自然言語処理　59－3
（1987. 1. 23）

漢字と読みの組を造語単位とした単語の造語モデル

福永 博信 ・ 松延 栄治 ・ 日高 達 ・ 吉田 将
（九州大学総合理工学研究科）

筆者らは，これまで日本語の文節文法の確率文法化の研究を行ってきた．確率文法を構成する上で問題となるのは，書き換え規則にいかにして適切な確率を付与するかである．解析の分野，語彙等を限定せず，汎用の文法を目指す立場においては，書き換え規則の数は膨大なものとなり，その中でも終端記号である単語への書き換え規則の数は数万に達し，それら個々に精度良く確率を付与するのは困難である．そこで，日本語においては漢字が意味的音韻的に結合して単語を造語するという性質に着目した単語の造語モデルを構成し，そのモデルから単語への書き換え規則の確率を推定している．

"A Marcov Model for Japanese Words
in terms of meaning and phonetic succession" (in Japanease)

Hironobu FUKUNAGA , Eiji MATSUNOBU , Toru HITAKA , Sho YOSHIDA
Interdisciplinary Graduate School of Enjineering Sciences, Kyushu University
6-1 kasuga-koen, kasuga-shi, fukuoka-ken 816 Japan

We have been trying to construct a Japanese stochastic regular grammar. Our stochastic grammar consists of words succession model and word model. The words succession model is based on Japanese BUNSETSU rules and estimates conditional probabilities of successive words in Japanese sentences. The word model estimates occurrence of words.

In this paper, we show our word model and how to get the word model from an ordinary sized Japanese word dictionary.

1．はじめに

自然言語の解析においては，常に曖昧さが問題となる．これに対する１つの解決策として筆者らは，文法の確率化を試み，日本語文の解析システムを作成している．このシステムは，入力文を単語の連接で規定し，単語の生起確率と隣接する単語間の連接確率の積からなる確率を評価値として採用することによって曖昧さを減少させようとするものである．単語の生起確率については，それを精度よく推定できるだけのテキストの収集が，現段階では困難であるため，日本語の漢字の性質に着目した単語の造語モデルを構成して確率の推定を行っている．

本来，自然言語というものは人間の意志を伝達するための手段であり，文字と音声による表現が用いられている．故に，単語の造語においても，意味的音韻的両面の考慮が必要であると考え，従来の漢字の意味的関係のみに注目したモデルを，音韻的関係を同時に考慮したものに拡張した．すなわち，漢字と読みの組を造語単位とした単語の造語モデルである．

2．確率文法

確率文法とは，従来の形式言語の生成規則に，その規則の適用される確率を付与したものである．但し，Xの生成規則

$$X \xrightarrow{p_1} Y_1$$
$$\vdots$$
$$X \xrightarrow{p_n} Y_n$$

において，適用確率の総和は１である．

$$\sum_{i=1}^{n} p_i = 1$$

図１のような解析木Tを考えると，

解析木T：

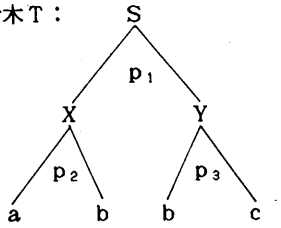

図１　確率文法による解析木

解析木Tの生起確率は，各生成規則の適用確率の積で与えられる．

$$P(T) = p_1 \cdot p_2 \cdot p_3$$

また，文字列の生起確率は，その文字列を生成する全ての解析木の生起確率の総和で与えられる．

$$P(abbc) = \sum_{\text{trees for abbc}} P(T)$$

確率文法の応用は，いくつか考えられるが，解析木の生起確率を用いたものとして，自然言語の解析における曖昧さの減少への利用がある．すなわち，与えられた入力文字列に関して，可能な全ての解析木中で最大の生起確率p(T)をもつ解析Tを妥当な解析とすることで，曖昧さを減少させることができる．文字列の生起確率を用いたものとしては，文字認識や音声認識等での言語処理への適用があり，認識率の向上が期待される．

3．文節構造規則と確率文節文法

日本語の統語規則は，単語の連接で文節を規定した規則(文節構造規則)と文節の連接で文を規定した規則(係り受け規則)の２つで捉えることができる．以下，確率文節文法の基礎となっている文節構造規則について述べる．文節構造規則は，文節内の単語の連接可能性を規定するであるが，ここではそれを文内の単語の連接を規定するものへ拡張したものを用いている．

