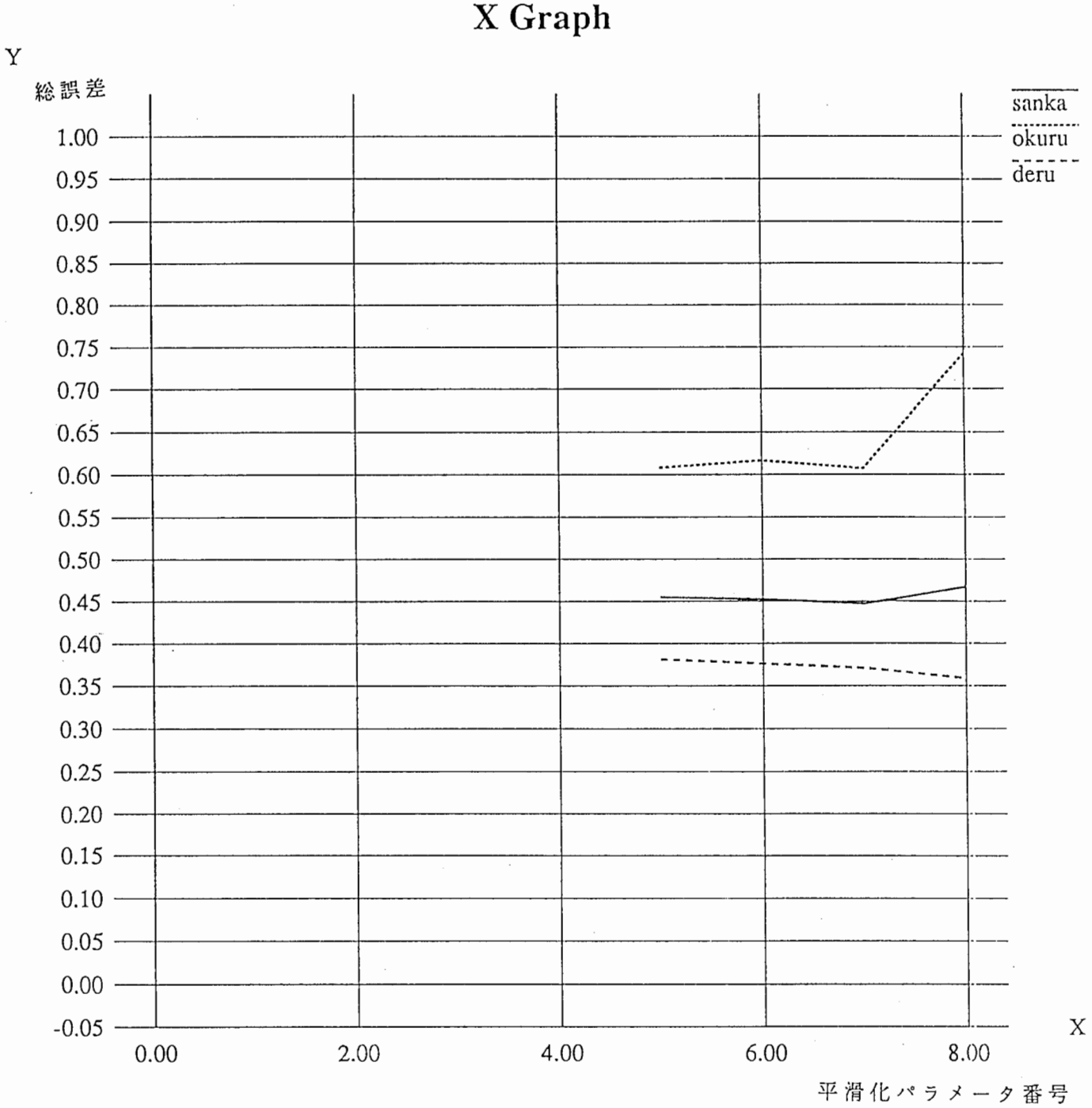


図３．１６　平滑化パラメータ番号による
総誤差の変化　２
述詞固定
実線：参加　点線：送る　破線：出る

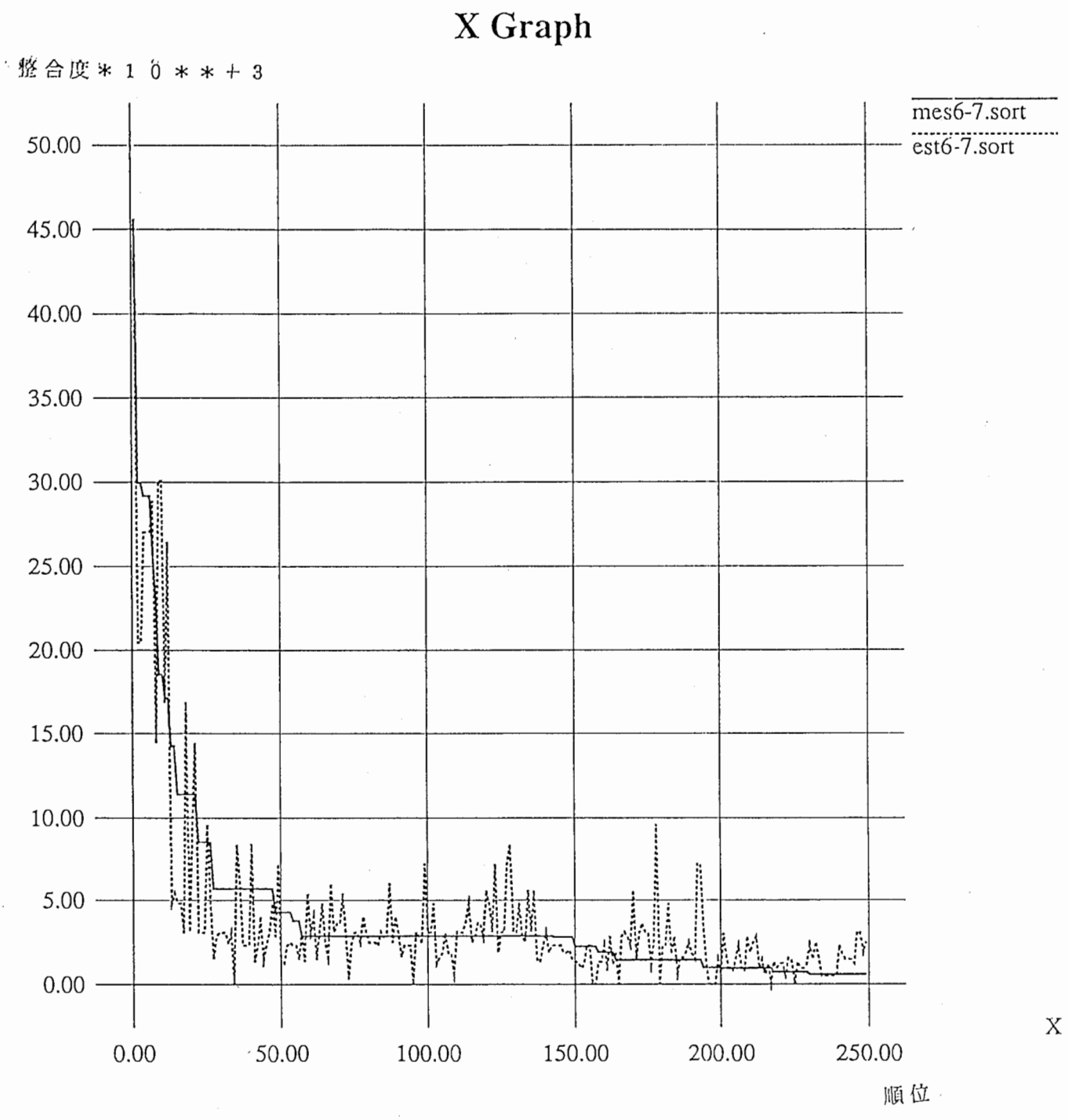


図３．１７　実測整合度（実線）と推定整合度（点線）
ｗ＝参加、ｋ＝６，　ｊ＝７

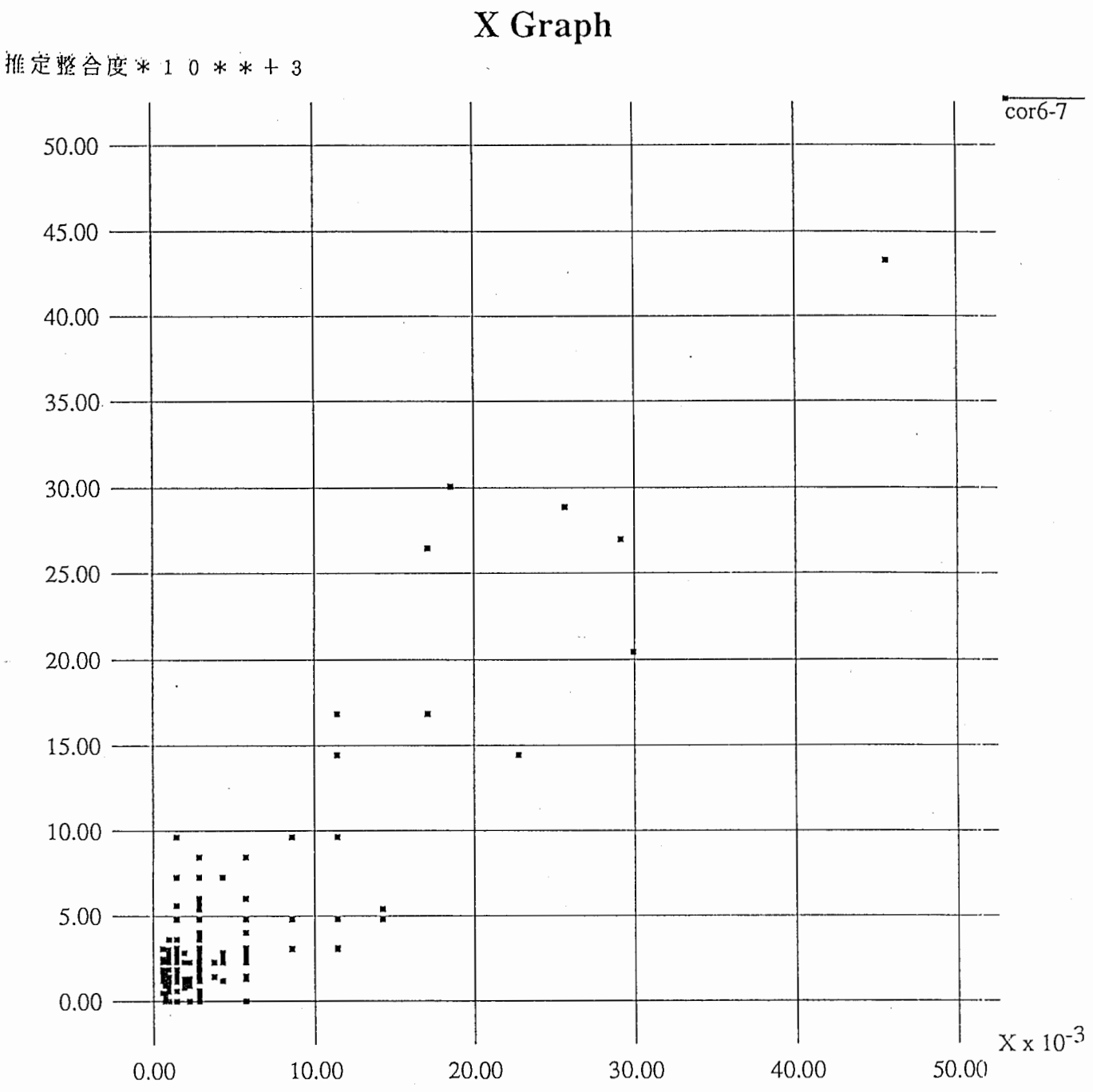


図３．１８　実測整合度と推定整合度の相関
ｗ＝参加、ｋ＝６，　ｊ＝７

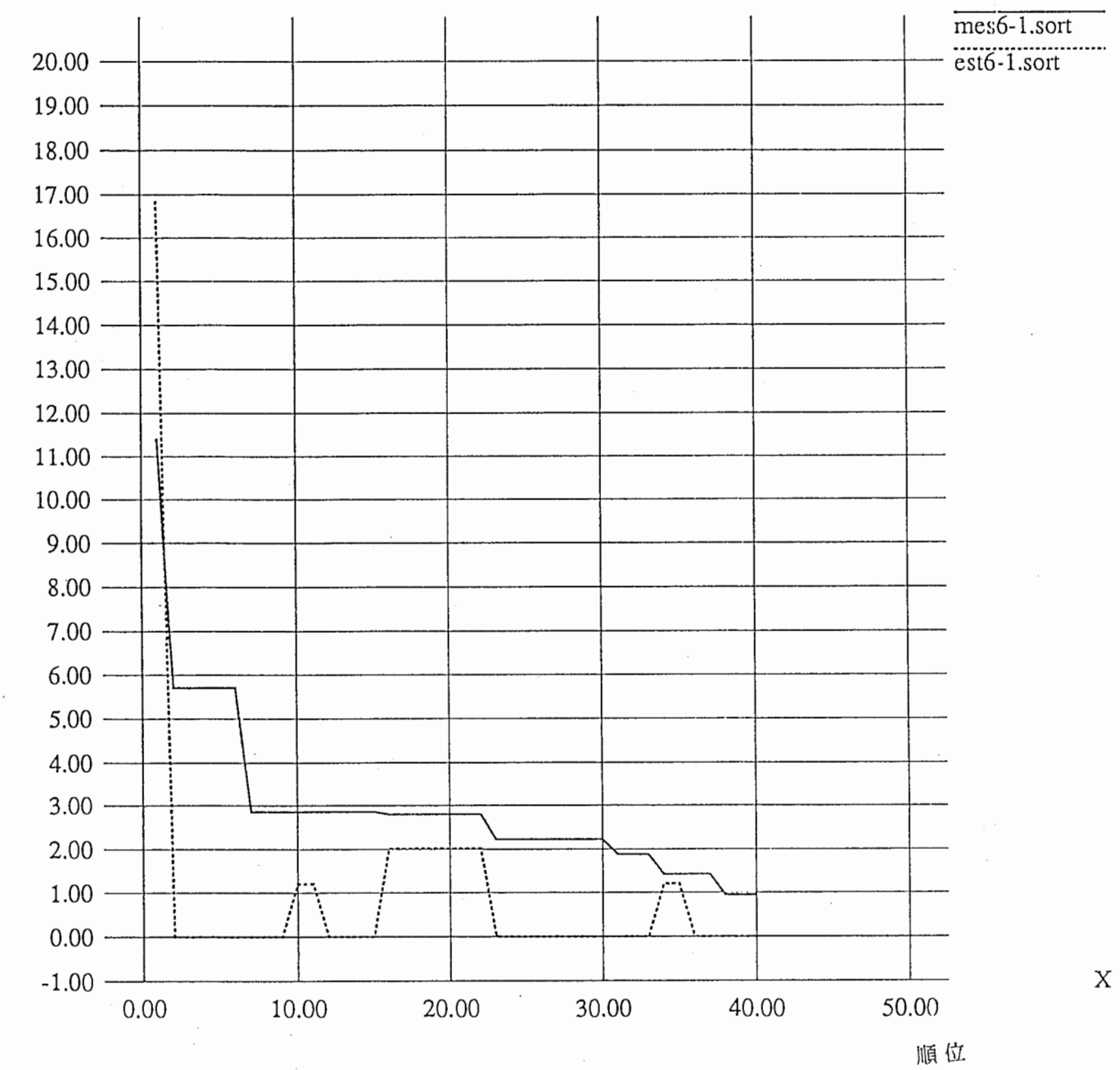


図３．１９　実測整合度（実線）と推定整合度（点線）
ｗ＝参加、ｋ＝６，　ｊ＝１
電話、キーに出現する係り元に限定

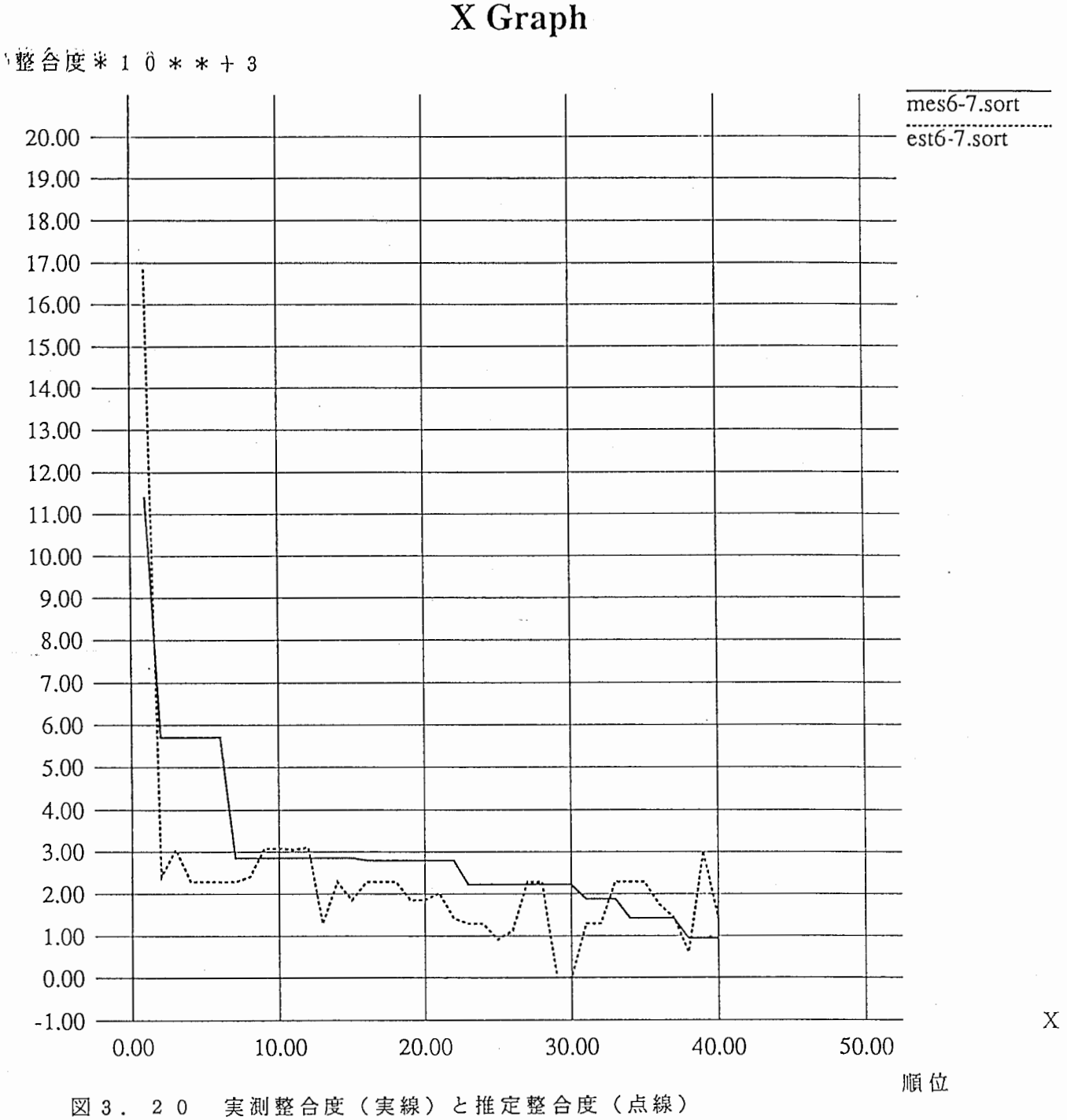


図３．２０　実測整合度（実線）と推定整合度（点線）
