日本文における付属語の統計的性質とその利用例

山内　佐敏　　林　大川　　舟部　敏行

（株）リコー　　技術本部

１．はじめに

現在、一般のオフィスで、文書処理の対象となる現代日本文は論文、報告書等で用いられ、一方的に意志を伝える“である”調と、手紙文等で用いられ、相手の意向を伺う丁寧語、謙譲語を多く含む“です、ます”調とでは、文体が大きく異なる。この文体の違いを特徴づけている要素は、一つには“お”“ご”の接頭辞が使用されるか、されないかの違いがあるが、もう一つには付属語の用い方に大きな違いがある。

また、日本文での単語の概念は、欧文のそれに比べてあまりはっきりせず、定義も論者によってまちまちである。そのため単語単位で分かち書きをさせようとすると、人によってのバラツキが大きいだけでなく、一個人でも時によって変化してしまう。それに対して文節の概念は、形式名詞、補助用言の定義で若干のゆれはあるものの付属語と自立語との境界は比較的わかりやすいため、文節単位での分かち書きをさせた場合、人によるバラツキは小さい。近年急速に普及してきた日本語文書処理機に用いられている入力方式には上記のような要素も考慮されてのことであろう、文節単位の仮名漢字変換方式を採用しているものが多い。

本稿では、日本文の文体を特徴づける付属語の用いられ方を頻度と付属語間の接続確率の面から調べ、その結果と、それを用いて仮名漢字変換ソフトの速度向上検討を行なった内容をあわせて報告する。

２．調査対象文と付属語の定義

報告調文体の文書として技術論文[1)]の要旨とまえがき部分４５文例をとり上げ、丁寧語を多く含む文体の文書として事務文例[2)]２２８例をとり上げた。

表１　事務文２２８例中の付属語出現度数

NO.	頻度	種類	付属語
1	2571	格助詞	の
2	1924	丁寧助動詞	ます
3	1830	格助詞	に
4	1274	格助詞	を
5	1055	副助詞	は
6	1048	接続助詞	て
7	812	形式名詞	うえ（他17語）
8	784	形式動詞	する
9	707	格助詞	と
10	645	断定助動詞	だ
11	582	過去助動詞	た
12	424	格助詞	が
13	269	形容動詞	だ
14	267	副助詞	も
15	265	補助用言的	ついて，つき
16	226	接続助詞	が
17	219	形式動詞	おる
18	218	格助詞	から
19	212	形式動詞	ある
20	205	準体助詞	の
21	138	形式動詞	よる
22	133	副助詞	か
23	128	接続助詞	と
24	128	形式動詞	なる
25	127	打消助動詞	ない
26	122	受身助動詞	れる
27	114	副助詞	まで
28	108	格助詞	して
29	94	断定助動詞	です
30	88	希望助動詞	たい
31	74	形式動詞	いる
32	64	使役助動詞	せる
33	61	形式動詞	おる
34	58	打消助動詞	ん(ぬ)
35	51	接続助詞	ながら
36	49	副助詞	なぞ
37	45	接続助詞	ば
38	38	接続助詞	ところ
39	37	格助詞	へ
40	35	可能動詞語尾	る
41	29	受身助動詞	られる
42	26	接続助詞	から
43	23	格助詞	で
44	20	接続助詞	に
45	18	形式動詞	くる
46	18	形式動詞	いう
47	15	格助詞	より
48	14	副助詞	ばかり
49	13	推量助動詞	う
50	13	用言性接尾辞	はらう

また、調査量の影響がどの程度あらわれるかをみるために、事務文例２２８例中の約１／１０にあたる２１例をとり出し比較した。

またここで付属語として扱うのは、一般の国語文法で言う助詞、助動詞の他、形式名詞、形式動詞、用言性接辞等、仮名表記する単語（自立語以外）を含めており、その総数は２１２種である。