【定義１】　単語構造

３組（w,H,K）を単語Wの単語構造と呼ぶ
ただし，w,H,Kは，それぞれ単語Wの綴り,品詞,活用形を表す．■

本稿でいう品詞とは，一般に用いられている品詞を活用の種類毎に細分類し，さらに付属語については１語１品詞に分類したものである．また，活用形は，活用の種類毎の活用形である．従って,正確にはK^Hであり品詞と活用形で単語が特定できるようになっている．

【定義２】　述語C

単語構造(w_i,H_i,K_i)，(w_j,H_j,K_j)である単語W_i,W_jが文中で連接可能であることを３変数述語Cを用いて，$C(H_i,K_i,H_j)$と記す．■

【定義３】　文節構造規則

単語列$W_1,W_2,\cdots W_n$において，単語W_kの単語構造を(w_k,H_k,K_k)で表すと，次の規則を満たす時単語列$W_1,W_2,\cdots W_n$は文をなすという．

$$\bigwedge_{k=0}^{n} C(H_k,K_k,H_{k+1}) \quad ■$$

ただし，(H_0, K_0),(H_{n+1}, K_{n+1})は文の前後の区切れを表すダミーの品詞(H_ω),活用形(K_ω)を表し，H_1が自立語の品詞であるときのみ$C(H_\omega, K_\omega, H_1)$=trueであり，$H_n$と$K_n$が文の末尾になり得る品詞,活用形であるときのみ$C(H_n, K_n, H_\omega)$=tureである．

文節構造規則は，正規文法(regular grammar)であり，品詞Hを内部状態とし，活用形Kによって状態遷移する非決定性有限オートマトンとして規定される．

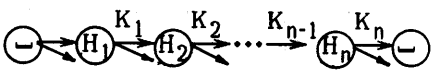

図2　文節構造規則のオートマトン表現

次に，文節構造規則に確率文法化した確率文節文法について述べる．確率文節文法を確率正規文法の枠組で記述すれば，次のようになる．

$G = <N, \Sigma, P_s, S>$

$N = \{S\} \cup \bigcup_i \{H_i, (H_i, K_1^{H_i}), (H_i, K_2^{H_i}), \cdots\}$

$\Sigma = \bigcup_i \{w_i\}$

P_s：次の２つのタイプがある

(I)　$S \xrightarrow{q} H$　　……①

$(H,K) \xrightarrow{r} H'$　　……②

$(H,K) \xrightarrow{s} \varepsilon$　　……③

(II)　$H \xrightarrow{t} w(H,K)$　　……④

タイプ(I)の生成規則①,②,③は文節構造規則を反映したものであり，$C(H_\omega, K_\omega, H)$, $C(H, K, H')$, $C(H, K, H_\omega)$に対応している．タイプ(II)の生成規則④は，品詞から終端記号すなわち具体的な単語の綴りwへの生成規則と考えられる．各生成規則の適用確率は次式で与えられる．

$q = p(H \mid \_)$　　……⑤

$r = p(H' \mid (H,K))$　　……⑥

$s = p(\_ \mid (H,K))$　　……⑦

$t = p(K \mid H) \cdot p(MW \mid H)$　　……⑧

ただし，MWは単語Wの語幹部分の綴り

タイプ(I)の生成規則に伴う確率(q,r,s)は，非終端記号の数が少ないため，ある程度の規模のテキスト集合より推定可能である．しかし，タイプ(II)の生成規則は数が多く，それに伴う確率を精度よく推定するためには，莫大な量のテキストデータを収集することが必要となる．実際には，一般単語数万語が最低１度以上出現し，単語の相対的な出現頻度が一般世界での使用頻度を適当に反映しているようなテキストの収集は困難である．⑧式においてp(K | H)の値は，ある程度の規模のテキスト集合より推定可能であるので，p(MW | H)の値を推定するためにモデルを構成する．MWは，活用語尾のない単語については，単語の綴りwに等しく，以後混同の恐れのない限りp(MW | H)を単語の生起確率と呼ぶ．

4．単語の造語モデル

自然言語で用いられる単語の数は膨大なものであり，それらの単語一つ一つに人間の主観にたよって人手で確率を付与するのは非常に困難な作業である．また，計算機でその確率を推定することを考えた場合，前述のとおり，テキストの収集およびその集計は現段階では事実上不可能であると考えられる．そこで筆者らは，日本語における漢字の性質に着目して単語の造語モデルを構成し，それによって単語の生起確率を推定している．