ｗ＝参加、ｋ＝６，ｊ＝７
電話、キーに出現する係り元に限定

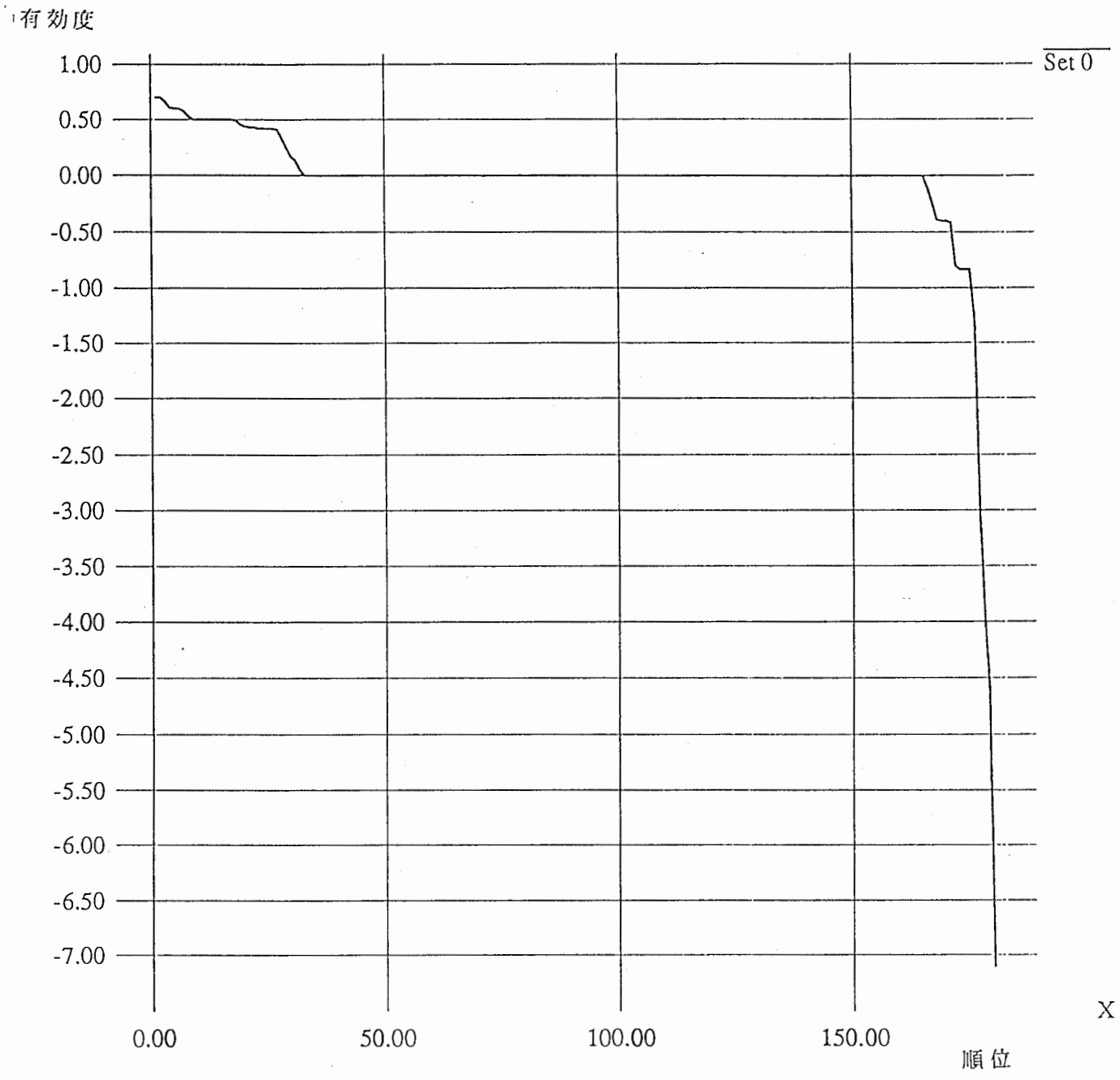


図３．２１　有効度の分布
ｗ＝参加、ｋ＝６，　ｊ＝２
分類番号４桁

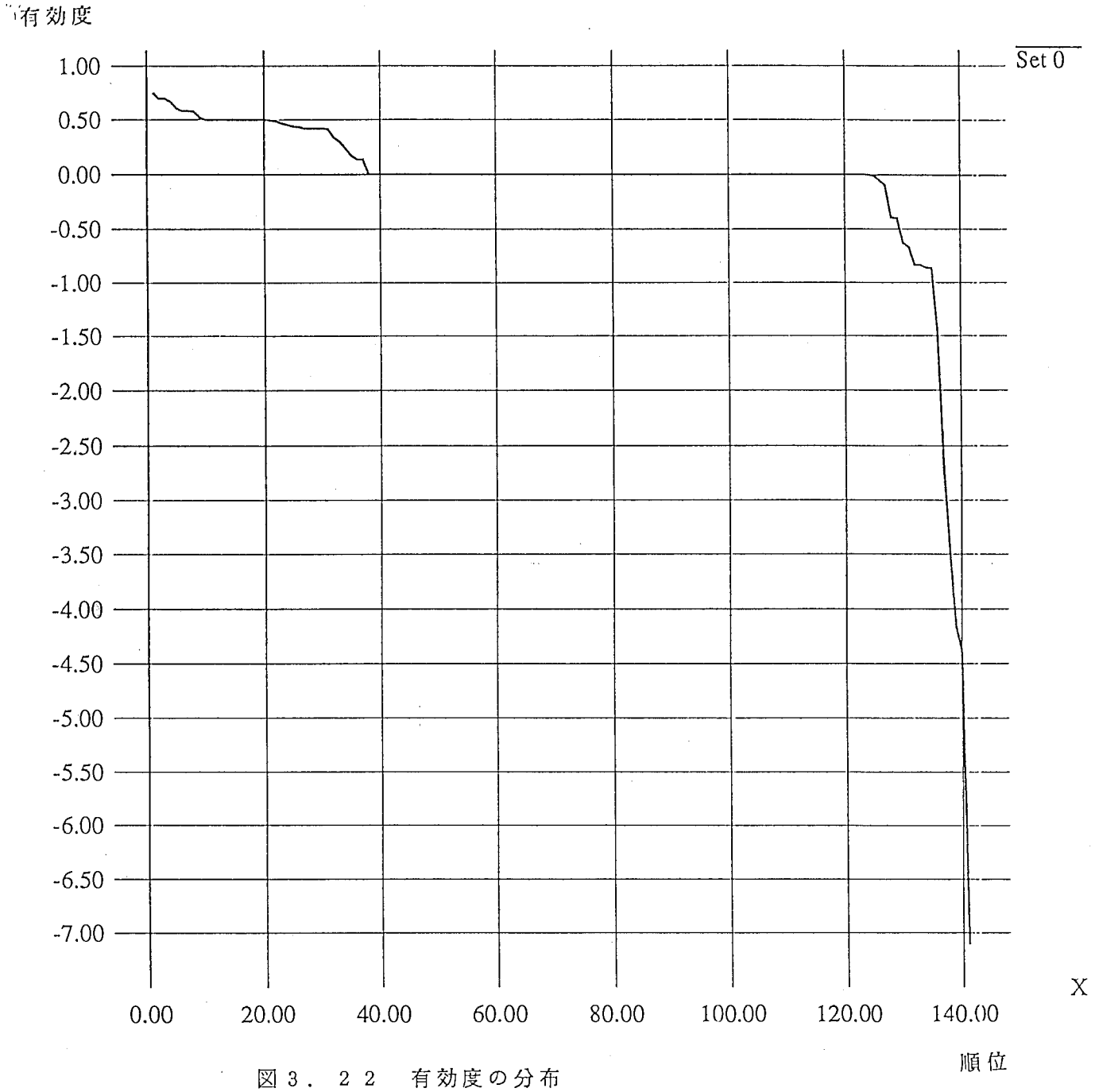


図３．２２　有効度の分布
ｗ＝参加、ｋ＝６，　ｊ＝３
分類番号３桁

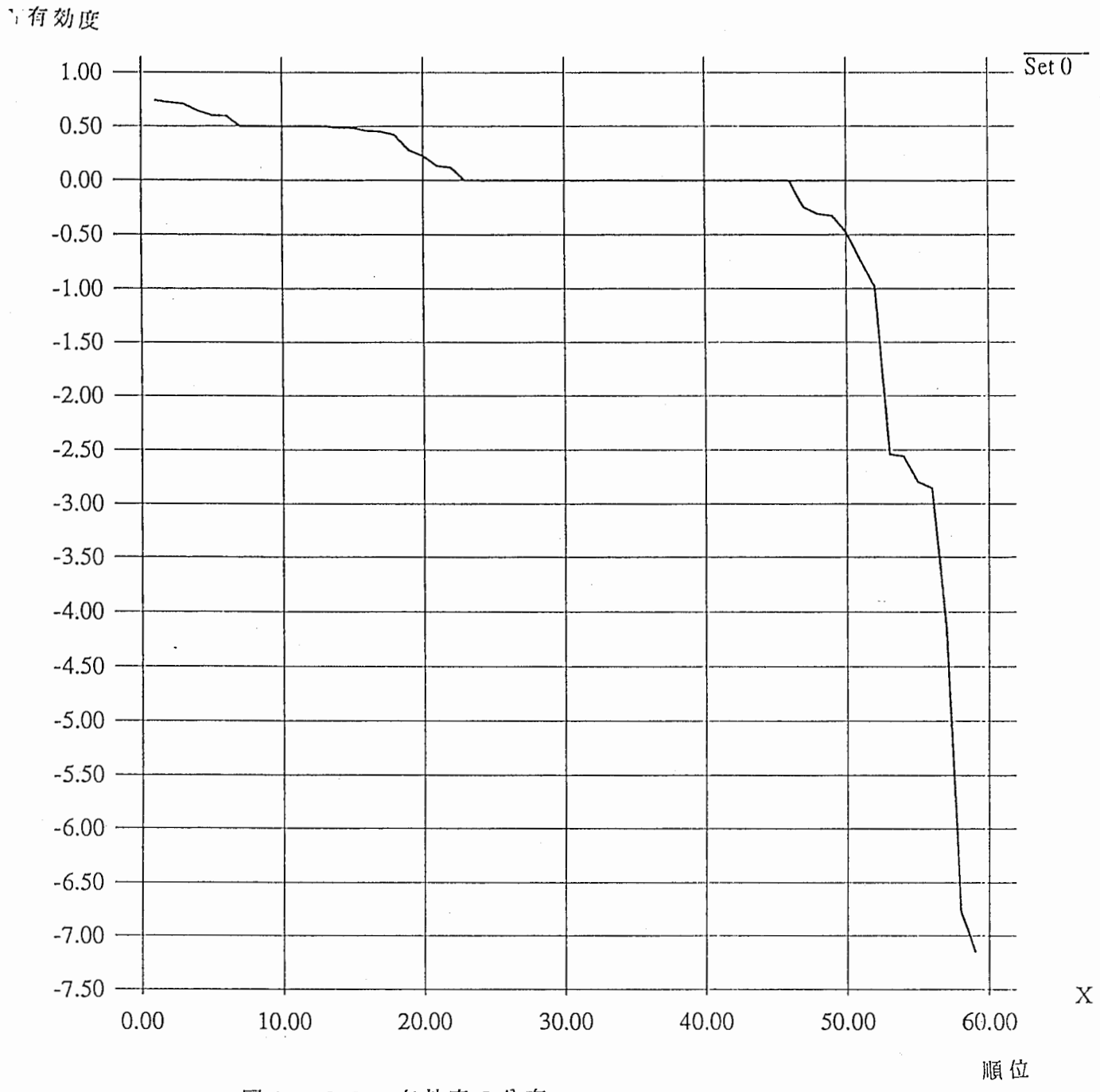


図３．２３　有効度の分布
ｗ＝参加、ｋ＝６，　ｊ＝４
分類番号２桁

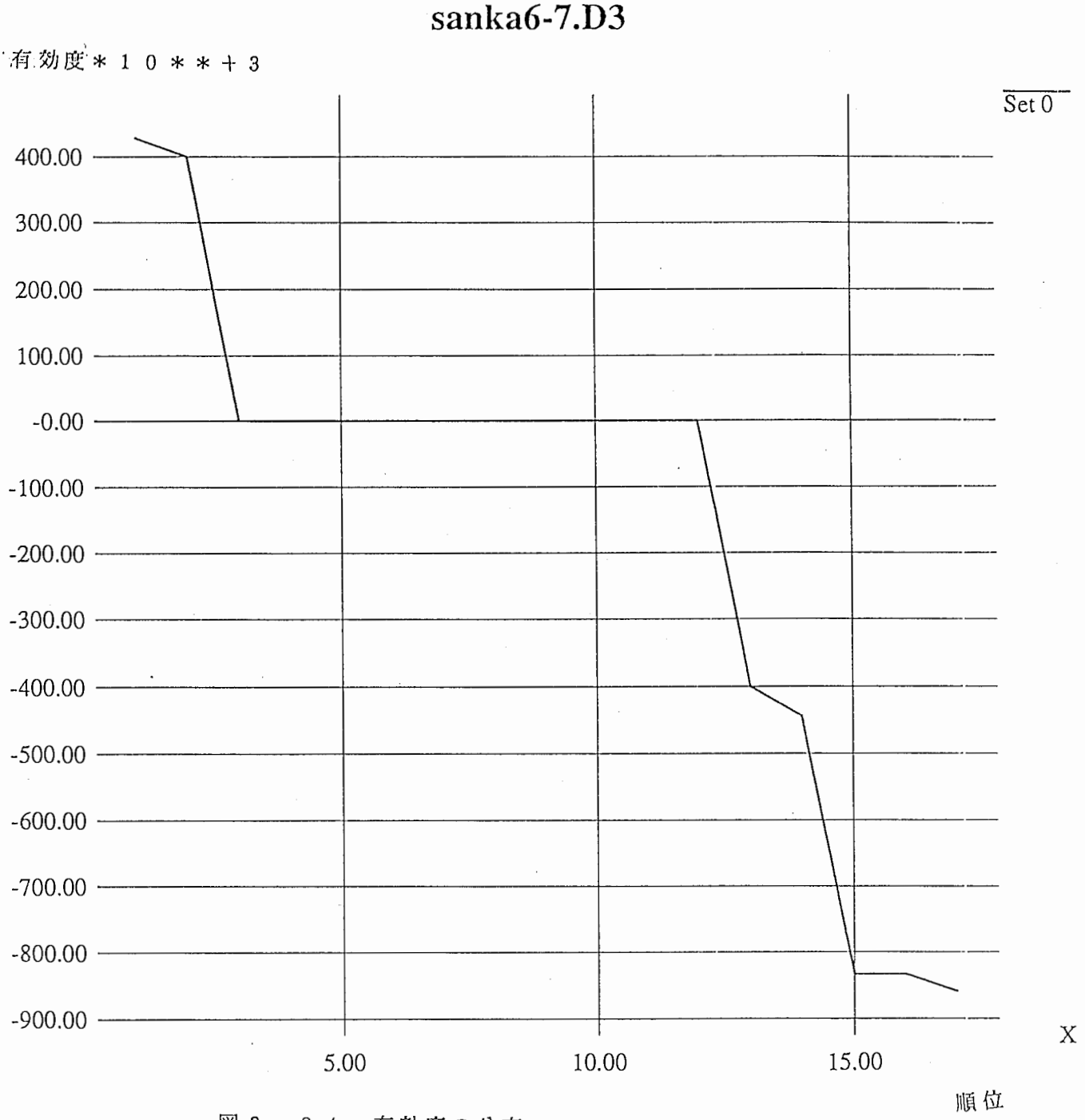


図３．２４　有効度の分布
ｗ＝参加、ｋ＝６，　ｊ＝７
分類番号３桁

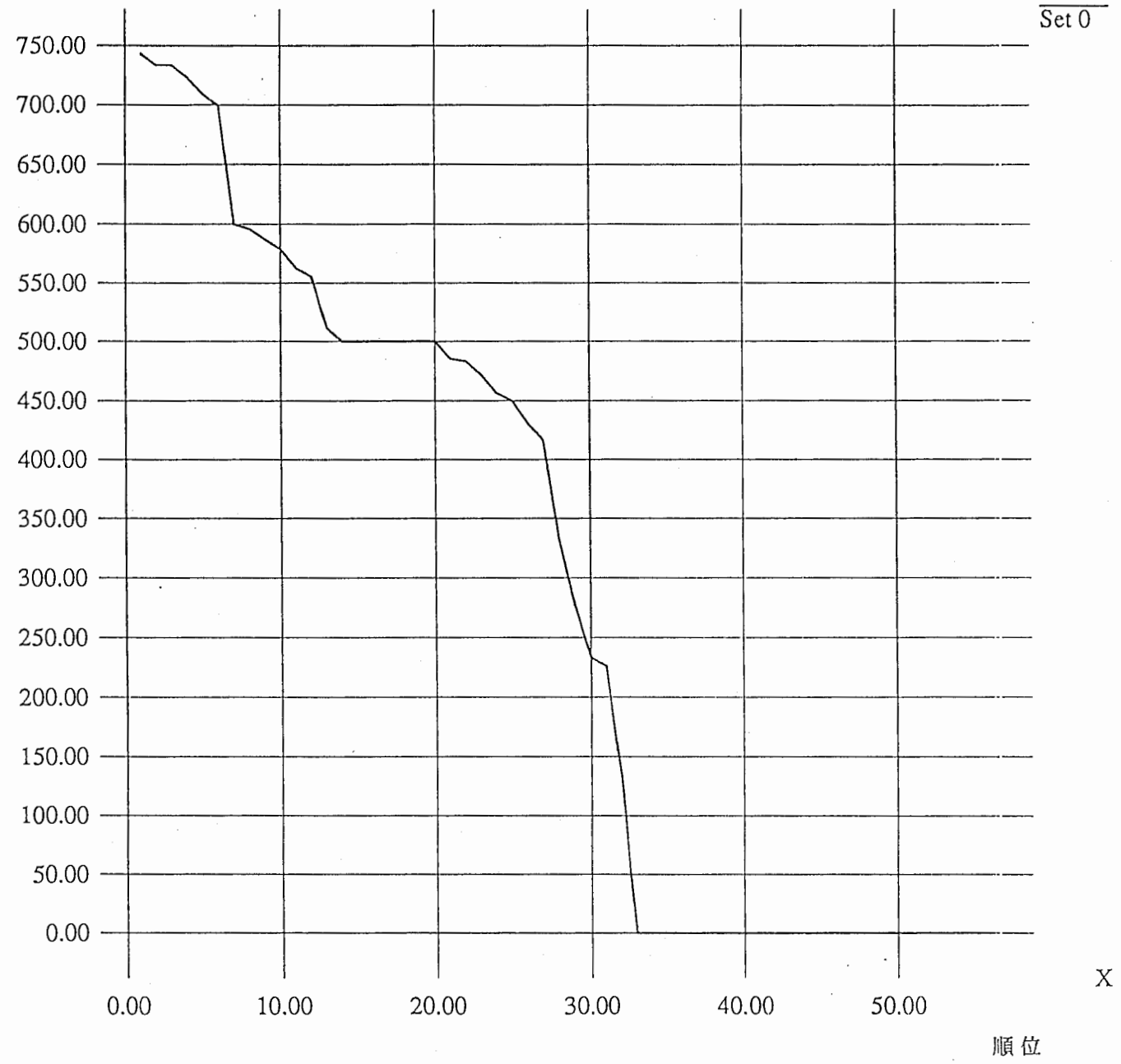


図３．２５　有効度の分布
ｗ＝参加、ｋ＝６，　ｊ＝７
分類番号２桁