３．付属語の出現確率分布

事務文２２８例中で使用された付属語は全部で１０７種であった。上位５０位の内容を表１に示す（総頻度数１７６３１）。またこの出現確率分布特性、およびその累積分布特性を図１に示す。

この出現確率分布特性は、指数分布

$$P(x)=a\,e^{-bx} \quad : a, b \text{は定数} \qquad (1)$$

の形で近似できる。

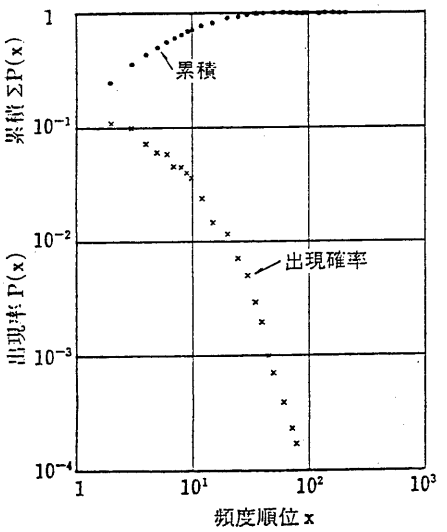


図１．頻度順付属語出現確率（事務文２２８例）

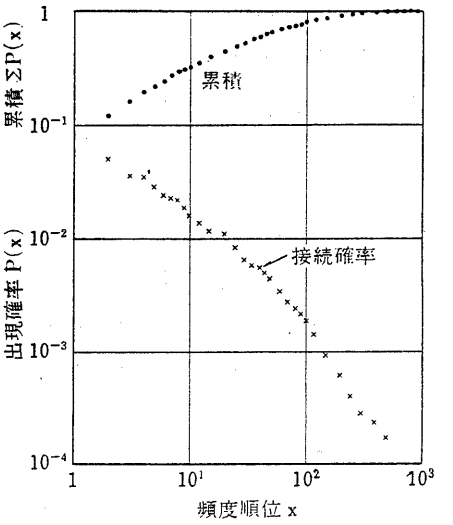


図３．頻度順付属語接続確率（事務文２２８例）

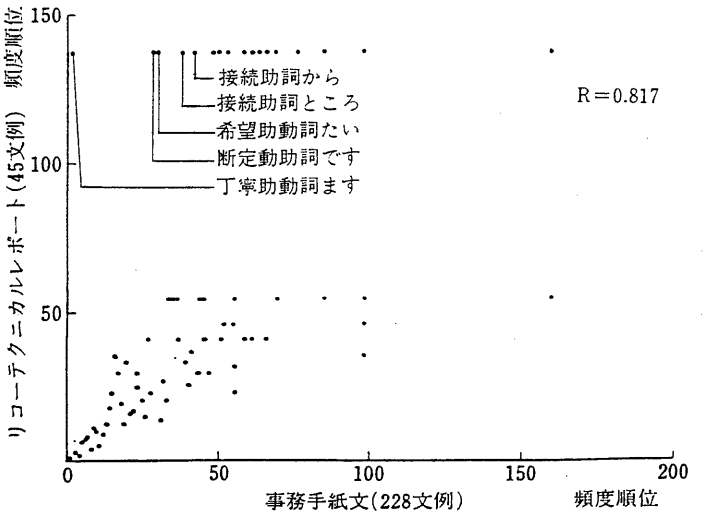
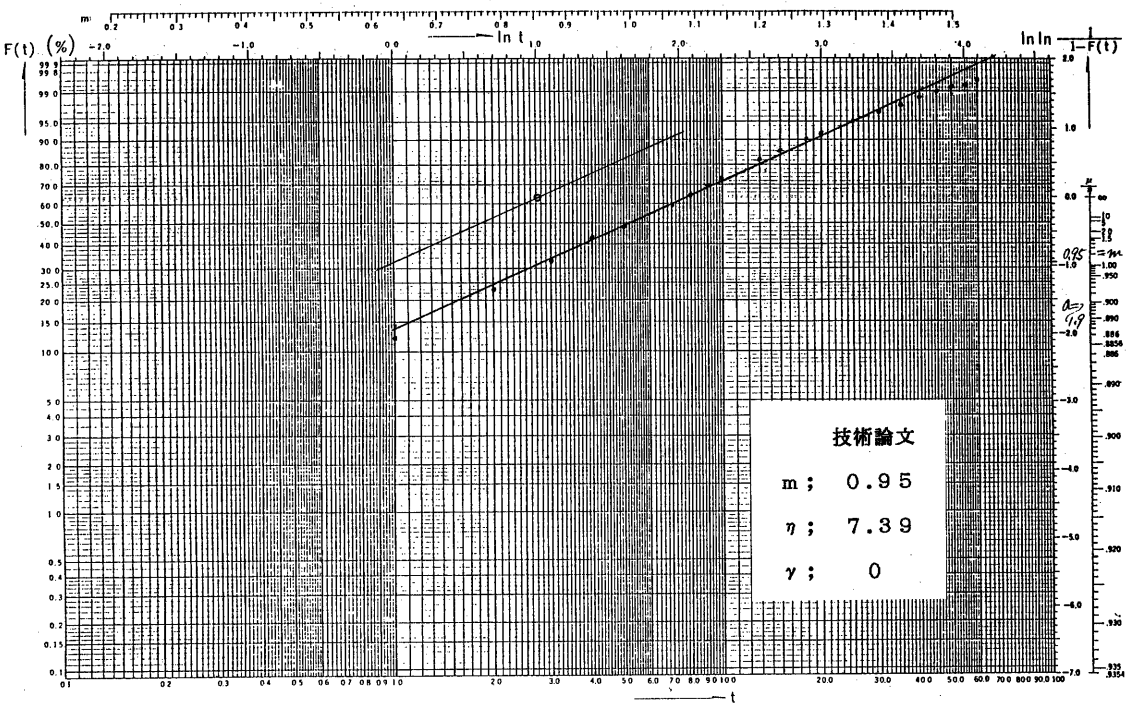
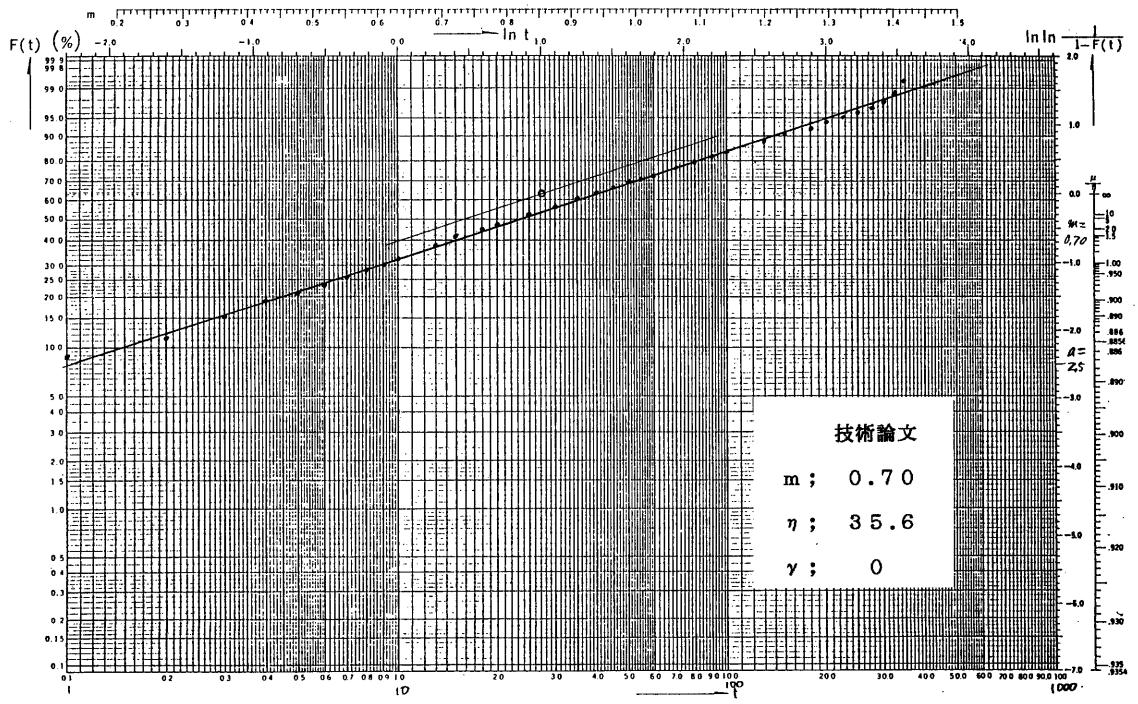