4-1　造語単位

単語の造語モデルを構成するにあたって，単語の造語に関して次の仮定をおく．

【仮定１】　漢字による単語の造語
日本語では，意味を担う最小単位（造語単位）は漢字であり，漢字が意味的音韻的に結合（造語）して単語が生成される．■

本来，自然言語と言うものは，人間がものごとを表現し，伝達するための手段であって，文書による表現と音声による表現が用いられており，単語の造語に関しても意味的結合と音韻的結合の両面を考慮することが必要であると考える．

単語の中には，漢字だけでは表現できないもの，即ち非漢字文字（ひらがな，カタカナ等）を含むものがある．それら非漢字文字は表音文字であって，表意文字である漢字とは異なり，１文字では音を表すのみで意味は表さない．そこで，非漢字文字については，文字種毎に一連の非漢字文字列を一まとめにして，便宜的に漢字１文字と同等に扱う．

ここで改めて単語を構成する最小単位である造語単位を定義する．

【定義４】　造語単位
造語単位は，綴りとその読みの組から構成される次のものを造語単位とする．
(1) 単語を構成する漢字１文字とそれに続く一連のひらがな列は造語単位である
(2) 単語中に存在する一連の同一文字種の非漢字文字列は造語単位である
■

定義4により，次の5つのタイプの造語単位が考えられる．

TYPE－1	漢字1文字	
TYPE－2	漢字1文字	＋　ひらがな列
TYPE－3	ひらがな列	
TYPE－4	カタカナ列	
TYPE－5	英数字・記号	

TYPE-2の造語単位は漢字の送り仮名を考慮したものである．送り仮名は，送り仮名部分だけでは意味を表さず，漢字と連接することによって音を整える役割を持つ．したがって，意味的にも音韻的にも漢字と送り仮名とは切り離して考えるべきではない．また，送り仮名の送り方は人によってまちまちであり，複数の送り仮名が存在するものもある．このように送り仮名を漢字に含めて取り扱えば，数種の送り仮名に対して造語単位の読みかえで比較的容易に対処することができる．

［例］　造語単位の例

計/けい　算/さん　機/き	・・・ 計算機
お/お　好み/このみ　焼き/やき	・・・ お好み焼き
テーブル/てーぶる　掛け/かけ	・・・ テーブル掛け
X/えっくす　線/せん	・・・ X線

造語単位の抽出は国語辞書の見出しを走査し，文字種に応じて切り出すことで行っている．すなわち，下図の文字種を状態とするオートマトンで容易に切り出すことができる．初期状態Sから出発し，再び状態Sへ戻るまでを1造語単位とする．4つの文字種状態への遷移は，その状態名の文字種と該当文字の文字種が一致した場合に起こり，その文字を消費する．状態Sへの遷移は，他の遷移が存在しない場合にのみ起こり，文字の消費はしない．

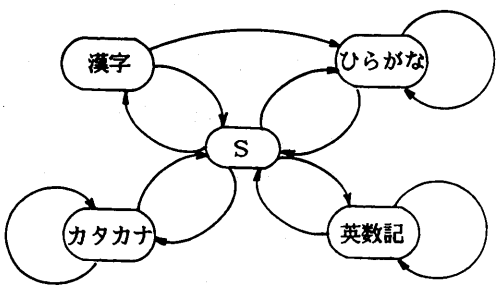

図3　造語単位抽出オートマトン

この作業は品詞毎の単語の生起確率p（w｜H）を推定するためのものであるから，造語単位は品詞別に抽出しなければならないことは言うまでもない．

4-2　造語モデル

4-1節では，単語を造語する最小単位である造語単位を定義し，非漢字文字列を一まとめにして，便宜的に漢字と同等に扱えるようにした．以下4-2節においては，特に混同の恐れのない限り'漢字'という表現は造語単位を表すものとする．