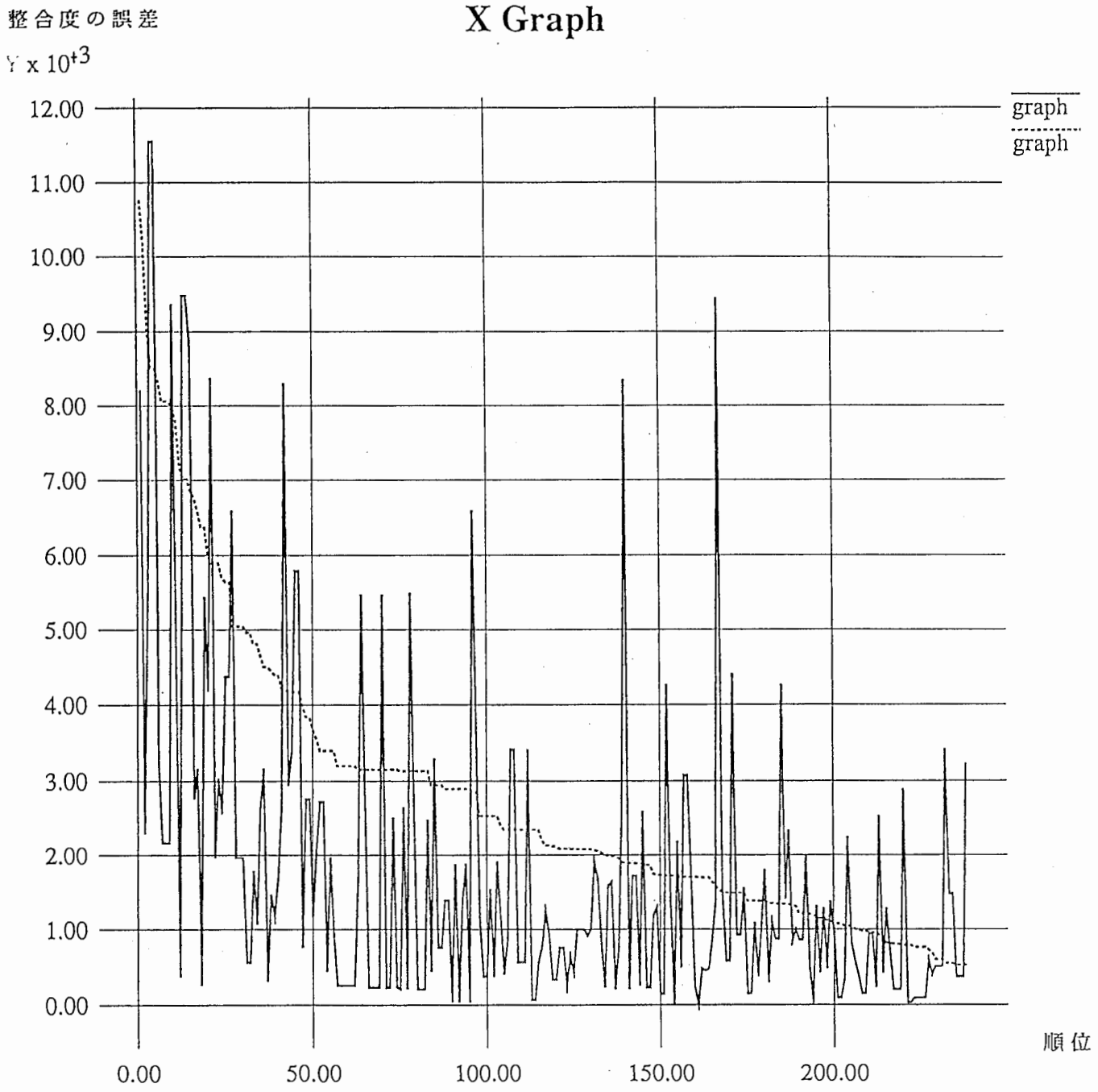


図３．２６　誤差推定値と誤差実測値の関係
ｗ＝参加、ｋ＝６，ｊ＝７
実線：実測値　点線：推定値

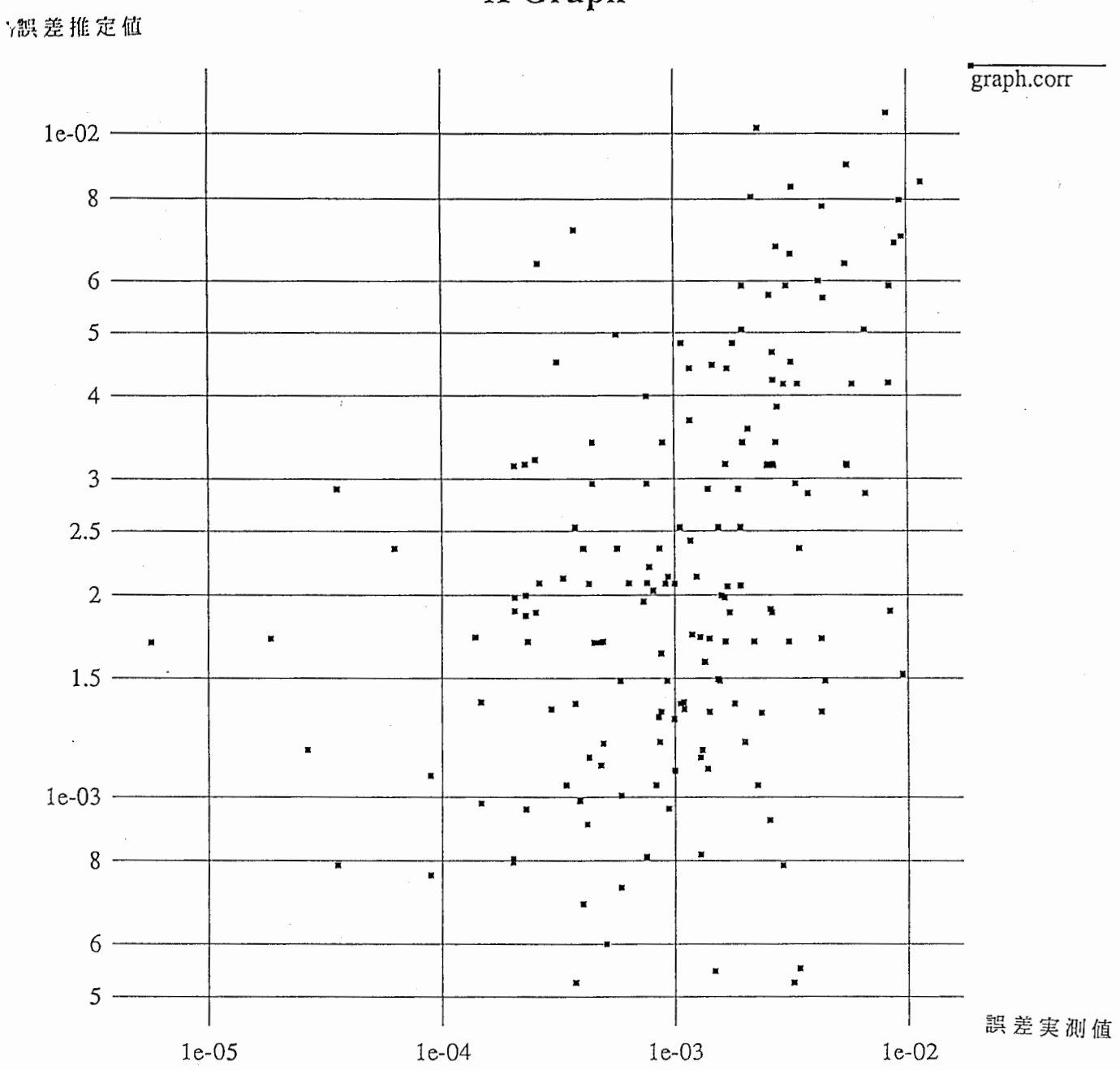


図３．２７　誤差推定値と誤差実測値の相関
ｗ＝参加、ｋ＝６，ｊ＝７

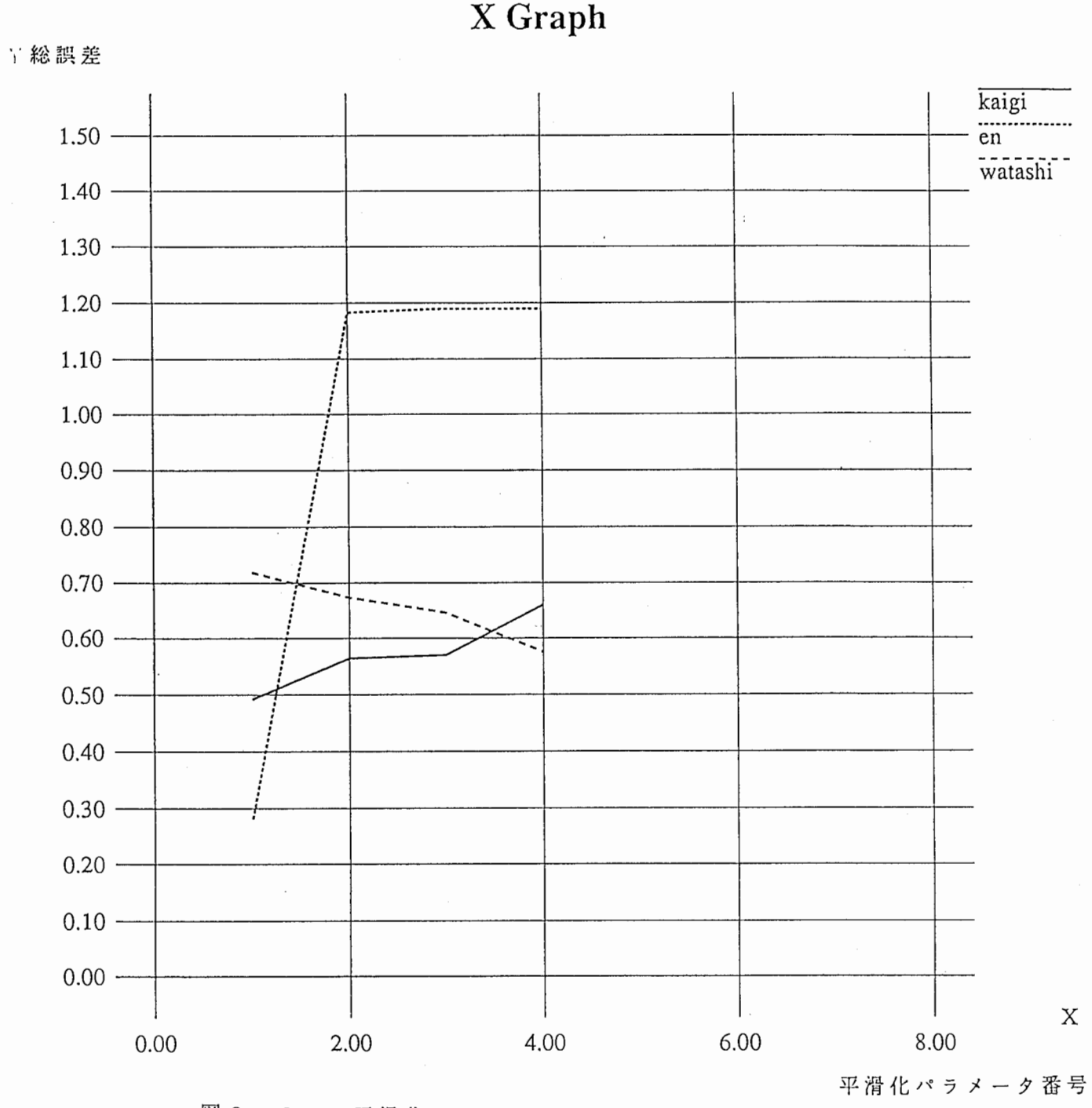


図３．２８　平滑化パラメータ番号による
総誤差の変化　1
名詞固定
実線：会議　点線：円　破線：私

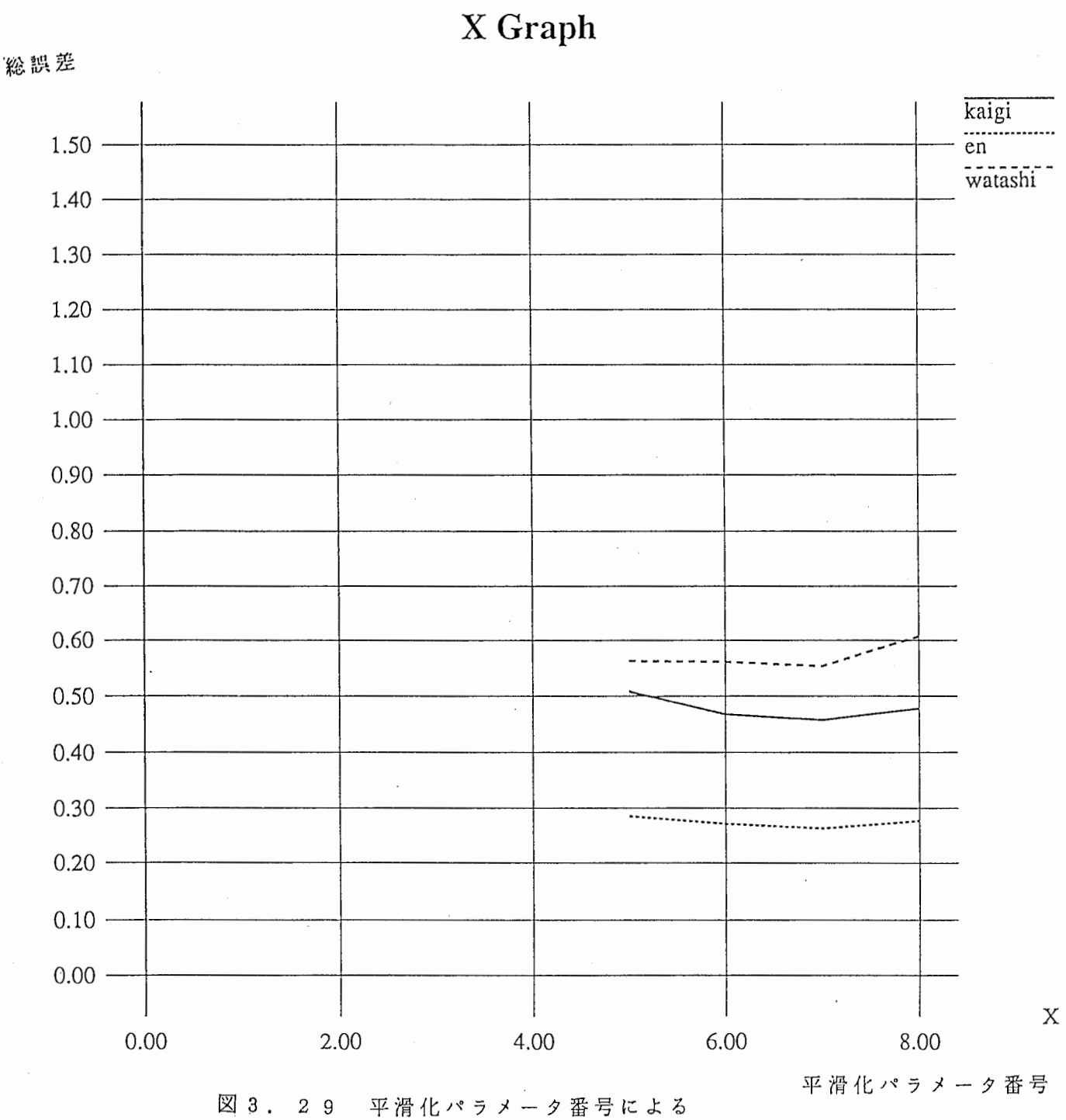


図３．２９　平滑化パラメータ番号による
総誤差の変化　２
名詞固定
実線：会議　点線：円　破線：私

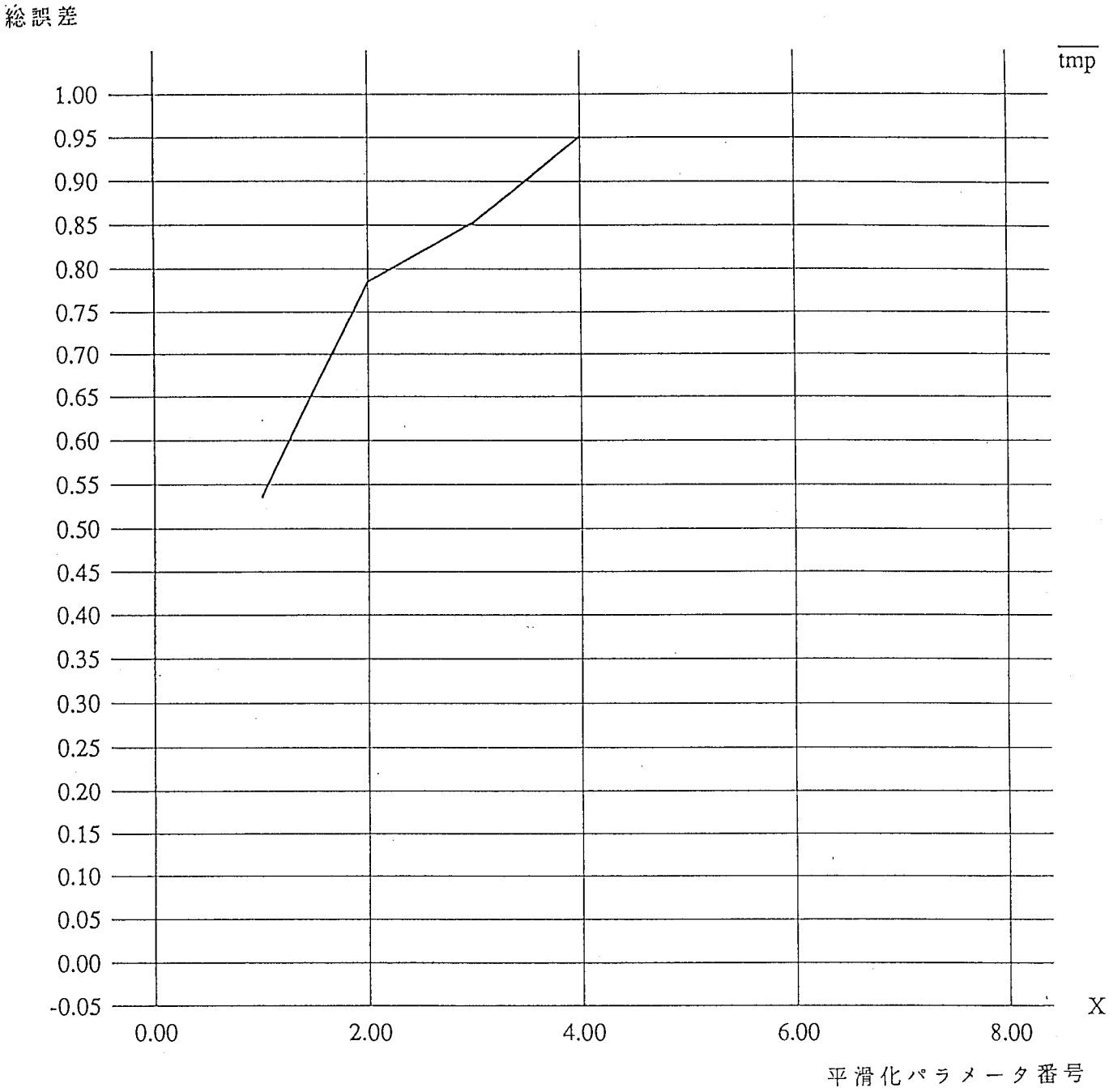


図３．３０　平滑化パラメータ番号による総誤差の変化　1