図２．付属語の事務文２２８例と技術論文間の相関図



a）出現頻度の累積分布



b）接続確率の累積分布

図４．ワイブル確率紙による累積分布（技術論文）

技術論文での出現確率分布特性および事務文２１文例でのそれもほとんど同形であるが、各順位に対応する個々の語は異なる。

回帰した結果の各パラメータと実データと回帰式との相関係数ｒは次のようになっている。

	技術論文	事務文228例	事務文21例
a；	8.7×10^{-2}	4.3×10^{-2}	6.1×10^{-2}
b；	1.0×10^{-1}	7.1×10^{-2}	8.2×10^{-2}
r；	0.98	0.98	0.98

また出現確率の累積分布特性は（1）式を積分しても良いがワイブル分布

$$F(x)=1-e\{-(x-\gamma)/\eta\}^{m} \quad (2)$$

ただし　m：形のパラメータ

η：尺度のパラメータ

γ：位置のパラメータ

で近似できる。日科技連ワイブル確率紙[3]で回帰した結果のｍとηの値を次に示す。

	技術論文	事務文228例	事務文21例
m；	0.95	0.85	0.90
η；	7.39	8.82	7.56
γ；	0	0	0

図２に事務文２２８例と技術文との順位相関図を示す。特徴的な付属語としては、事務文２２８例で上位となった丁寧助動詞“ます”、断定助動詞“です”、希望助動詞“たい”等は、技術論文中には一度も使われていないが、その逆に技術論文中で上位となった付属語で事務文に使われていない語はない。

他の組合せの順位相関も調べたので、それぞれの相関係数Ｒを次に示す。

	R
事務文２１例－事務文２２８例	0.880
事務文２１例－技術論文	0.838
事務文２２８例－技術論文	0.817

４．付属語の接続確率分布

図３に事務文２２８例の付属語の接続確率分布特性と累積分布特性を示す。

この接続確率分布特性も付属語の出現確率分布と同様指数分布式（1）で近似できる。回帰した結果の各パラメータと回帰式との相関係数ｒを同様の形で次に示す。

	技術論文	事務文228例	事務文21例
a；	7.7×10^{-3}	7.1×10^{-3}	5.0×10^{-3}
b；	1.1×10^{-2}	4.8×10^{-3}	7.6×10^{-3}
r；	0.92	0.91	0.90

接続確率の累積分布特性もワイブル分布式（2）で近似すると各パラメータは

	技術論文	事務文228例	事務文21例
m；	0.70	0.59	0.63
η；	35.6	47.0	45.1
γ；	0	0	0

となる（このうち技術論文のワイブル分布図を、先の付属語の出現確率分布のそれとあわせて図４に示す）。

これも出現確率分布と同様、各分布は似たような分布となっているが、各順位に対応する個々の語は極端に異なる。各分布間の順位相関係数Ｒは

	R
事務文２１例－事務文２２８例	0.630
事務文２１例－技術論文	0.143
事務文２２８例－技術論文	0.325

となり、事務文と技術論文の間では無相関の方に近いことがわかる。

特徴的な接続の相違としては事務文で上位となった一般動詞連用形＋丁寧助動詞“ます”、丁寧助動詞連用形“まし”＋過去助動詞“た”、あるいは接続助詞“て”等の繋がりが技術論文中には一度も出