モデルを構成するにあたって，漢字（造語単位）の結合，即ち単語の造語過程に関して次の仮定をおく．

【仮定2】　造語過程
造語過程はマルコフ過程である．　■

仮定2より単語の造語過程をマルコフ過程でモデル化する．N重マルコフモデルでNを大きくすれば，より正確なモデルとなる反面，指数的に増大する遷移の状態を用意する必要がある．本モデルにおいては，基礎となる漢字の種類が多いため，2重以上のモデルでは状態数が膨大な数となり実際的ではない．ゆえに造語モデルは，各造語単位を状態とする1重マルコフモデルを採用する．すなわち，隣接する造語単位間の関係のみを考慮したモデルである．

このモデルは言うなれば単語中の隣接する漢字間の意味的音韻的な親和性に着目したモデルである．漢字が結合して単語を造語する場合，意味的な結合のしやすさや連続発声の困難さ等の傾向があると考えられる．それら漢字間結合の難易傾向を国語辞書の見出しから読み取り，単語の生起確率を推定するのが，本稿の試みである．

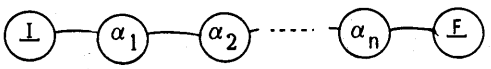

図4　1重マルコフモデルによる
単語の造語モデル

図4に示すように単語を構成する造語単位α_1，α_2，・・・，α_nの前後にダミーの造語単位I，Fを仮定し，隣接する造語単位間での遷移で表されたモデルで考える．

音韻的には，常に直前直後の音のつながりのみが影響を及ぼし合うので，このモデルで妥当であると思われる．

意味的見地からは，多少問題がある．2造語単位以下から構成される単語については，隣接しない造語単位間の関係は存在しないので，このモデルで妥当である．しかし3造語単位以上からなる単語については，隣接しない造語単位間の関係も存在しうる．3造語単位からなる単語について考察する．3造語単位からなる単語は次の3つに分類される．

Ⅰ．隣接するどの造語単位も一まとまりに
ならないもの　　［例］　新生児

Ⅱ．前２つの造語単位が一まとまりとなって
他の１つと関係を持つもの　［例］　情報学

Ⅲ．後２つの造語単位が一まとまりとなって
他の１つと関係を持つもの　［例］　打楽器

Ⅰ型は，順次隣接した造語単位間の関係で表されるので問題はない．Ⅱ型，Ⅲ型は，２つの造語単位が一まとまりとなって他の１つとの関係を持つものであるが，２つの造語単位がまとまって１つの意味を表す際には，後ろの造語単位がその意味を代表する場合が多い．そのように考えると，Ⅱ型はⅠ型で近似されるが，Ⅲ型は，隣接しない造語単位間の関係を持ち，このモデルでは表現できない．しかし，Ⅰ型とⅡ型，Ⅲ型の分類は非常に微妙な部分を含み，完全に意味が分化しているⅡ型，Ⅲ型は少ないと考えられる．

４造語単位以上からなる単語については，極端に数が少なくなる．したがって，一部表現しきれないものを含んではいるが，数量的に見てほぼ妥当なモデルであるといえよう．

４-３　造語単位の細分類

４-２節では，１重マルコフモデルによる単語の造語モデルとその妥当性について述べた．しかし，単純な１重マルコフモデルでは，造語能力（モデルが造語できる全ての単語の数）が大きくなりすぎる．つまり，未知語が多数造語され，それらは単語とは認めがたいものが多い．また，解析のモデルとみた場合，長い漢字列を単語と同定してしまう傾向がある．そこで一般の国語辞書にある見出しは全て造語でき，それ以外の未知語の発生を極力抑制するようなモデルを考える．

造語単位の中には，１造語単位のみで単語を構成するものや，前後のいずれか，或は両方で他の造語単位と結合して単語を構成するという傾向を持つものがある．そこに着目して造語単位を次の４つの型に細分する．ただし，述語B(α)，E(α)はそれぞれ造語単位αが単語の先頭，末尾の造語単位であることを表す．

(1)「α」･･･　B(α) ∧ E(α)
(2)「α　･･･　B(α) ∧～E(α)
(3)　α　･･･　～B(α)∧～E(α)
(4)　α」･･･　～B(α)∧ E(α)