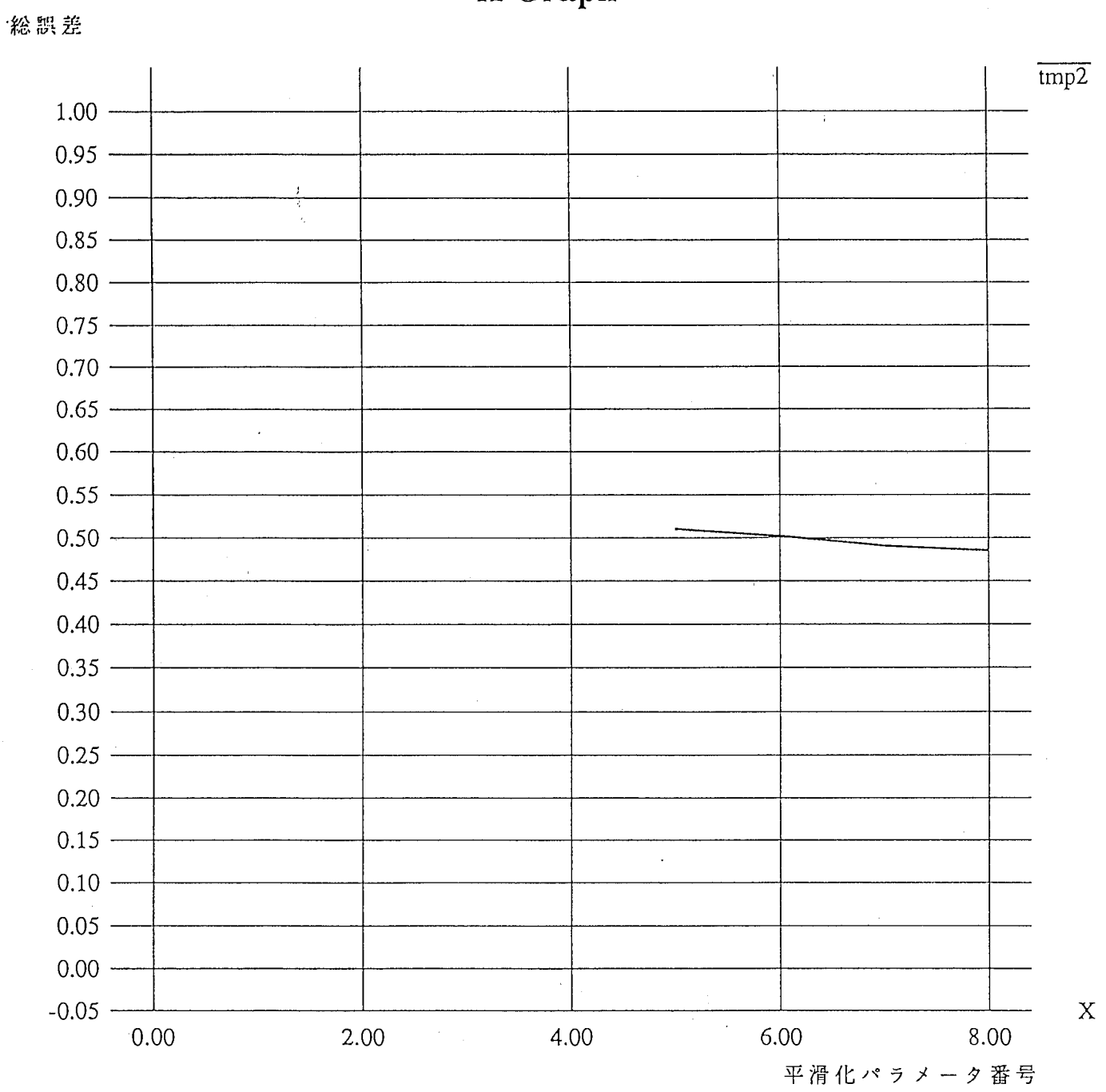


図３．３１　平滑化パラメータ番号による総誤差の変化　2

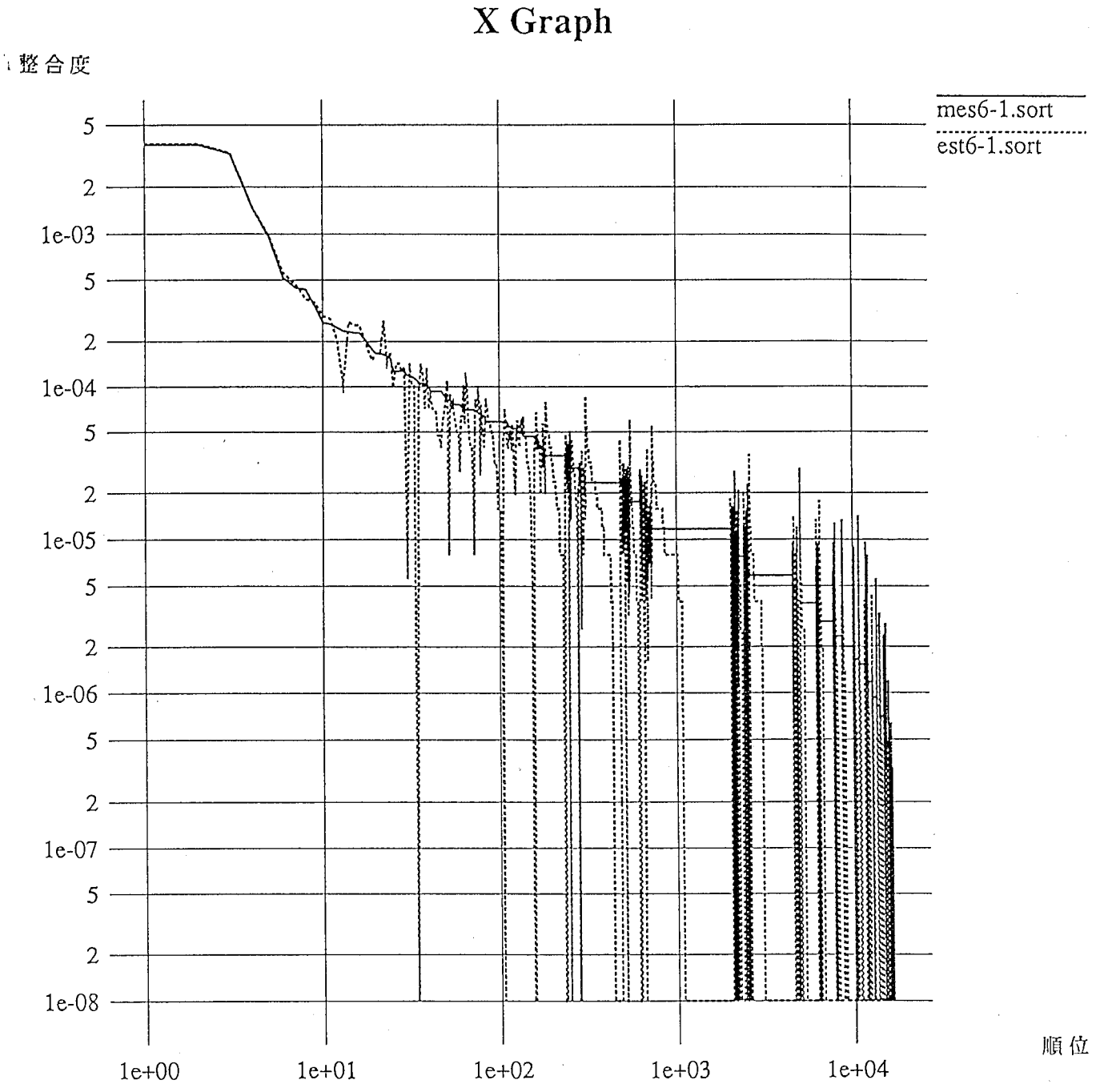


図３．３２　実測整合度（実線）と推定整合度（点線）
k = 6，　j = 1

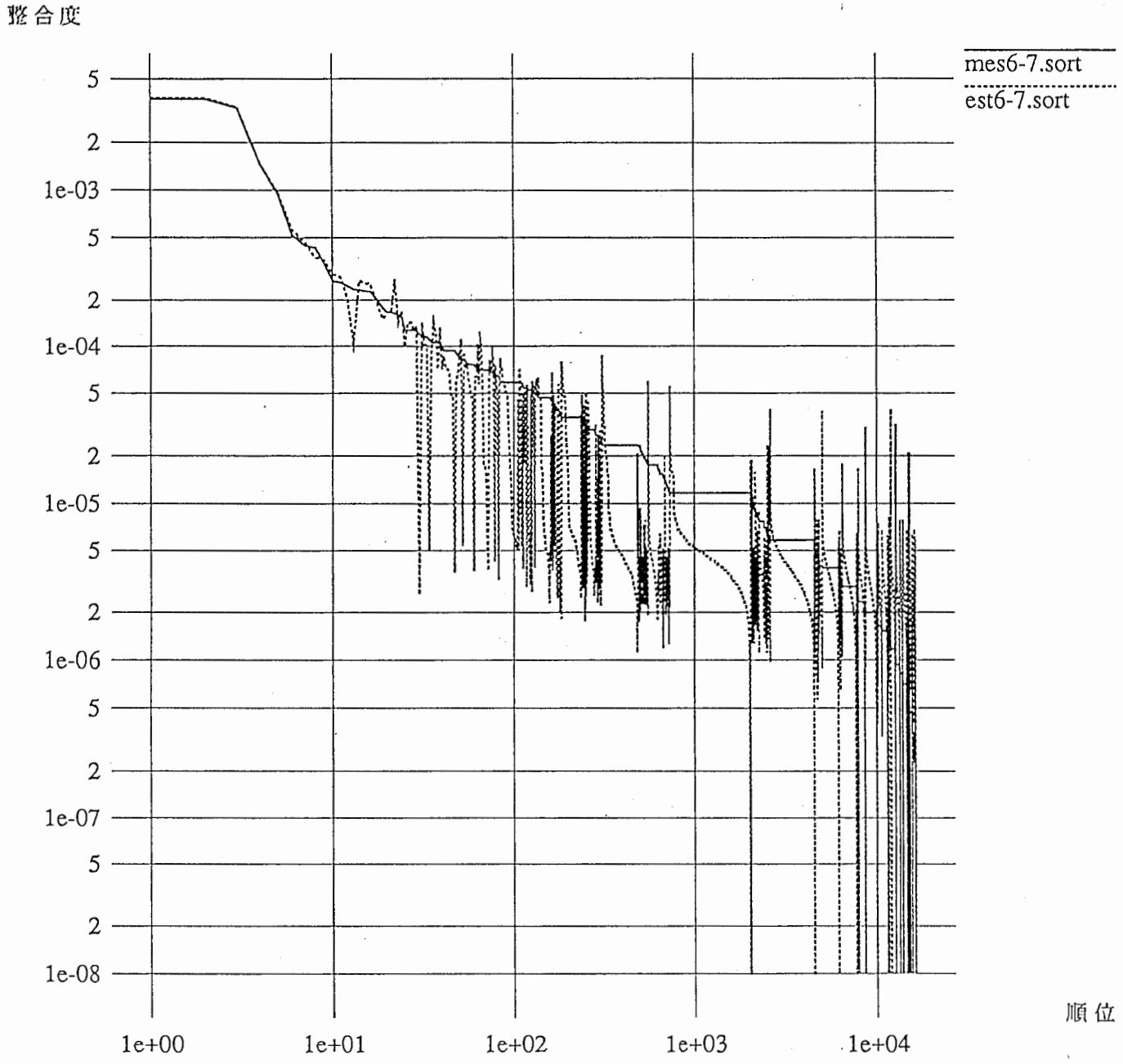


図３．３３　実測整合度（実線）と推定整合度（点線）
k = 6，　j = 7

４．曖昧性解消への利用

本手法は冒頭にも述べたように、構文的曖昧性の解消に利用し得る。ちなみに、１章の（１．２）（１．３）の例をデータ組合せ番号ｋ＝６を用いて、整合度計算をして見ると次の様になる。ただし、ここでは、分類番号が複数付与されているものの中から、正しく付与されているもののみ取り出して計算した。平滑化パラメータ番号ｊ＝１、つまり、語形と分類番号４桁を用いて整合度を求めると、