現せず、技術論文中で上位の、た行五段活用動詞連用形促音便“っ”＋過去助動詞“た”、サ変動詞未然形“さ”＋使役助動詞“せる”等の繋がりは事務文中には一度も出現していない。

5．文字種の頻度特性との比較

筆者等[4]は以前に日本文事務文書における文字種の調査を行なった。そのおり、文字種の頻度分布をワイブル分布で回帰している。それによると各パラメータは表6のようになる。

表6　日本文事務文書中の各文字種の
ワイブル分布パラメータ

区分	m	η
記号	1．1	4．32
数字	1．1	4．53
英字	0．84	15．5
ひらがな	0．97	12．0
カタカナ	0．83	16．0
漢字	0．70	150．0
総合	0．70	72．2

注）原典ではα（$=\eta^{-m}$）で表示されている。

この内容と付属語の内容を比較してみると、付属語の頻度分布特性では文字種の少ない文字セット、すなわち記号、数字、英字、ひらがな、カタカナに近く、付属語間の接続確率分布特性では文字種の多い文字セット、すなわち漢字、総合（日本文全体で使われる文字の総合）に近い特性であることがわかる。

6．本調査の利用例

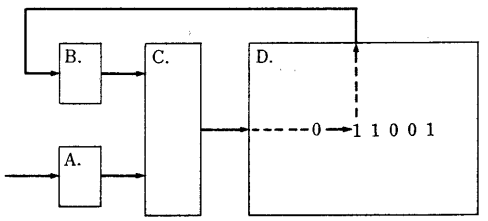
以上のような調査結果は、かな漢字変換装置等で行なう文法解析の効率向上に利用することができる。例えば、相沢ら[5]が報告しているような方法、すなわち、自立語を最長一致法で仮確定し、接続行列表の検索により順次接続可となっている付属語と後続の文字列を比較していき、文節末までの一致を確認する方法には有効である。

この方法では、接続行列表の検索において、接続の成功確率が大きいものを優先的に検索するほうが、検索時間を短縮させられる。そこで、最も適切な配列を決定する基準として、後続する付属語の並びを表わす列番号を、ランダムに配列した場合と使用される頻度の高い順に配列した場合とを比較し、どの程度検索時間が短縮されるかを検討した。

6－1．接続行列表検索モデル

図5に検索モデルを示す。以下に検索モデルの動作を説明する。

まず、自立語が決定された段階で、その品詞をキーとして、Aを検索しCへ行く。Cにおいて、Dの行番号が決定されDへ行く。ただし、自立語が動詞の場合、活用形を決定してから、Dへ行く。Dの検索において、「1」の立っている列番号のところで、それをキーとして、Bを検索しCへ行く。Cにおいて付属語を検索する。まだ付属語候補の文字列が残っている場合はDへ行き、それ以外は、検索を終了する。