この細分化（ラベル付け）された造語単位を用いたモデルをラベル付き１重マルコフモデルと呼ぶ．

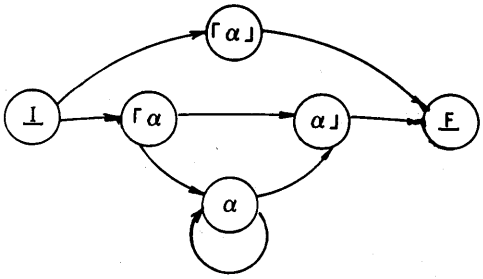

図５　ラベル付き１重マルコフモデル
による単語の造語モデル

造語能力抑制に対するラベル付けの効果を少数単語世界の例を用いて示す．単語集合W＝｛文/ぶん,国語/こくご,文法/ぶんぽう,学問/がくもん,言語/げんご,問題/もんだい,調子/ちょうし,英語圏/えいごけん,文語調/ぶんごちょう,調教師/ちょうきょうし｝（10語）を考える．単純マルコフモデルでは，単語集合Wより造語単位を抽出し,造語単位集合U'＝｛文/ぶん,国/こく,語/ご,法/ぼう,学/がく,問/もん,言/げん,題/だい,英/えい,圏/けん,調/ちょう,子/し,教/きょう,師/し｝（14造語単位）を得る．従って，図６のようなモデル（状態数18）を構成する．

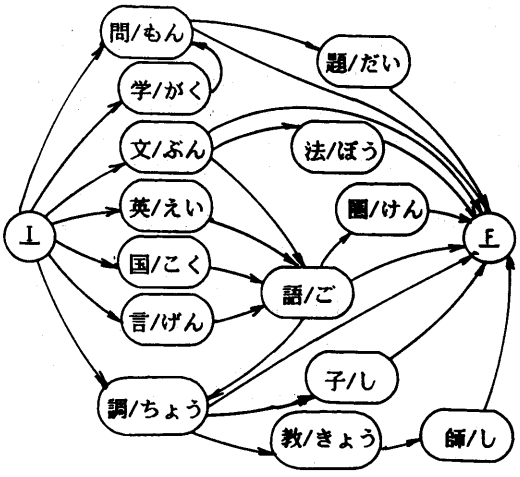

図６　単純マルコフモデル例

一方ラベル付きマルコフモデルでは，ラベル付き造語単位集合U＝｛「文/ぶん」,「国/こく,語/ご」,「文/ぶん,法/ぼう」,「学/がく,問/もん」,「言/げん,「問/もん,題/だい」,「調/ちょう,子/し」,「英/えい,語/ご,圏/けん」,調/ちょう」,教/きょう,師/し」｝（18造語単位）を得る．従って，図7のようなモデル（状態数20）を構成する．

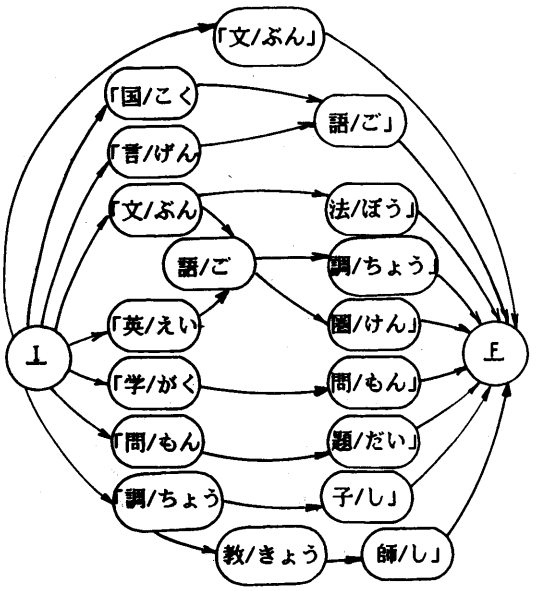

図7　ラベル付きマルコフモデル例

単純マルコフモデル，ラベル付きマルコフモデル共に単語集合Wの単語は全て造語できる．単純マルコフモデルでは，Wに属する単語の他に，１７語の未知語(問,学問題,文語,文語圏,文語調子,文語調教師,英語,英語調,英語調子,英語調教師,国語圏,国語調,国語調教師,言語圏,言語調,言語調教師,調:読み省略)を造語するが，ラベル付きマルコフモデルでは，２語の未知語(文語圏,英語調:読み省略)を造語するにとどまっている．

このラベル付けによって，２造語単位以下の単語の存在に起因する未知語の発生はなくなり，国語辞書に載っている単語のほとんどが３造語単位以内で構成されていることを鑑みれば，造語能力をかなり抑制することができる．一方，モデルの遷移の状態数は，単純な１重マルコフモデルに比べて高々４倍であり実現化に支障はない．