Ｓ（鳥 061c、が、飛ぶ 218_）＝ 0.00000103
Ｓ（鳥 061c、が、見る 331_）＝ 0
Ｓ（鳥 061c、が、曇る 022c）＝ 0　　（４．１）

となる。これでは、（１．２）の曖昧性は解消できても、（１．３）の曖昧性は解消できない。しかし、ｊ＝７の場合は

Ｓ（鳥 061c、が、飛ぶ 218_）＝ 0.00000181
Ｓ（鳥 061c、が、見る 331_）＝ 0.00000150
Ｓ（鳥 061c、が、曇る 022c）＝ 0.00000103　　（４．２）

となる。よって、（１．５）（１．６）が成立しており、（１．２）（１．３）とも、構文的曖昧性が正しく解消できる。

本手法は、依存文法［児玉］では自然に利用できるが、構成文法［児玉］でも、主辞が定義されたものでは利用可能である。この点を説明する。文脈自由形の構成文法の規則は、例えば、

Ａ　－＞　Ｘ[1]，Ｘ[2]，・・・，Ｘ[n]　　（４．３）

となる。Ａは非終端記号であり、Ｘ[i]（ｉ＝１，．．．，ｎ）は終端または非終端記号である。構成素Ｘ[i]（ｉ＝１，．．．，ｎ）の中でＸ[j]がこの規則の主辞であるとする。この様に各規則に主辞が１つ定義されていると、構文解析木に、主辞の連鎖が定義でき、木の各節点から主辞をたどって行くことで、ついには、語彙項目に到達する。この語彙項目のことを当該節点の主語彙項目と呼ぶ。このとき、次の規則によって、語彙項目間の係り受け関係が定義できる。

規則：　構文解析木の全ての節点Ａに対して、Ａを展開する規則の主辞Ｘ[j]以外の構成素の主語彙項目は、すべて、Ｘ[j]の主語彙項目に係る。
（４．４）

という規則を用いる。

例えば、文（１．３）が以下に示す主辞付構成文法で解析される場合を見てみよう。ここで、規則の主辞はアンダーラインで示されている。ただし、語彙規則は構成素が１つしかないため主辞が明かであるから、アンダーラインは除いてある。

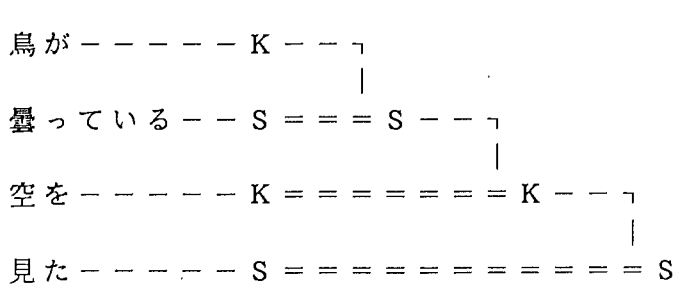
Ｇ＝（Ｖ，Ｔ，Ｐ，Ｓ）

Ｖ＝｛Ｓ，Ｋ｝
Ｔ＝｛鳥が、曇っている、空を、見た｝

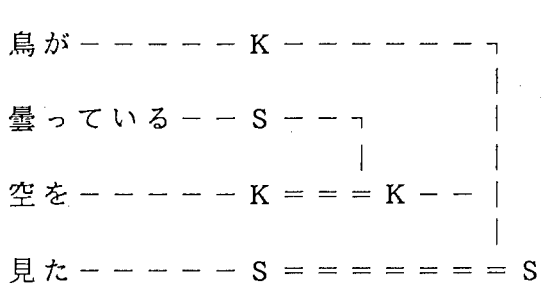
P = {
K －＞鳥が
K －＞空を
S －＞曇っている
S －＞見た
S －＞K　S
S －＞K　K　S
K －＞S　K
}

解析結果は以下の様になり、主辞が2重線で示されている。

解析結果A



解析結果B



そうすると、（4．4）の規則によって、語彙項目間の係り受け関係が次のように定義される。

解析結果A

「鳥が、曇っている」
「曇っている、空を」
「空を、見た」

解析結果B

「鳥が、見た」
「曇っている、空を」
「空を、見た」

AとBの差は「鳥が、曇っている」と「鳥が、見た」だけであるから、（4．2）式の整合度を用いて、Bが選択される。

構文的曖昧性解消と並んで、自然言語処理の課題として、語彙的曖昧性解消がある。例えば、「教授」には、「人間としての教授（572_）」と「教えることそのもの（455d）」の語義がある。「教授が参加する」は前者の語義が正しい。本実験では、これらの意味分類番号が均等に出現すると仮定して、データを作成しているので、語彙的曖昧性は解消できないと考えるのがもっともであるが、「教授」以外の語の係り受け度数との関連で解消できる可能性がある。つまり、分類番号572_を持つ他の語が「参加する」に係る度数の方が、分類番号455dを持つ他の語が「参加する」に係る度数よりも大きければ、分類番号を利用することによって、語「教授」に対しても語彙的曖昧性が解消できる。そこで、ｋ＝６、ｊ＝７の場合に両者の整合度を計算すると、

S（教授 572_、が、参加 712_）＝ 0.00000466
S（教授 455d、が、参加 712_）＝ 0.00000308

である。そこで、前者の整合度の方が大きく、実際に語彙的曖昧性が正しく解消されることが分かる。

もちろん、これらは１例に過ぎず、大量の試験データを用いて実験をしなければ、本手法を曖昧性解消に利用したときの精度は評価できない。

５．関連研究との比較

カテゴリーを用いて、まばらなデータを平滑化する手法には、種々の方法が提案されている。これらの方法はカテゴリーをどのように定義するかということと、定義されたカテゴリーをどのように利用して整合度を求めるかということで分類できる。

［Church］で述べられているいくつか方法は、Bigramまたは Unigramの度数に基づいてカテゴリーを定義している。整合度は各カテゴリー毎に推定したものを用いている。Bigram を対象にはしているが、係り受けにも利用し得る方法である。しかし、これらの方法は意味カテゴリーを利用していないため、意味的な整合性が考慮されないことになり、精度に問題が生ずるのではないかと思われる。

一方、［Bahl］の方法は、カテゴリーを構文的に定義している。構文的カテゴリーに変えて、意味カテゴリーを用いれば、本手法と類似の方法となる。しかし、［Bahl］では、標本の度数によらず、カテゴリーを定義しているので、度数の大きなデータと、小さなデータが共存するカテゴリーでは、整合度推定値の精度が悪くなると思われる。［Bahl］は推定整合度として、カテゴリーを何種類か定義して、カテゴリー毎に推定した整合度の１次結合を用いている。結合係数は削除補完法によって求めている。本報告で述べる方法は、この結合係数が０か１に制限されている。本手法でも削除補完法で連続値を取る係数を定めることは可能であろうが、推定すべきパラメータの数が多く成りすぎて推定精度に問題が生じると思われる。

６．今後の課題

今後の課題として、まず、学習用例のデータ量の問題がある。データベースからの統計に基づく知識ベースを利用する手法では、データ量と被覆率の関係が一般的に問題になる。辞書の語数と被覆率の関係は有名であるが、係り受けの場合もデータ量が少ないと、被覆率も小さいし、本手法のような方法を用いても、結果の精度は低くなることが予想される。データ量と精度との関係を明らかにする

ことが課題である。

今回の実験では、ＡＴＲ対話データベースに加えて、新聞データベースを利用した。新聞データベースには、今回利用したものも含めて、

・「が」格関係　１９万３千件　朝日新聞１年＋８４日分
・「を」格関係　１０万１千件　朝日新聞１年＋８４日分
・「に」格関係　　５万７千件　朝日新聞８４日分
・「で」格関係　１０万７千件　朝日新聞１年＋８４日分
・「の」格関係　２０万１千件　朝日新聞８４日分

がＡＴＲで利用可能であり（atr-ln:/data6/KYOUKI）、また、

・「は」格関係

も入手予定である。しかし、新聞データベースには、現在のところ係り受け関係として格関係のみが収集されており、

・述語による名詞の修飾関係
・述語による述語の修飾関係

などは収集されていない。

他の公開利用の可能性がある日本語の係り受けデータとしては、ＥＤＲの共起データが考えられる［Nakao］。このデータは１，０００万文からの抽出を目標に構文解析を半自動で行って収集している。

係り受けデータに意味カテゴリーを付与するためには、語を意味分類した辞書が必要である。日本語に対するこの種の既存の辞書として、一般的に利用可能で、かつ、ある程度の規模があるものは、

類語国語辞典［浜西］
分類語彙表［国研］
ＥＤＲ概念辞書［ＥＤＲ］

である。意味分類体系と整合度推定の精度との関係は明かでなく、今後の課題である。また、最適あるいはより良い意味分類体系を構築することも課題である。なお、分類語彙表は増補を行っており、作業途中のものではあるがＡＴＲで使用可能である。

係り受けデータベースが構築できたとしても、整合度計算上の課題がある。初めに、最適な平滑化パラメータの値を設定する必要がある。本報告で述べた平滑化パラメータの決定は経験的であり、最適な平滑化パラメータを決定する方法を得ることは、今後の課題である。

次の問題点として、機能語による係り受け関係の変化がある［木村］。例えば、

「花瓶　が　置いてある」

という文は可能である。しかし、

＊「花瓶　が　置く」

は不可能である。これは、機能語「てある」による格関係の変化である。そこで、前者の係り受け整合度として、

S（花瓶、が、置く）

を取るのは妥当ではなく、

S（花瓶、を、置く）

を取らなければならない。この様な機能語による格関係の変化も考慮して、整合度計算をする必要がある。

次に、述語による連体修飾に関する整合度を求めようとしたとき、係り受け関係が陽に表現されていない。例えば、

「食べた　リンゴは　おいしかった」

という文で、「食べた」が「リンゴは」に係る整合度を計算するときに、係り受け関係が「述語による名詞の修飾」としか規定できず、制約が少なすぎると思われる。「食べた」の結合価である「名詞$_1$が、名詞$_2$を、食べる」を利用して、

S（リンゴ、が、食べる）
S（リンゴ、を、食べる）

を求め、後者を採用するような機構が必要であると思われる。しかし、主名詞が述語を含む節の格要素に成っていない、疑似関係節の場合は、この方法を取ることもできず、問題が残る。

次に、述語が述語を修飾する

「外国からの参加者もかなり居られますので、
使用言語は英語と成っております」

の場合、

S（居る、ので、成る）

を用いるだけでは、これも、制約が少なすぎるように思える。このような、場合は、語と語の係り受け関係ではなく、節と節の係り受け関係

S（［参加者も、かなり、居る］、ので、［使用言語は、英語と、成る］）

を利用する方が精度が良いと思われる。しかし、節と節の場合、用例データの被覆率が一層低くなると予想され、より強力な対策が必要になるであろう。例えば、［工藤］の方法などが考えられる。

その他の問題として、語と語が並列する場合：

「ご住所と　お名前を　お願いします。」
└ー ー ー ー ー ー┘

副詞が内容語でなく機能語に係る場合：

「決して　ご心配　なさらないように」
└ー ー ー ー ー ー ー ー ー ー ー ー┘

などが挙げられるが、これらの問題は、特別な処理で対応できると思われる。

7．おわりに

係り受け整合度を用いて、構文的曖昧性を解消する手法において、意味カテゴリーと標本における度数を利用して、整合度を平滑化する手法を提案し、実験によって、有効性を実証した。また、実際に整合度を利用する場合の問題点について述べた。

今後は、構成文法による日本語の構文解析器に本手法を適用して、構文的曖昧性の解消がどの程度の精度で可能であるかを実験する予定である。

本研究の機会を与えられた、槇松明社長、森元逞室長に感謝する。また、プログラムの開発と実験の実施に協力された、松田伸洋氏（漢字情報サービス）［ＫＩＳ］、武谷典子さん（日進ソフト）［日進］に感謝する。また、ＡＴＲ対話データベースを作成された、小倉健太郎　元主任研究員（現ＮＴＴ）、橋本一男　元研究員（現オージス）はじめ関係各位、係り受け情報を作成された、井ノ上直己　元研究員（現ＫＤＤ）、新聞データを作成された、田中康仁教授（愛知淑徳大学）に敬意を表する。

参考文献

［Atkinson］ Martin Atkinson et al.: Foundations of General Linguistics, p.33, Geogre Allen & Unwin, London, 1982.

［Bahl］ Lalit Bahl et al.: A Maximum Likelihood Approach to Continuous Speech Recognition, IEEE Transaction, Vol.PAMI1-5, No.2, March, 1983.

［Church］ Kenneth Church and William Gale: A Comparison of the Enhanced Good-Turing and Deleted Estimation Methods for Estimating Probabilities of English Bigrams, Computer Speech and Language, No.5, pp.19-54, 1991.

［Nakao］ Yoshio Nakao: How to Extract Dictionary Data from the EDR Corpus, Proceedings of International Workshop on Electronic Dictionaries, Oiso, Nov. 8-9, 1990.

［ＥＤＲ］日本電子化辞書研究所：概念辞書、ＥＤＲテクニカルレポート、ＴＲ－０２０、１９９０年４月。

［ＫＩＳ］漢字情報サービス：係り受け整合度の計算法評価作業報告書、納品書類、１９９０年９月。

［井ノ上］井ノ上直己、竹澤寿幸：階層的クラスタリング手法の訳語選択への応用、ＴＲ－Ｉ－０２０４、１９９１年３月。

［井ノ上２］井ノ上直己、森元逞：クラスタリング手法と既存のシソーラスとの組合せ手法、４２回情全大、７Ｅ－６、１９９１。

［江原］江原暉将ほか：ＡＴＲ対話データベースの内容、ＴＲ－Ｉ－０１８６、１９９０年１０月。

［木村］木村睦子、空閑茂起：態による格助詞変換、計量国語学、Ｖｏｌ．１５，Ｎｏ．２，１９８５。

［工藤］工藤育男：文と文の結束性を捉えるための知識、ＴＲ－Ｉ－０１３４、１９９０年２月。

［国研］国立国語研究所：「分類語彙表」増補モニター通信、１９８９。

［児玉］児玉徳美：依存文法の研究、研究社出版、１９８７。

［田中（英）］田中英輝、江原暉将：日本語「が」格の係り受け整合度の一計算

法、情処研資、８９－ＮＬ－７０、１９８９年１月。
[田中（康）] 田中康仁：語と語の関係解析用資料　「が」を中心とした、１９８９。
[日進] 日進ソフトウェア：係り受け整合度計算プログラムの作成報告書、納品書類、１９９１年２月。
[浜西] 浜西正人、大野普：類語国語辞典、角川書店、１９８５。
[水谷] 水谷静夫：朝倉日本語新講座２　語彙、４７頁、朝倉書店、１９８３。

付録1

係り受けの頻度の問題に入る前に、まず、標本調査法の一般論を述べておく。有限母集団をＤＰとし、その大きさをＮＰとする。母集団は次の様にして確率空間である。

$$SP = (DP,\ \beta(DP),\ P) \qquad (A1.\ 1)$$

ここで、β（ＤＰ）はＤＰの部分集合の全体、確率Ｐ：β（ＤＰ）－＞Ｒ　（Ｒは実数の集合）はβ（ＤＰ）の要素Ｄに対して

$$P(D) = \frac{|D|}{NP} \qquad (A1.\ 2)$$

から定められる写像である。ここで、｜Ｄ｜はＤの要素数である。この母集団から、大きさＮの無作為標本を非重複抽出するとする。この様な標本の全体は、標本空間という（ＳＰとは異なる）確率空間

$$SS = (DD,\ \beta(DD),\ PD) \qquad (A1.\ 3)$$

になる。ここで、ＤＤは全ての標本の全体（その個数は${}_{NP}C_{N}$である）、β（ＤＤ）はＤＤの部分集合の全体、確率ＰＤ：β（ＤＤ）－＞Ｒ　はβ（ＤＤ）の要素ＤＫに対して

$$PD(DK) = \frac{|DK|}{{}_{NP}C_{N}} \qquad (A1.\ 4)$$

である。

さて、確率空間ＳＰ上の実数値をとる確率変数をＸとし、Ｘの平均値をＥ［Ｘ］、分散をＶ［Ｘ］、標準偏差をσ［Ｘ］と書く。ここで、Ｘから確率空間ＳＳ上の確率変数ＸＤを次のようにして定義する。

$$XD: DD \rightarrow R;\ XD(DK) = E[X \mid DK] \qquad (A1.\ 5)$$

ここで、Ｅ［Ｘ｜ＤＫ］はＸのＤＫ上での条件付平均値である。ＸＤとＸとは別の確率空間上の確率変数であるので、ＸＤの平均値を取る作用素をＥと区別してＥＳと書くと

$$ES[XD] = E[X] \qquad (A1.\ 6)$$

が成立する。また、ＸＤの分散をＶＳ［ＸＤ］、標準偏差をσＳ［ＸＤ］と書くと、

$$VS[XD] = \frac{NP-N}{NP-1} * \frac{V[X]}{N} \qquad (A1.\ 7)$$

$$\sigma S[XD] = (VS[XD])^{1/2} \qquad (A1.8)$$

となる。NP>>Nと仮定すると、

$$VS[XD] = \frac{V[X]}{N} \qquad (A1.9)$$

$$\sigma S[XD] = \frac{\sigma[X]}{N^{1/2}} \qquad (A1.10)$$

と近似できる。今後、この近似式を用いる。

さて、次に我々の場合を一般論に沿って議論する。まず、母集団ＤＰとしては、解析対象となる係り受けの全体とする。ここで、関係ｋと係り先ｖは固定している場合を考える。この場合、ＤＰは係り元語の単位語としての集合である。これを

$$DP = \{d_1, d_2, \cdots, d_{NP}\} \qquad (A1.11)$$

とする。ＤＰ中の異なり語の全体を

$$E = \{n_1, n_2, \cdots, n_M\} \qquad (A1.12)$$

とする。まず、カテゴライズを行わない場合について考察する。Ｅの各要素n_iに対して、ＳＰ上の確率変数Xn_iを次の様に定義する。

$$Xn_i(d_j) = \begin{cases} 1, & d_j = n_i \\ 0, & \text{else} \end{cases} \qquad (A1.13)$$