A：自立語後続索引
B：付属語索引
C：付属語表
D：接続行列表

図5　続接行列表検索モデル

6－2．検出時間の算出

検索モデルにおいて、一文節に対する平均検索時間を求める。平均検索時間$\tilde{T}$は、次のように定義できる。

$$\tilde{T}=\tilde{T}_1+\tilde{T}_2+\tilde{T}_3+\tilde{T}_4 \quad (3)$$

$\tilde{T}_1$ ；接続行列表（D）内の平均検索時間

$\tilde{T}_2$ ；D→B→C→Dの平均検索時間

$\tilde{T}_3$ ；A→C→Dの平均検索時間

$\tilde{T}_4$ ；自立語の切り出しに失敗した時の平均検索時間

$\tilde{T}_1$は

$$\tilde{T}_1=\tilde{n}\,E(x)\,t_1 \quad (4)$$

$\tilde{n}$ ；Dの最初の列番号から検索する平均回数

E(x)；接続が成功する列番号の期待値

t_1 ；x列からスタートして、x＋1列を検索する時間

である。

$\tilde{T}_2$は

$$\tilde{T}_2=\tilde{n}\,(3\,\tilde{f}\,t_2+\tilde{h}\,t_3) \quad (5)$$

$\tilde{f}$ ；接続が成功する列（その列も含む）までの１の平均個数

$\tilde{h}$ ；C内部で実際に検索する平均付属語候補数

t_2 ；表間、索引・表間の移動の検索時間

t_3 ；付属語候補１個の検索時間

である。

$\tilde{T}_3$は

$$\tilde{T}_3=2\,t_2+\tilde{h}\,t_3 \quad (6)$$

である。

$\tilde{T}_4$は

$$\tilde{T}_4=(\tilde{T}_1+\tilde{T}_2+\tilde{T}_3)\tilde{m} \quad (7)$$

$\tilde{m}$ ；自立語の切り出しに失敗した平均回数

よって式(3)は、

$$\tilde{T}=(1+\tilde{m})\{\tilde{n}(E(x)\,t_1+3\,\tilde{f}\,t_2+\tilde{h}\,t_3)+2\,t_2+\tilde{h}\,t_3\} \quad (8)$$

となる。ここで$t_1=t$、$t_2=at$、$t_3=bt$とすると（a、bは定数）

$$\tilde{T}=(1+\tilde{m})\{\tilde{n}(E(x)+3\,a\,\tilde{f}+b\,\tilde{h})+2\,a+b\,\tilde{h}\}\,t \quad (9)$$

となる。

6－3．接続行列表の構成

前述の検索時間のうちで接続行列表のみが配置順によって、大きく変動をうける。

1）ランダム構成

接続行列表をランダムに配列した時の接続成功列番号の期待値（Er(x)），標準偏差（Dr(x)）を求める。

期待値（E(x)）は次式で算出する。

$$E(x)=\sum_{i=0}^{n} x_i\cdot p_i \quad (10)$$

x_i：列番号

p_i：各列番号で接続が成功する確率

標準偏差（D(x)）は次式で求める。

$$D(x)=\sqrt{E(x^2)-\{E(x)\}^2} \quad (11)$$

ランダム構成において、p_iは各列とも等しいので、式（１０）は次のように変形できる。

$$Er(x)=\frac{1}{I}\sum X_i \quad (12)$$

I：列番号数

ここで、I＝２１２として計算すると、

Er(x)＝１０７

となる。列を検索した場合、平均１０７番目の列で接続が成功することが期待できる。

標準偏差は、

$Dr(x)=61$

となる。

2）頻度順構成

接続行列表を頻度順に配列した時の接続成功列番号の期待値（$E_0(x)$），標準偏差（$D_0(x)$）を同様に算出する。

式(10)は　次のように変形できる。

$$E_0(x)=\sum_{i=0}^{n} x_i \cdot P(x_i) \qquad (13)$$

$P(x_i)$は、図1の確率曲線で示される確率である。

式(1)の回帰式は、事務文228例を用いると、

$$P(x_i)=4.3\times10^{-2}\times e^{[-7.1\times10^{-2}x_i]} \qquad (14)$$

よって

$E_0(x)=15$，$D_0(x)=14$となる。

列を検索した場合、平均15番目の列で接続が成功することが期待できる。

3）実用構成

上記のように接続行列表を頻度順に配列することにより、検索時間を短縮できるので頻度順構成にするのが良い。ただし、接続行列表のメンテナンス等を考慮すると、多少頻度順を犠牲にせざるをえない。以下に、実用を考慮した接続行列表の期待値（$E_m(x)$），標準偏差（$D_m(x)$）を算出する。