以下単に造語単位といった場合は，このラベル付きの造語単位を表すものとする．

５．遷移確率と単語の生起確率の推定

４節では，造語単位を状態とする状態間の遷移で単語を表現した造語モデルを示した．一般に，品詞Hの語幹MWが$\alpha_1\alpha_2\cdots\alpha_n$の造語単位列からなる単語の生起確率p(MW | H)は，次の式によって計算できる．

$$p(MW \mid H) = p(\alpha_1 \mid \underline{\perp}) \cdot p(\alpha_2 \mid \alpha_1) \cdot \cdots \cdot p(\alpha_n \mid \alpha_{n-1}) \cdot p(\underline{F} \mid \alpha_n)$$

ただし，p(β | α)は品詞Hにおける造語単位α,β間の遷移確率の値である．たとえば，単語(名詞)"計算機(けいさんき)"は，

⊥ →「計/けい → 算/さん → 機/き」→ F

のように状態遷移して生成されると考える．したがって，単語の生起確率は，次に示すようなそれらの状態間の遷移確率の積の形で表される．

p("計算機(けいさんき)" | 名詞)
＝p(「計/けい | ⊥)・p(算/さん |「計/けい)
・p(機/き」| 算/さん)・p(F | 機/き」)

全ての単語とそれらの単語の正確な生起確率が与えられていれば，遷移確率は計算によって求めることができる．逆に，遷移確率が与えられれば，単語の生起確率の計算は可能である．また，モデルが完全なものであれば，単語の生起確率と遷移確率の計算を繰り返しても，それらの値は変化しないはずである．しかし，そのどちらも与えられていないので，次のような仮定をおき，遷移確率の推定を行う．

【仮定３】　国語辞書の見出し
国語辞書の見出しには生起確率の高い単語が載っており，載っていない単語は確率が無視しうるほど小さい．　■

仮定３により，遷移確率の推定は，国語辞書の見出しを用いて次のような手順で行っている．

(1) 各見出し語に適当な初期確率を付与
(2) 繰り返し計算による確率値の更新
　(2-1) 見出し語を重み付きの状態（造語単位）間の遷移に分解
　(2-2) 遷移確率の集計，正規化
　(2-3) 見出し語の生起確率の計算

国語辞書の見出し語集合W，Wより抽出した造語単位集合Uにおいて，見出し語w（∈W）が造語単位α_1，α_2,$\cdots$,α_nからなり，初期確率p_wが与えられているとする．前述のとおり，このモデルでは，モデル上の状態は造語単位に等しいので，以下の説明では「モデルの状態」の代わりに「造語単位」を用いる．

・(2-1)では，見出し語wを造語単位間の遷移に分解し，それぞれの遷移関係に見出し語の確率p_wの重みを持たせる．

$$(\underline{\perp}\rightarrow\alpha_1 : p_w),(\alpha_1\rightarrow\alpha_2 : p_w),\cdots$$
$$\cdots,(\alpha_{n-1}\rightarrow\alpha_n : p_w),(\alpha_n\rightarrow\underline{F} : p_w)$$

・(2-2)では，同じ遷移関係を集計し，

$$(\alpha\rightarrow\beta : q_1),(\alpha\rightarrow\beta : q_2),\cdots$$
$$\Longrightarrow\ (\alpha\rightarrow\beta : q(\alpha\rightarrow\beta))$$
ただし，$q(\alpha\rightarrow\beta)=q_1+q_2+\cdots$

遷移前の造語単位が等しいものについて正規化して，遷移確率を求める．

$$p(\beta \mid \alpha) = q(\alpha\rightarrow\beta) \Big/ \sum_{\alpha'\in U} q(\alpha\rightarrow\alpha')$$

・(2-3)では，(2-2)で計算した造語単位間の遷移確率によって見出し語の生起確率を再計算し，更新する．

$$p_w = p(\alpha_1 \mid \underline{\perp})\cdot p(\alpha_2 \mid \alpha_1)\cdot \cdots$$
$$\cdot p(\alpha_n \mid \alpha_{n-1})\cdot p(\underline{F} \mid \alpha_n)$$

この手順(2)を繰り返して，見出し語の生起確率および造語単位間の遷移確率を集束させていくのだが，集束値は(1)で与えた見出し語の初期生起確率に依存する．そのため適当な確率を付与する必要がある．