すると、n_iのＤＰ中での度数をfp_iとするとき

$$E[Xn_i] = \frac{fp_i}{NP} = sp_i \qquad (A1.14)$$

となる。つまり、係り元n_iが関係ｋで係り先ｖに係る係り受けの整合度sp_iは確率変数Xn_iの平均値であることが分かる。また、分散$V[Xn_i]$は

$$V[Xn_i] = E[Xn_i] * (1 - E[Xn_i]) \qquad (A1.15)$$

となる［水谷］。さらに、

$$1 >> E[Xn_i] \qquad (A1.16)$$

と仮定すると

$$V[Xn_i] = E[Xn_i] \quad (A1.17)$$

と近似できる。今後はこの近似式を用いる。

次に、大きさNの無作為標本の1つをDとする。上記Xn_iから一般論の（A1.5）式によって定義される

$$Xn_iD(D) = E[Xn_i \mid D] \quad (A1.18)$$

は、次の通りである。標本D中の係り元n_iの度数をf_iとすると、条件付平均値の定義から

$$Xn_iD(D) = -\frac{f_i}{N}- \quad (A1.19)$$

となる。（2.2）式はこれに等しい。つまり、標本Dから得られる係り元n_iの整合度sp_iの推定値s_iは、確率変数Xn_iDのDでの値に等しい。そこで、整合度の誤差推定値e_iを

$$e_i = ES[(Xn_iD - sp_i)^2]^{1/2}$$

と定義するのが妥当であり、これはまた（A1.6）を用いて

$$\begin{aligned} e_i &= ES[(Xn_iD - E[Xn_i])^2]^{1/2} \\ &= ES[(Xn_iD - ES[Xn_iD])^2]^{1/2} \\ &= VS[Xn_iD]^{1/2} \\ &= \sigma S[Xn_iD] \quad (A1.20) \end{aligned}$$

となる。この式の最右辺は（A1.10）と（A1.17）から

$$e_i = \frac{\sigma[Xn_i]}{N^{1/2}} \quad (A1.21)$$

$$= \frac{E[Xn_i]^{1/2}}{N^{1/2}}- \quad (A1.22)$$

となる。さらに、Xn_iの平均値$E[Xn_i]$を標本Dでの条件付平均値$E[Xn_i \mid D]$で近似すると、（A1.19）から

$$e_i = \frac{f_i^{1/2}}{N}- \quad (A1.23)$$

となる。これで、（2.5）式が求まった。

次に、（２．７）式のカテゴリーＦを用いる場合を考察する。係り元n_iを含むカテゴリーをc_kとするとき、ＳＰ上の確率変数Yn_iを

$$Yn_i = \frac{1}{|c_k|} \Sigma [n_j \in c_k] \, Xn_j \qquad (A1.24)$$

と定義する。つまり、Xn_jのc_k上での算術平均である。つぎに、$Xn_i D$と同様に$Yn_i D$をＳＳ上の確率変数として定義する。つまり

$$Yn_i D(D) = E[Yn_i \mid D] \qquad (A1.25)$$

である。カテゴライズしたときの推定誤差を

$$e_i' = ES[(Yn_i D - sp_i)^2]^{1/2} \qquad (A1.26)$$

と定義する。これは

$$\begin{aligned} e_i' &= ES[\{(Yn_i D - ES[Yn_i D]) \\ &\quad + (ES[Yn_i D] - sp_i)\}^2]^{1/2} \\ &= \{ES[(Yn_i D - ES[Yn_i D])^2] \\ &\quad + 2ES[Yn_i D - ES[Yn_i D]] * (ES[Yn_i D] - sp_i) \\ &\quad + (ES[Yn_i D] - sp_i)^2\}^{1/2} \\ &= \{VS[Yn_i D] \\ &\quad + (ES[Yn_i D] - sp_i)^2\}^{1/2} \qquad (A1.27) \end{aligned}$$

となる。（Ａ１．９）より、

$$VS[Yn_i D] = \frac{V[Yn_i]}{N}$$

$$= \frac{1}{N} E\left[\left\{\frac{1}{|c_k|} \Sigma [n_j \in c_k] \, Xn_j - \frac{1}{|c_k|} \Sigma [n_j \in c_k] \, E[Xn_j]\right\}^2\right]$$

$$= \frac{1}{N * |c_k|^2} \Sigma [n_j \in c_k] \Sigma [n_l \in c_k] \{E[Xn_j Xn_l] - E[Xn_j] E[Xn_l]\} \qquad (A1.28)$$

ところが、$j \neq 1$ならば、（A1．13）から

$$X n_j X n_l = 0 \qquad (A1．29)$$

であるから、

$$VS[Y n_i D] = \frac{1}{N * |c_k|^2} \Sigma [n_j \in c_k] V[X n_j] -$$

$$\frac{1}{N * |c_k|^2} \Sigma [n_j \in c_k] \Sigma [n_l \in c_k,\ j \neq 1] E[X n_j] E[X n_l]$$

（A1．30）

となる。この右辺第2項は2次の微小量であるから省略する。次に、（A1．17）を用いて、さらに、母集団での平均を標本での平均で近似すると、

$$VS[Y n_i D] = \frac{1}{N * |c_k|^2} \Sigma [n_j \in c_k] V[X n_j]$$

$$= \frac{1}{N^2 * |c_k|^2} \Sigma [n_j \in c_k] f_j \qquad (A1．31)$$

となる。一方、（A1．27）の最右辺{}内の第2項は、（A1．6）などより、

$$(ES[Y n_i D] - sp_i)^2$$

$$= \{\frac{1}{|C_k|} \Sigma [n_j \in c_k] E[X n_j] - E[X n_i]\}^2$$

$$= \frac{1}{N^2} \{\frac{1}{|C_k|} \Sigma [n_j \in c_k] f_j - f_i\}^2 \qquad (A1．32)$$

である。よって、（A1．27）は

$$e_i' = \{\frac{1}{N^2 * |c_k|^2} \Sigma [n_j \in c_k] f_j +$$

$$\frac{1}{N^2} (\frac{1}{|c_k|} \Sigma [n_j \in c_k] f_j - f_i)^2\}^{1/2}$$

（A1．33）

となる。これで（２．９）式が求まった。カテゴリー内の誤差の２乗和を

$$ec_k^2 = \Sigma\,[n_i \in c_k]\; e_i^2 \qquad (A1.34)$$

$$ec_k'^2 = \Sigma\,[n_i \in c_k]\; e_i'^2 \qquad (A1.35)$$

と置くと、（Ａ１．２３）（Ａ１．３３）より、

$$ec_k^2 = \Sigma\,[n_i \in c_k]\; \frac{f_i}{N^2} \qquad (A1.36)$$

$$ec_k'^2 = \Sigma\,[n_i \in c_k]\; \{\frac{1}{N^2 * |c_k|^2}\Sigma\,[n_j \in c_k]\; f_j + \frac{1}{N^2}\,(\frac{1}{|c_k|}\Sigma\,[n_j \in c_k]\; f_j - f_i)^2\}$$

$$= \frac{1}{|c_k|}\Sigma\,[n_i \in c_k]\; \frac{f_i}{N^2} + \Sigma\,[n_i \in c_k]\; \frac{1}{N^2}\,(\frac{1}{|c_k|}\Sigma\,[n_j \in c_k]\; f_j - f_i)^2 \qquad (A1.37)$$

となる。ここで

$$E[c_k] = \frac{1}{|c_k|}\Sigma\,[n_i \in c_k]\; \frac{f_i}{N^2}$$

$$V[c_k] = \frac{1}{|c_k|}\Sigma\,[n_i \in c_k]\; \frac{1}{N^2}\,(\frac{1}{|c_k|}\Sigma\,[n_j \in c_k]\; f_j - f_i)^2$$

と置くと、

$$ec_k'^2 - ec_k^2 = E[c_k] + |c_k| * V[c_k] - |c_k| * E[c_k] \qquad (A1.38)$$

となる。カテゴライズによって誤差が減少する条件は（Ａ１．３８）≦０であり、これは

$$\frac{(|c_k|-1)}{|c_k|} \geqq \frac{V[c_k]}{E[c_k]} \qquad (A1.39)$$

である。これで、（２．１５）式が求まった。

付録２　係り受け元データ　（文例ファイル）

データ形式

出典｜関係名｜通番｜度数｜係り元語形／関係語形／係り先語形｜係り元標準形／関係語標準形／係り先標準形｜係り元品詞コード／関係語品詞コード／係り先品詞コード

出典　１：電話　２：キー　３：新聞
品詞コードはＡＤＤのもの［江原］を利用

データ例

電話データ例　（bunrei.tel）

1|が|1|1|登録 用紙/が/ある ん です けれども|登録 用紙/が/有る ん です けれども|04 04/15/32 34 12 14
1|が|1|1|所/が/ある ん です が|所/が/有る ん です が|04/15/32 34 12 14
1|が|1|1|者/が/いる ん です けれども|者/が/居る ん です けれども|04/15/32 34 12 14
1|が|1|1|支払い/が/し やすい と|支払/が/為る 易い と|04/15/32 17 15
1|が|1|1|こと/が/ある ん です が|事/が/有る ん です が|04/15/32 34 12 14
1|が|1|1|エージェント みたい の/が/ある ん でしょ う か|エージェント みたいだ の/が/有る ん です う か|04 12 34/15/32 34 12 12 16
1|が|1|1|方/が/速い でしょ う か|方/が/速い です う か|04/15/01 12 12 16
1|が|1|1|方/が/速い と|方/が/速い と|04/15/01 15
1|が|1|1|リムジン・バス/が/出 て おり まし て|リムジン・バス/が/出る て おる ます て|04/15/32 14 19 12 14
1|が|1|1|こと/が/載っ て た ん です けれども|事/が/乗る て た ん です けれども|04/15/32 14 12 34 12 14

キーデータ例　（bunrei.key）

2|が|1|1|点/が/ございまし たら|点/が/御座います たら|04/15/32 14
2|が|1|1|内容/が/明記 さ れ て おり ます が|内容/が/明記 する れる て おる ます が|04/15/05 19 12 14 19 12 14
2|が|1|1|自信/が/無い の です が|自信/が/無い の です が|04/15/01 34 12 14
2|が|1|1|同時 通訳/が/付き ます ので|同時 通訳/が/付く ます ので|04 04/15/32 12 14
2|が|1|1|こと/が/出来 ます から|事/が/出来る ます から|04/15/32 12 14
2|が|1|1|同時 通訳/が/付き ます ので|同時 通訳/が/付く ます ので|04 04/15/32 12 14
2|が|1|1|分科 会/が/ございます|分科 会/が/御座います|04 04/15/32
2|が|1|1|御 婦人/が/お 見え に なる と|御 婦人/が/御 見える に なる と|18 04/15/18 32 15 19 15

2|が|1|1|必要/が/あり ます が|必要/が/有る ます が|04/15/32 12 14
2|が|1|1|研究 生/が/言っ て おり ます|研究 生/が/言う て おる ます|04 17
/15/32 14 19 12

新聞データ例　(bunrei.ga1,bunrei.ga2)

3|が|9006|3|運動/が/盛んだ|運動/が/盛んだ|05/15/31
3|が|9007|10|運動/が/盛んである|運動/が/盛んだ|05/15/31
3|が|9008|6|運動/が/盛んに なる|運動/が/盛んだ なる|05/15/31 32
3|が|9009|1|運動/が/相次ぐ|運動/が/相次ぐ|05/15/32
3|が|9010|2|運動/が/続く|運動/が/続く|05/15/32
3|が|9011|1|運動/が/大きく なる|運動/が/大きい なる|05/15/01 19
3|が|9012|1|運動/が/大 切|運動/が/大 切|05/15/18 04
3|が|9013|1|運動/が/中心 である|運動/が/中心 である|05/15/04 12
3|が|9014|1|運動/が/注目 を 集める|運動/が/注目 を 集める|05/15/05 15 32
3|が|9015|1|運動/が/定着 する|運動/が/定着 する|05/15/05 19
3|が|9016|1|運動/が/凍結 する|運動/が/凍結 する|05/15/05 19
3|が|9017|1|運動/が/動き 始める|運動/が/動く 始める|05/15/32 32

付録３　意味カテゴリー付与済みデータ

データ形式

出典｜関係名｜通番｜度数｜係り元標準形／関係語標準形／係り先標準形｜係り元品詞コード／関係語品詞コード／係り先品詞コード｜係り元分類語彙表番号／係り先分類語彙表番号｜係り元類語番号／係り先類語番号

データ例

電話データ例　（tel.jisho2）

```
1|が|0000347|0000001|インタビュー/が/独占 インタビュー だ ある と|04/15/04 04 12 19 15|13520. 1. 50/13701. 2. 20|781a/379 |1/10000|1/10000
1|が|0000348|0000001|インタビュー/が/入る て おる ます が|04/15/32 14 19 12 14|13520. 1. 50/21530. 6. 20,21530. 7.220|781a/230b,233b,313b,370d,712 ,745a|1/10000|1/10000
1|が|0000366|0000001|エキスパート/が/発表 いたす ます て|04/15/05 19 12 14|12340. 4.380/13140. 3. 10|582 /752 |1/1000|1/1000
1|が|0000386|0000001|カード 自身/が です ね/使う と|04 04/15 12 16/32 15|12010. 1. 30/23852. 2. 10|505a/363c,381 |01/10|01/10
1|が|0000403|0000001|カメラマン/が/プール する て おく ようだ|04/15/05 19 14 19 12|12410.12. 80/14700.13.150|574a/035 ,379b,723a,724 |1/10000|1/10000
1|が|0000413|0000001|キャンプ 場/が/有る ます て|04 04/15/32 12 14|11600. 1. 50,11600.99. 20,11700. 4. 90,12620. 1. 20,12620. 1. 10/21200. 1. 20|702 ,724a,829a,883a,883b/125a,379 |01/100|01/100
```

キーデータ例　（key.jisho2）

```
2|が|0000314|0000001|コピー 機/が/利用 できる か|04 17/15/05 19 16|13151. 1.130/13852. 3.160|336a/381 |10/100|10/100
2|が|0000325|0000001|コンサート/が/有る と|04/15/32 15|13510. 5.110/21200. 1. 20|711b/125a,379 |1/10|1/10
2|が|0000354|0000001|サービス 料/が/付く ます|04 04/15/32 12|13740. 3. 20,14100. 2. 10/21560. 2. 10|748a/224 ,263a,291a,783a|01/10|01/10
2|が|0000355|0000001|サービス/が/受ける られる ので|04/15/32 12 14|13541. 1. 50,13541. 1. 40/21503. 2.130,21515. 5. 40,23300. 1. 30,23770. 7. 10|743c,787 ,795 /295a,370d|1/100|1/100
2|が|0000356|0000001|サービス/が/受ける られる ます ようだ|04/15/32 12 12 12|13541. 1. 50,13541. 1. 40/21503. 2.130,21515. 5. 40,23300. 1. 30,23770. 7. 10|743c,787 ,795 /295a,370d|1/1000|1/1000
2|が|0000377|0000001|シングル/が/成る ます|04/15/32 12|11981. 6. 60/21220. 1. 40,21500. 1.130,25810.11.100|128a/250 ,260 ,273a|1/10|1/10
2|が|0000388|0000001|スケジュール/が/厳しい て|04/15/01 14|11680. 2. 30/33600. 5. 40,33680. 5. 10|281a,824b/163 ,655b|1/10|1/10
```

新聞データ例　（news1.jisho2, news2.jisho2）

```
3|が|0003060|0000001|アップ/が/目的|99/15/04|11540. 1. 90/11113. 2. 10,1
1731. 1. 50|616a/818 |1/1|1/1
3|が|0003061|0000001|アップ/が/目立つ|99/15/32|11540. 1. 90/25010. 3. 30
|616a/299 |1/1|1/1
3|が|0003064|0000001|アトキンズ 氏/が/辞める|99 04/15/32|12020. 6. 70,12
100. 5. 40,12100. 5. 10/23320. 5. 10,23630. 4.120|503 ,509 ,822b/776a|01
/1|01/1
3|が|0003065|0000001|アドバイス/が/効く|99/15/32|13640. 3.110/21310. 1.1
00|455 /294d|1/1|1/1
3|が|0003068|0000001|アナウンサー/が/か も する 出す|99/15/13 36 19 32|1
2410. 7. 80/21210. 1.400,21220. 2.130,21502. 1.170,21530. 1. 20,23770. 3
.110|573 /230a,233a,272 ,282 ,314a,333c,335 ,386b|1/0001|1/0001
3|が|0003069|0000001|アナウンサー/が/質問 する|99/15/05 19|12410. 7. 80/
23130.11. 30|573 /346a|1/11|1/10
3|が|0003070|0000001|アナウンス/が/響く|99/15/32|13123. 5. 30/21110. 4.
20,25030. 1. 90|750a/096b,295 ,754c|1/1|1/1
3|が|0003075|0000001|アパート/が/焼ける|99/15/32|12660. 2. 50,14400. 1.2
30/25161. 1. 10|728 /022c,092 ,256 ,355a|1/1|1/1
3|が|0003078|0000001|アフリカ 人/が/利用 できる|99 04/15/05 32|11960. 3.
 70,11960. 3. 40,12010. 2. 60,12020. 5. 10,12020. 1. 20,12020. 5. 20/138
52. 3.160|505b,507 ,508 ,537 ,580 ,660c,829d/381 |01/10|01/10
3|が|0003079|0000001|アブルッチ 公/が/指揮 する|99 04/15/05 19|12020. 6.
150,12500. 1. 30,12500. 1. 10/13600. 2. 10|506a,509 ,733a/452 |01/10|01/
10
3|が|0003082|0000001|アボット 社/が/開発 する|99 04/15/05 19|12630. 2. 6
0,12640. 1. 90,12800. 2. 80/13822. 3. 10|713b,727a/395a,455b|01/10|01/10
3|が|0003084|0000001|アマ/が/勝つ|99/15/32|12340.13. 40/21990. 6. 10,235
30.11. 10|582a/299 ,463a|1/1|1/1
3|が|0003085|0000001|アマチュア/が/使用 する|99/15/05 19|12340.13. 30/23
852. 2. 20|582a/363c,381 |1/11|1/10
3|が|0003086|0000001|アマ 気分/が/抜ける ない|99 04/15/32 12|11302. 1. 8
0,13004. 5. 40/21250. 2.120,21531. 5. 20|168 ,690 /213a,231a,260a,264b,3
13a|01/10|01/10
```