$Em(x)=26$，$Dm(x)=21$

平均26番目の列で接続が成功することが期待できる。

また、参考に、当初作成した（以下「初期構成」と言う）接続行列表を用いた場合の期待値（$En(x)$），標準偏差（$Dn(x)$）を同様の手法で算出する。

$En(x)=50$，$Dn(x)=37$

平均50番目の列で接続が成功することが期待できる。

4）各構成の比較

接続行列表内の平均検索時間$\tilde{T}_1$について上記4つを比較する。表3にランダム構成の$\tilde{T}_1$を基準とした各$\tilde{T}_1$の比率を示す。初期構成を実用構成と比較すると、

初期構成：実用構成　＝　1：0.51

で実用構成の平均検索時間$\tilde{T}_1$は約半分になることが期待できる。

表3　平均検索時間の比較

構　成	K
ランダム構成	1.0
頻度順構成	0.14
初期構成	0.47
実用構成	0.24

K…接続行列表内の平均検索時間の比率

6－4．実文による検索処理速度向上の確認

事務文例より10文節をランダムに抽出し（表4にしめす）、初期構成と実用構成の接続行列表を用いて、検索回数を調べ、比較し、どの程度検索時間が短縮するかを検討した。

表4　抽出した文節

No.	文　節	No.	文　節
1	皆さまに	6	部分にも
2	上りましたが	7	郵便に
3	願えれば	8	検討の
4	拝見いたし	9	実積を
5	申し上げます	10	対しまして

表５．接続行列表の検索時間（単位；検索回数）

NO. 構成	1	2	3	4	5	6	7	8	9	10	計	回数比
初期構成	46	83	82	234	33	79	46	45	48	86	782	1
実用構成	25	53	56	118	11	37	25	24	26	52	429	0.55

注：２個以上付属語を持つものは最後の付属語の検索時間を取る。

表６．全検索時間（単位；検索回数）

NO. 構成	1	2	3	4	5	6	7	8	9	10	計	回数比
初期構成	67	538	335	351	46	198	67	60	73	1184	2919	1
実用構成	37	365	196	225	14	101	37	33	39	678	1725	0.59

１）接続行列表内の検索時間

表５は、表４の１０文節について、初期構成と、実用構成の各接続行列表を用い、接続行列表内の実際の検索回数を調べた結果である。表５から、実用構成は初期構成と比較すると、検索時間は０．５５倍に減少していることがわかる。これは６－３で述べた期待値０．５１倍に近い。

２）全検索時間

表６は、表４の１０文節について、初期構成と、実用構成の各接続行列表を用い、実際の全検索回数を調べた結果である。表６から、実用構成は、初期構成と比較すると。全検索時間は、平均０．５９倍に減少していることがわかる。

以上のように、配列に頻度順を考慮した接続行列表を採用することで、当初のものと比較し実用的にも半分近くに検索時間を短縮している。

７．おわりに

以上述べたように、付属語の使われ方が日本文中の文字の使われ方に似ていることがわかった。また、このような調査を行なうことにより、具体的な装置の改善を机上で行なえることを示した。今後は自立語や接辞の特性も調査し、日本文中の文字や単語のふるまいについて考察していきたい。

最後に本調査について御協力いただいた関係各位に、特にデータの整理に献身的につくしてくれた島田美佐子嬢に深謝いたします。

参考文献

１）リコー：“RICOH TECHNICAL REPORT” NO.1～NO.6，(1979 ～ 1980)

２）山城，安田：“企業経営文例全書” １～４ ぎょうせい（昭５３）

３）日科技連編：信頼性データの解析（日科技連ワイブル確率紙の使い方)、日科技連ライブラリー １２

４）村山：日本文事務文書における字種の解析と応用，RICOH TECHNICAL REPORT NO.6 （昭和５６年１１月２５日）

５）相沢，江原：計算機によるカナ漢字，ＮＨＫ技術研究，Vol ２５，NO. ５，pp. ２６１～２９８（１９７３）