付録４　係り受け語形ファィル

データ形式

係り元語形｜係り先語形｜係り元分類番号｜係り先分類番号｜データ１度数｜・・・｜データ７度数

データ例

```
ディスカッション|為す|345 |361 |   2.00|   0.00|   0.00|   0.00|   0.00|   0.00|   0.00
ディスカッション|出来る|345 |071a|   0.14|   0.00|   0.00|   0.00|   0.00|   0.00|   0.00
ディスカッション|出来る|345 |260 |   0.14|   0.00|   0.00|   0.00|   0.00|   0.00|   0.00
ディスカッション|出来る|345 |272 |   0.14|   0.00|   0.00|   0.00|   0.00|   0.00|   0.00
ディスカッション|出来る|345 |273a|   0.14|   0.00|   0.00|   0.00|   0.00|   0.00|   0.00
ディスカッション|出来る|345 |361a|   0.14|   0.00|   0.00|   0.00|   0.00|   0.00|   0.00
ディスカッション|出来る|345 |391 |   0.14|   0.00|   0.00|   0.00|   0.00|   0.00|   0.00
ディスカッション|出来る|345 |774 |   0.14|   0.00|   0.00|   0.00|   0.00|   0.00|   0.00
ディスク|高熱|968e|074b|   0.00|   0.00|   0.50|   0.00|   0.00|   0.00|   0.00
ディスク|高熱|968e|093 |   0.00|   0.00|   0.50|   0.00|   0.00|   0.00|   0.00
ディナー|終わる|354 |282a|   0.00|   0.00|   0.00|   0.00|   0.00|   0.00|   1.00
ディレクター|出る|552 |230a|   0.00|   0.00|   0.13|   0.00|   0.00|   0.00|
 0.00
ディレクター|出る|552 |233a|   0.00|   0.00|   0.13|   0.00|   0.00|   0.00|
 0.00
ディレクター|出る|552 |272 |   0.00|   0.00|   0.13|   0.00|   0.00|   0.00|
 0.00
ディレクター|出る|552 |312a|   0.00|   0.00|   0.13|   0.00|   0.00|   0.00|
 0.00
ディレクター|出る|552 |313a|   0.00|   0.00|   0.13|   0.00|   0.00|   0.00|
 0.00
ディレクター|出る|552 |314a|   0.00|   0.00|   0.13|   0.00|   0.00|   0.00|
```

0.00
ディレクター|出る|552 |745b|　0.00|　0.00|　0.13|　0.00|　0.00|　0
.00|
0.00
ディレクター|出る|552 |745b|　0.00|　0.00|　0.13|　0.00|　0.00|　0
.00|
0.00
ディレクター|出る|552 |785a|　0.00|　0.00|　0.13|　0.00|　0.00|　0
.00|
0.00
ディレクター|製作|552 |391 |　0.00|　0.00|　1.00|　0.00|　0.00|　0
.00|
0.00
デザイナー|つ|574a|829b|　0.00|　0.00|　0.00|　1.00|　0.00|　0.00|
1.00
デザイナー|開発|574a|395a|　0.00|　0.00|　0.50|　0.00|　0.00|　0.0
0|　0.00
デザイナー|開発|574a|455b|　0.00|　0.00|　0.50|　0.00|　0.00|　0.0
0|　0.00
デザイナー|結集|574a|227a|　0.00|　0.00|　0.00|　0.00|　1.00|　0.0
0|　0.00
デザイナー|参加|574a|712 |　0.00|　0.00|　0.00|　0.00|　0.00|　1.0
0|　0.00
デザイナー|制作|574a|808a|　0.00|　0.00|　0.00|　1.00|　0.00|　0.0
0|　1.00
デザイナー|得意|574a|535c|　0.00|　0.00|　0.00|　0.00|　0.00|　0.0
0|　0.20

付録５　有効度の計算に用いたファイル

学習用データ　ＤＬ　（sanka6.DL）

データ形式　語形｜分類番号｜度数

雪|024 |　1.00
湾|034c|　2.00
私鉄|047g|　1.00
馬|061b|　1.00
生|070 |　0.13
生|071b|　0.13
場所|100 |　0.33
余|101d|　0.66
社中|103a|　0.50
前後|104 |　0.25
近い|108b|　0.50
北|109d|　1.00
大|117a|　0.33
十|120b|　1.00
千|120b|　0.50
八|120b|　1.00
百|120b|　1.50
六|120b|　1.00
多数|121a|　5.00
半数|121a|　1.00
兵員|121c|　1.00
二十|124b|　2.00
大|126a|　0.33
千|126c|　0.50
百|126c|　1.50
以上|126g|　3.00
余|126g|　0.66
全体|127 |　0.50
全部|127 |　1.00
一部|127c|　1.00
大半|127d|　2.00
大部分|127d|　2.00
生|139a|　0.13
前後|152 |　0.25
前後|156 |　0.25
全体|158d|　0.50
近い|158f|　0.50
相|160 |　1.50
生|175c|　0.13
曲|179a|　0.33
派|184b|　0.50
近い|189 |　0.50
以上|191a|　3.00

クラス|192 | 1.00
生|194b| 0.13
大|195 | 0.33
前後|198a| 0.25
余|198a| 0.66
層|228a| 1.00
余る|261 | 3.00
余る|265a| 3.00
直す|277b| 0.25
直す|277c| 0.25
関係|290 | 0.20
関係|295 | 0.20
代表|297b| 12.50
行|311a| 0.33
生|350 | 0.13
企業|369 | 12.00
工業|369a| 1.00
給油|371b| 1.00
支持|387b| 0.50
製鋼|391b| 1.00
直す|392 | 0.25
直す|417a| 0.25
行|421b| 0.33
研究|424 | 7.00
支持|445a| 0.50
共済|458a| 1.00
支援|458b| 1.00
攻撃|464a| 0.50
攻撃|473 | 0.50
諸氏|502b| 1.00
氏|503 | 0.99
両氏|503 | 1.00
自ら|505a| 0.33
自身|505a| 1.00
人|505b| 0.84
余人|505b| 3.00
個人|506b| 3.00
民間|506b| 0.50
者|507 | 0.50
人|507 | 0.84
万人|507a| 2.00
家|508 | 0.20
子|508 | 0.20
者|508 | 0.50
人|508 | 0.84
生|508 | 0.13
氏|509 | 0.99
生|509 | 0.13
女性|511b| 5.00
婦人|511b| 1.00

若者|514 | 1.00
子|514a| 0.20
患者|518 | 1.00
父母|522 | 1.00
母親|522b| 1.00
子|523 | 0.20
家|528a| 0.20
社中|530 | 0.50
党|530 | 0.50
教徒|530b| 1.00
委員|531 | 3.00
議員|531 | 7.00
社員|531 | 1.00
代議員|531 | 1.00
党員|531 | 3.00
全員|531a| 2.00
市民|536 | 4.00
住民|536 | 3.00
人|537 | 0.84
大衆|537 | 1.00
国民|538 | 1.00
市民|538 | 4.00
会頭|542 | 1.00
学長|542 | 1.00
議長|542 | 1.00
社長|542 | 2.00
党首|542a| 1.00
総裁|542b| 1.00
総長|542b| 1.00
外相|543 | 1.00
閣僚|543 | 2.00
首相|543 | 2.00
大臣|543 | 2.00
大統領|543 | 2.00
知事|543 | 1.00
三一|544 | 1.00
先生|545a| 0.50
労働者|549b| 1.00
スタッフ|551 | 1.00
首脳|551 | 3.00
指導者|551a| 1.00
担当者|552 | 2.00
役員|552 | 2.00
理事|552 | 1.00
助手|552a| 1.00
当事者|553 | 1.00
代表|553a| 12.50
奏者|556 | 1.00
記者|557 | 0.50
読者|557 | 1.00

ランナー|559 | 1.00
選手|559 | 18.00
メーカー|560 | 1.00
業者|560 | 3.00
実業家|560 | 1.00
大使|570 | 1.00
領事|570 | 1.00
弁護士|570b| 1.00
教師|572 | 2.00
先生|572 | 0.50
科学者|572a| 3.00
学生|572b| 2.00
生徒|572b| 2.00
アナウンサー|573 | 1.00
記者|573 | 0.50
音楽家|574b| 1.00
僧侶|576 | 1.00
保母|577a| 1.00
ＯＬ|578 | 1.00
収入役|578 | 1.00
人|580 | 0.84
博士|581 | 0.50
専門家|582 | 4.00
強豪|583 | 0.50
棋士|585a| 1.00
登山家|585a| 1.00
相|611b| 1.50
自ら|642 | 0.33
自ら|642a| 0.33
近い|651 | 0.50
人|660c| 0.84
強豪|670c| 0.50
精鋭|670c| 1.00
技術|677 | 1.00
博士|682b| 0.50
層|682d| 1.00
地域|700a| 3.00
場所|702 | 0.33
国|704a| 11.22
市|705a| 1.00
国|708 | 11.22
欧|709b| 1.00
ソ連|709c| 2.00
中国|709c| 5.00
米国|709c| 1.00
日本|709d| 6.00
群衆|710 | 1.00
教団|713 | 2.00
組合|713 | 1.00
団体|713 | 11.00

労組|713 | 1.00
会社|713b| 0.50
社|713b| 8.50
空軍|714 | 1.00
軍|714 | 4.00
ゲリラ|714a| 1.00
デモ隊|714a| 1.00
部隊|714a| 1.00
派|715 | 0.50
共産党|715a| 1.00
党|715a| 0.50
野党|715a| 1.00
クラス|715b| 1.00
チーム|715b| 6.00
業界|716 | 1.00
財界|716 | 1.00
家|717 | 0.20
家庭|717 | 1.00
世帯|717 | 1.00
農家|717 | 0.50
民間|718 | 0.50
我が国|719 | 1.00
国|719 | 11.22
大国|719 | 1.00
各国|719b| 3.00
諸国|719b| 4.00
塾|722 | 1.00
大学|722 | 5.00
病院|723 | 1.00
会社|724 | 0.50
場所|724a| 0.33
社|727a| 8.50
家|728 | 0.20
農家|728 | 0.50
飲食店|729a| 1.00
市|729b| 1.00
機関|733 | 1.50
政府|733a| 1.00
銀行|740 | 3.00
資本|747b| 1.00
子|748c| 0.20
スカウト|777a| 1.00
関係|780c| 0.20
関係|790 | 0.20
関係|790a| 0.20
作品|808 | 2.00
氏|822b| 0.99
子|827b| 0.20
隻|829c| 4.00
人|829d| 0.84

行|829e|　0.33
コーラス|873 |　1.00
曲|874 |　0.33
曲|878 |　0.33
家|940 |　0.20
縄|987a|　1.00
機関|991 |　1.50
艦船|998 |　2.00
艦艇|998b|　1.00
航空機|999 |　2.00

試験用データ　ＤＴ　（sanka6.DT）

データ形式　語形｜分類番号｜度数

湾|034c|　2.00
産|071a|　0.25
方方|101a|　1.00
何方|101b|　0.66
方|101d|　0.78
双方|101d|　2.00
前後|104 |　0.25
近い|108b|　0.25
方|109 |　0.78
何方|109e|　0.66
大|117a|　0.66
二|120b|　0.33
五つ|120b|　1.00
十|120b|　1.00
多数|121a|　1.00
半数|121a|　1.00
兵員|121c|　1.00
割|121i|　0.50
少年|124a|　0.50
二十|124b|　1.00
大|126a|　0.66
多い|126a|　1.00
以上|126g|　1.50
大勢|126h|　0.50
皆|127 |　0.50
全部|127 |　1.00
一部|127c|　2.00
一人|128a|　0.50
二人|128b|　1.00
組|128c|　0.33
前後|152 |　0.25
方|152 |　0.78
前後|156 |　0.25
二|156f|　0.33

近い|158f| 0.25
相|160 | 1.00
体|160a| 0.34
美|178a| 0.33
体|180 | 0.34
職種|181 | 1.00
近い|189 | 0.25
以上|191a| 1.50
級|192 | 0.50
大|195 | 0.66
前後|198a| 0.25
方|198a| 0.78
軍機|233d| 2.00
余る|261 | 0.50
余る|265a| 0.50
大勢|270 | 0.50
局|270a| 0.33
勢力|271a| 1.00
関係|290 | 0.20
関係|295 | 0.20
代表|297b| 6.50
行|311a| 0.33
居住|351 | 1.00
従業|364 | 1.00
企業|369 | 9.00
支持|387b| 1.50
産|390 | 0.25
行|421b| 0.33
研究|424 | 4.00
支持|445a| 1.50
教育|455a| 1.00
教授|455d| 2.00
兵|460a| 0.20
美|470a| 0.33
自分|501 | 1.00
私|501 | 2.00
私共|501b| 1.00
氏|503 | 1.32
何方|504 | 0.66
誰|504 | 2.00
自ら|505a| 0.33
自分|505a| 1.00
自身|505a| 2.00
人|505b| 0.98
余人|505b| 1.00
個人|506b| 1.00
民間|506b| 1.50
私|506b| 2.00
者|507 | 0.50
方|507 | 0.78

人|507 | 0.98
皆|507a| 0.50
方方|507a| 1.00
家|508 | 0.20
子|508 | 0.20
者|508 | 0.50
人|508 | 0.98
方|508a| 0.78
氏|509 | 1.32
男性|511a| 1.00
女性|511b| 1.00
少年|513 | 0.50
子|514a| 0.20
娘|514a| 0.50
家族|520 | 1.00
遺族|520 | 2.00
父母|522 | 2.00
母親|522b| 1.00
子|523 | 0.20
娘|523a| 0.50
家|528a| 0.20
組|530 | 0.33
党|530 | 0.50
信者|530b| 1.00
会員|531 | 1.00
職員|531 | 1.00
代議士|531 | 1.00
党員|531 | 3.00
メンバー|531 | 4.00
委員|531 | 4.00
議員|531 | 6.00
全員|531a| 5.00
友人|533 | 1.00
町民|536 | 1.00
市民|536 | 2.00
住民|536 | 2.00
人|537 | 0.98
大衆|537 | 1.00
国民|538 | 1.00
市民|538 | 2.00
外国人|538a| 1.00
人種|539a| 1.00
一人|540a| 0.50
廷臣|540c| 1.00
議長|542 | 1.00
局長|542 | 1.00
長官|542 | 1.00
会長|542 | 3.00
教頭|542a| 1.00
首相|543 | 1.00

大統領|543 | 1.00
閣僚|543 | 2.00
政治家|543 | 3.00
上司|544 | 1.00
先生|545a| 1.00
司令官|546a| 1.00
幕僚|546a| 1.00
兵|546b| 0.20
幹部|551 | 1.00
首脳|551 | 4.00
オルグ|551a| 1.00
顧問|552 | 1.00
担当者|552 | 1.00
役員|552 | 2.00
責任者|553 | 1.00
代理人|553a| 1.00
代表|553a| 6.50
被告|553b| 1.00
記者|557 | 0.50
選手|559 | 16.00
メーカー|560 | 1.00
業者|560 | 1.00
作業員|561 | 1.00
飛行士|564 | 2.00
公務員|570 | 1.00
兵|571 | 0.20
教師|572 | 1.00
先生|572 | 1.00
教授|572 | 2.00
科学者|572a| 2.00
学者|572a| 3.00
新入生|572b| 1.00
学生|572b| 4.00
記者|573 | 0.50
デザイナー|574a| 1.00
画家|574a| 1.00
歌手|574b| 2.00
女優|575 | 1.00
人|580 | 0.98
人材|580 | 1.00
権威|582 | 0.50
専門家|582 | 3.00
兵|583 | 0.20
ファン|585 | 0.50
体|600 | 0.34
相|611b| 1.00
体|620a| 0.34
権威|623 | 0.50
美|624 | 0.33
自ら|642 | 0.33

自ら|642a| 0.33
近い|651 | 0.25
人|660c| 0.98
国|704a| 10.23
市町村|705a| 1.00
国|708 | 10.23
ソ連|709c| 1.00
韓国|709c| 1.00
中国|709c| 5.00
日本|709d| 4.00
群|710 | 1.00
農協|713 | 1.00
団体|713 | 6.00
会社|713b| 1.00
社|713b| 10.50
兵|714 | 0.20
空軍|714 | 1.00
軍|714 | 1.00
全軍|714 | 1.00
部隊|714a| 2.00
党|715a| 0.50
組|715b| 0.33
級|715b| 0.50
チーム|715b| 8.00
業界|716 | 1.00
財界|716 | 1.00
家|717 | 0.20
民間|718 | 1.50
国|719 | 10.23
大学|722 | 1.00
会社|724 | 1.00
職場|724 | 1.00
社|727a| 10.50
家|728 | 0.20
店|728 | 0.50
店|729 | 0.50
局|733 | 0.33
機関|733 | 0.50
産|747 | 0.25
子|748c| 0.20
関係|780c| 0.20
関係|790 | 0.20
関係|790a| 0.20
主義|803 | 1.00
体|812 | 0.34
方|819 | 0.78
氏|822b| 1.32
二|823a| 0.33
子|827b| 0.20
割|828h| 0.50

局|829a|　0.33
叟|829c|　4.00
体|829d|　0.34
人|829d|　0.98
行|829e|　0.33
方|834e|　0.78
コーラス|873 |　1.00
産|900d|　0.25
家|940 |　0.20
ファン|959 |　0.50
機関|991 |　0.50
戦車|997a|　1.00
艦船|998 |　1.00
艦艇|998b|　1.00
戦闘機|999 |　1.00

意味カテゴリーデータ　j＝2にたいするD4　（sanka6-2.D4）

データ形式　分類番号（4桁）|カテゴリー要素数|平均度数|分散

024 |1|　1.00|　0.0000
034c|1|　2.00|　0.0000
047g|1|　1.00|　0.0000
061b|1|　1.00|　0.0000
070 |1|　0.13|　0.0000
071b|1|　0.13|　0.0000
100 |1|　0.33|　0.0000
101d|1|　0.66|　0.0000
103a|1|　0.50|　0.0000
104 |1|　0.25|　0.0000
108b|1|　0.50|　0.0000
109d|1|　1.00|　0.0000
117a|1|　0.33|　0.0000
120b|5|　1.00|　0.1000
121a|2|　3.00|　4.0000
121c|1|　1.00|　0.0000
124b|1|　2.00|　0.0000
126a|1|　0.33|　0.0000
126c|2|　1.00|　0.2500
126g|2|　1.83|　1.3689
127 |2|　0.75|　0.0625
127c|1|　1.00|　0.0000
127d|2|　2.00|　0.0000
139a|1|　0.13|　0.0000
152 |1|　0.25|　0.0000
156 |1|　0.25|　0.0000
158d|1|　0.50|　0.0000
158f|1|　0.50|　0.0000
160 |1|　1.50|　0.0000

175c	1	0.13	0.0000
179a	1	0.33	0.0000
184b	1	0.50	0.0000
189	1	0.50	0.0000
191a	1	3.00	0.0000
192	1	1.00	0.0000
194b	1	0.13	0.0000
195	1	0.33	0.0000
198a	2	0.46	0.0420
228a	1	1.00	0.0000
261	1	3.00	0.0000
265a	1	3.00	0.0000
277b	1	0.25	0.0000
277c	1	0.25	0.0000
290	1	0.20	0.0000
295	1	0.20	0.0000
297b	1	12.50	0.0000
311a	1	0.33	0.0000
350	1	0.13	0.0000
369	1	12.00	0.0000
369a	1	1.00	0.0000
371b	1	1.00	0.0000
387b	1	0.50	0.0000
391b	1	1.00	0.0000
392	1	0.25	0.0000
417a	1	0.25	0.0000
421b	1	0.33	0.0000
424	1	7.00	0.0000
445a	1	0.50	0.0000
458a	1	1.00	0.0000
458b	1	1.00	0.0000
464a	1	0.50	0.0000
473	1	0.50	0.0000
502b	1	1.00	0.0000
503	2	1.00	0.0000
505a	2	0.67	0.1122
505b	2	1.92	1.1664
506b	2	1.75	1.5625
507	2	0.67	0.0289
507a	1	2.00	0.0000
508	5	0.37	0.0706
509	2	0.56	0.1849
511b	2	3.00	4.0000
514	1	1.00	0.0000
514a	1	0.20	0.0000
518	1	1.00	0.0000
522	1	1.00	0.0000
522b	1	1.00	0.0000
523	1	0.20	0.0000
528a	1	0.20	0.0000

530	2	0.50	0.0000
530b	1	1.00	0.0000
531	5	3.00	4.8000
531a	1	2.00	0.0000
536	2	3.50	0.2500
537	2	0.92	0.0064
538	2	2.50	2.2500
542	4	1.25	0.1875
542a	1	1.00	0.0000
542b	2	1.00	0.0000
543	6	1.67	0.2222
544	1	1.00	0.0000
545a	1	0.50	0.0000
549b	1	1.00	0.0000
551	2	2.00	1.0000
551a	1	1.00	0.0000
552	3	1.67	0.2222
552a	1	1.00	0.0000
553	1	1.00	0.0000
553a	1	12.50	0.0000
556	1	1.00	0.0000
557	2	0.75	0.0625
559	2	9.50	72.2500
560	3	1.67	0.8889
570	2	1.00	0.0000
570b	1	1.00	0.0000
572	2	1.25	0.5625
572a	1	3.00	0.0000
572b	2	2.00	0.0000
573	2	0.75	0.0625
574b	1	1.00	0.0000
576	1	1.00	0.0000
577a	1	1.00	0.0000
578	2	1.00	0.0000
580	1	0.84	0.0000
581	1	0.50	0.0000
582	1	4.00	0.0000
583	1	0.50	0.0000
585a	2	1.00	0.0000
611b	1	1.50	0.0000
642	1	0.33	0.0000
642a	1	0.33	0.0000
651	1	0.50	0.0000
660c	1	0.84	0.0000
670c	2	0.75	0.0625
677	1	1.00	0.0000
682b	1	0.50	0.0000
682d	1	1.00	0.0000
700a	1	3.00	0.0000
702	1	0.33	0.0000

704a	1	11.22	0.0000
705a	1	1.00	0.0000
708	1	11.22	0.0000
709b	1	1.00	0.0000
709c	3	2.67	2.8889
709d	1	6.00	0.0000
710	1	1.00	0.0000
713	4	3.75	17.6875
713b	2	4.50	16.0000
714	2	2.50	2.2500
714a	3	1.00	0.0000
715	1	0.50	0.0000
715a	3	0.83	0.0556
715b	2	3.50	6.2500
716	2	1.00	0.0000
717	4	0.68	0.1169
718	1	0.50	0.0000
719	3	4.41	23.2108
719b	2	3.50	0.2500
722	2	3.00	4.0000
723	1	1.00	0.0000
724	1	0.50	0.0000
724a	1	0.33	0.0000
727a	1	8.50	0.0000
728	2	0.35	0.0225
729a	1	1.00	0.0000
729b	1	1.00	0.0000
733	1	1.50	0.0000
733a	1	1.00	0.0000
740	1	3.00	0.0000
747b	1	1.00	0.0000
748c	1	0.20	0.0000
777a	1	1.00	0.0000
780c	1	0.20	0.0000
790	1	0.20	0.0000
790a	1	0.20	0.0000
808	1	2.00	0.0000
822b	1	0.99	0.0000
827b	1	0.20	0.0000
829c	1	4.00	0.0000
829d	1	0.84	0.0000
829e	1	0.33	0.0000
873	1	1.00	0.0000
874	1	0.33	0.0000
878	1	0.33	0.0000
940	1	0.20	0.0000
987a	1	1.00	0.0000
991	1	1.50	0.0000
998	1	2.00	0.0000
998b	1	1.00	0.0000

999 |1| 2.00| 0.0000

意味カテゴリーデータ　j＝3に対するD3　（sanka6-3.D3）

データ形式　分類番号（3桁）｜カテゴリー要素数｜平均度数｜分散

024|1| 1.00| 0.0000
034|1| 2.00| 0.0000
047|1| 1.00| 0.0000
061|1| 1.00| 0.0000
070|1| 0.13| 0.0000
071|1| 0.13| 0.0000
100|1| 0.33| 0.0000
101|1| 0.66| 0.0000
103|1| 0.50| 0.0000
104|1| 0.25| 0.0000
108|1| 0.50| 0.0000
109|1| 1.00| 0.0000
117|1| 0.33| 0.0000
120|5| 1.00| 0.1000
121|3| 2.33| 3.5556
124|1| 2.00| 0.0000
126|5| 1.20| 0.9737
127|5| 1.30| 0.3600
139|1| 0.13| 0.0000
152|1| 0.25| 0.0000
156|1| 0.25| 0.0000
158|2| 0.50| 0.0000
160|1| 1.50| 0.0000
175|1| 0.13| 0.0000
179|1| 0.33| 0.0000
184|1| 0.50| 0.0000
189|1| 0.50| 0.0000
191|1| 3.00| 0.0000
192|1| 1.00| 0.0000
194|1| 0.13| 0.0000
195|1| 0.33| 0.0000
198|2| 0.46| 0.0420
228|1| 1.00| 0.0000
261|1| 3.00| 0.0000
265|1| 3.00| 0.0000
277|2| 0.25| 0.0000
290|1| 0.20| 0.0000
295|1| 0.20| 0.0000
297|1| 12.50| 0.0000
311|1| 0.33| 0.0000
350|1| 0.13| 0.0000
369|2| 6.50| 30.2500
371|1| 1.00| 0.0000

387	1	0.50	0.0000
391	1	1.00	0.0000
392	1	0.25	0.0000
417	1	0.25	0.0000
421	1	0.33	0.0000
424	1	7.00	0.0000
445	1	0.50	0.0000
458	2	1.00	0.0000
464	1	0.50	0.0000
473	1	0.50	0.0000
502	1	1.00	0.0000
503	2	1.00	0.0000
505	4	1.29	1.0331
506	2	1.75	1.5625
507	3	1.11	0.4124
508	5	0.37	0.0706
509	2	0.56	0.1849
511	2	3.00	4.0000
514	2	0.60	0.1600
518	1	1.00	0.0000
522	2	1.00	0.0000
523	1	0.20	0.0000
528	1	0.20	0.0000
530	3	0.67	0.0556
531	6	2.83	4.1389
536	2	3.50	0.2500
537	2	0.92	0.0064
538	2	2.50	2.2500
542	7	1.14	0.1224
543	6	1.67	0.2222
544	1	1.00	0.0000
545	1	0.50	0.0000
549	1	1.00	0.0000
551	3	1.67	0.8889
552	4	1.50	0.2500
553	2	6.75	33.0625
556	1	1.00	0.0000
557	2	0.75	0.0625
559	2	9.50	72.2500
560	3	1.67	0.8889
570	3	1.00	0.0000
572	5	1.90	0.6400
573	2	0.75	0.0625
574	1	1.00	0.0000
576	1	1.00	0.0000
577	1	1.00	0.0000
578	2	1.00	0.0000
580	1	0.84	0.0000
581	1	0.50	0.0000
582	1	4.00	0.0000

583	1	0.50	0.0000
585	2	1.00	0.0000
611	1	1.50	0.0000
642	2	0.33	0.0000
651	1	0.50	0.0000
660	1	0.84	0.0000
670	2	0.75	0.0625
677	1	1.00	0.0000
682	2	0.75	0.0625
700	1	3.00	0.0000
702	1	0.33	0.0000
704	1	11.22	0.0000
705	1	1.00	0.0000
708	1	11.22	0.0000
709	5	3.00	4.4000
710	1	1.00	0.0000
713	6	4.00	17.2500
714	5	1.60	1.4400
715	6	1.67	3.8056
716	2	1.00	0.0000
717	4	0.68	0.1169
718	1	0.50	0.0000
719	5	4.04	14.2237
722	2	3.00	4.0000
723	1	1.00	0.0000
724	2	0.42	0.0072
727	1	8.50	0.0000
728	2	0.35	0.0225
729	2	1.00	0.0000
733	2	1.25	0.0625
740	1	3.00	0.0000
747	1	1.00	0.0000
748	1	0.20	0.0000
777	1	1.00	0.0000
780	1	0.20	0.0000
790	2	0.20	0.0000
808	1	2.00	0.0000
822	1	0.99	0.0000
827	1	0.20	0.0000
829	3	1.72	2.6350
873	1	1.00	0.0000
874	1	0.33	0.0000
878	1	0.33	0.0000
940	1	0.20	0.0000
987	1	1.00	0.0000
991	1	1.50	0.0000
998	2	1.50	0.2500
999	1	2.00	0.0000

意味カテゴリーデータ　j＝4に対するD2　（sanka6-4.D2）

データ形式　分類番号（2桁）｜カテゴリー要素数｜平均度数｜分散

02|1|　1.00|　0.0000
03|1|　2.00|　0.0000
04|1|　1.00|　0.0000
06|1|　1.00|　0.0000
07|2|　0.13|　0.0000
10|6|　0.54|　0.0596
11|1|　0.33|　0.0000
12|19|　1.39|　1.1506
13|1|　0.13|　0.0000
15|4|　0.38|　0.0156
16|1|　1.50|　0.0000
17|2|　0.23|　0.0100
18|2|　0.50|　0.0000
19|6|　0.90|　0.9696
22|1|　1.00|　0.0000
26|2|　3.00|　0.0000
27|2|　0.25|　0.0000
29|3|　4.30|　33.6200
31|1|　0.33|　0.0000
35|1|　0.13|　0.0000
36|2|　6.50|　30.2500
37|1|　1.00|　0.0000
38|1|　0.50|　0.0000
39|2|　0.63|　0.1406
41|1|　0.25|　0.0000
42|2|　3.66|　11.1222
44|1|　0.50|　0.0000
45|2|　1.00|　0.0000
46|1|　0.50|　0.0000
47|1|　0.50|　0.0000
50|19|　0.95|　0.6851
51|5|　1.64|　2.9184
52|4|　0.60|　0.1600
53|15|　2.19|　3.0872
54|16|　1.28|　0.2490
55|14|　3.29|　25.3112
56|3|　1.67|　0.8889
57|15|　1.27|　0.4289
58|6|　1.31|　1.4936
61|1|　1.50|　0.0000
64|2|　0.33|　0.0000
65|1|　0.50|　0.0000
66|1|　0.84|　0.0000
67|3|　0.83|　0.0556
68|2|　0.75|　0.0625
70|10|　4.18|　15.4412

```
71|30|    2.28|    8.7128
72|10|    1.90|    6.5685
73|2|    1.25|    0.0625
74|3|    1.40|    1.3867
77|1|    1.00|    0.0000
78|1|    0.20|    0.0000
79|2|    0.20|    0.0000
80|1|    2.00|    0.0000
82|5|    1.27|    1.9489
87|3|    0.55|    0.0998
94|1|    0.20|    0.0000
98|1|    1.00|    0.0000
99|4|    1.63|    0.1719
```

意味カテゴリーデータ　j＝7に対するＤＷ　（sanka6-7.DW）

データ形式　語形｜分類番号（４桁）｜度数

```
チーム|715b|    6.00
企業|369 |   12.00
議員|531 |    7.00
研究|424 |    7.00
国|704a|   11.22
国|708 |   11.22
国|719 |   11.22
社|713b|    8.50
社|727a|    8.50
選手|559 |   18.00
代表|297b|   12.50
代表|553a|   12.50
団体|713 |   11.00
日本|709d|    6.00
```

意味カテゴリーデータ　j＝7に対するＤ４　（sanka6-7.D4）

データ形式　分類番号（４桁）｜カテゴリー要素数｜平均度数｜分散

```
582 |1|    4.00|    0.0000
829c|1|    4.00|    0.0000
```

意味カテゴリーデータ　j＝7に対するＤ３　（sanka6-7.D3）

データ形式　分類番号（３桁）｜カテゴリー要素数｜平均度数｜分散

```
034|1|    2.00|    0.0000
121|3|    2.33|    3.5556
124|1|    2.00|    0.0000
```

191	1	3.00	0.0000
261	1	3.00	0.0000
265	1	3.00	0.0000
511	2	3.00	4.0000
531	5	2.00	0.8000
536	2	3.50	0.2500
538	2	2.50	2.2500
700	1	3.00	0.0000
709	4	2.25	2.6875
719	4	2.25	1.6875
722	2	3.00	4.0000
740	1	3.00	0.0000
808	1	2.00	0.0000
999	1	2.00	0.0000

意味カテゴリーデータ　j＝7に対するD2　（sanka6-7.D2）

データ形式　分類番号（2桁）｜カテゴリー要素数｜平均度数｜分散

02	1	1.00	0.0000
04	1	1.00	0.0000
06	1	1.00	0.0000
07	2	0.13	0.0000
10	6	0.54	0.0596
11	1	0.33	0.0000
12	15	1.17	0.4934
13	1	0.13	0.0000
15	4	0.38	0.0156
16	1	1.50	0.0000
17	2	0.23	0.0100
18	2	0.50	0.0000
19	5	0.47	0.1001
22	1	1.00	0.0000
27	2	0.25	0.0000
29	2	0.20	0.0000
31	1	0.33	0.0000
35	1	0.13	0.0000
36	1	1.00	0.0000
37	1	1.00	0.0000
38	1	0.50	0.0000
39	2	0.63	0.1406
41	1	0.25	0.0000
42	1	0.33	0.0000
44	1	0.50	0.0000
45	2	1.00	0.0000
46	1	0.50	0.0000
47	1	0.50	0.0000
50	19	0.95	0.6851
51	3	0.73	0.1422

52	4	0.60	0.1600
53	5	0.77	0.0513
54	16	1.28	0.2490
55	12	1.29	0.4358
56	3	1.67	0.8889
57	15	1.27	0.4289
58	5	0.77	0.0513
61	1	1.50	0.0000
64	2	0.33	0.0000
65	1	0.50	0.0000
66	1	0.84	0.0000
67	3	0.83	0.0556
68	2	0.75	0.0625
70	2	0.67	0.1122
71	22	1.03	0.5394
72	7	0.65	0.1025
73	2	1.25	0.0625
74	2	0.60	0.1600
77	1	1.00	0.0000
78	1	0.20	0.0000
79	2	0.20	0.0000
82	4	0.59	0.1105
87	3	0.55	0.0998
94	1	0.20	0.0000
98	1	1.00	0.0000
99	3	1.50	0.1667