4節で用いた少数単語世界の例に，この手順を適用して得られた結果を表1に示す．ただし，初期確率は全て等確率0.1で与えるものとする．

計算回数	初期値	1	2	3	4
文	0.100	0.100	0.100	0.100	0.100
国語	0.100	0.100	0.100	0.100	0.100
文法	0.100	0.100	0.100	0.100	0.100
学問	0.100	0.100	0.100	0.100	0.100
言語	0.100	0.100	0.100	0.100	0.100
問題	0.100	0.100	0.100	0.100	0.100
調子	0.100	0.100	0.100	0.100	0.100
英語圏	0.100	0.050	0.025	0.013	0.006
文語調	0.100	0.050	0.025	0.013	0.006
調教師	0.100	0.100	0.100	0.100	0.100

表1　ラベル付きマルコフモデル計算例

6．単語と複合語

4節，5節で単語の生起確率を求める方法を示した．本節では，単語(名詞)の連接によって発生した複合語の生起確率について考察する．

単語(名詞)$W_1,W_2,\cdots,W_n$が存在し，それぞれの綴りを$w_1,w_2,\cdots,w_n$，生起確率を$P_w(w_1),P_w(w_2),\cdots,P_w(w_n)$とするとき，綴り$w_1+w_2+\cdots+w_n$からなる複合語の生起確率$P_c$を次式で考える．

$$P_c(w_1w_2\cdots w_n)$$
$$= P_w(w_1)\cdot P_w(W_2 \mid W_1)\cdot P_w(w_2)\cdot\cdots$$
$$\cdot P_w(W_n \mid W_{n-1})\cdot P_w(w_n)$$
$$\fallingdotseq P_w(w_1)\cdot P_w(w_2)\cdot\cdots\cdot P_w(w_n)\cdot P(N \mid N)^{n-1}$$

ここで，$P(N \mid N)$は名詞-名詞間の連接確率である．$P(N \mid N)$は，綴りwにおいて単語である場合には，$P_w(w)>P_c(w)$，複合語である場合には，$P_w(w)<P_c(w)$となるように決定する．例えば，「資本主義」という綴りを単語と認定する場合には，

$$P_w(\text{資本主義})>P_w(\text{資本})\cdot P_w(\text{主義})\cdot P(N \mid N)$$

となるように$P(N \mid N)$を決定する．すなわち，

$$P(N \mid N)< \frac{P_w(\text{資本主義})}{P_w(\text{資本})\cdot P_w(\text{主義})}$$

の値の範囲で$P(N \mid N)$を設定する．

7．実験と考察

辞書中より名詞74,454語について造語単位の抽出，単語の生起確率の推定実験を行った．

構成造語単位数	語数
1	9，203
2	43，729
3	20，293
4	1，184
5	43
6	2
合計	74，454

表2　単語分布

単純マルコフモデル

$\underline{\perp}\rightarrow\alpha$	13798
$\alpha\rightarrow\underline{F}$	75762
合計	89560

ラベル付きマルコフモデル

$\underline{\perp}\rightarrow$「$\alpha$」	9203
$\underline{\perp}\rightarrow$「$\alpha$	7034
「$\alpha\rightarrow\alpha$」	43729
「$\alpha\rightarrow\alpha$	13058
$\alpha\rightarrow\alpha$」	15886
$\alpha\rightarrow\alpha$	1009
合計	89919

表3　造語単位間遷移分布

実験に用いた辞書の内容を表2に示す．この辞書より16,771種類の造語単位を抽出した．造語単位間の遷移の種類(遷移確率が0でないもの)は，単純マルコフモデル，ラベル付きマルコフモデルそれぞれ表3に示すとおりである．遷移の種類は，元の単語数の2割増し程度であり，繰り返し計算も支障なく行えた．

５節の表１に示したように繰り返し計算を行っても未知語を発生しない部分の相対的な確率は変化しない．したがって，ラベル付きマルコフモデルでは，２造語単位以下からなる単語の部分については，相対的には初期確率で与えられた値をそのまま反映し，繰り返し計算の効果が現れない．故に，単純マルコフモデルによる繰り返し計算(初期確率は等確率)を行った結果を初期値として採用し，ラベル付きマルコフモデルによる繰り返し計算を行っている．

次に結果に対する考察を述べる．個々の単語の生起確率，あるいは造語単位間の遷移確率について妥当であるかどうかを調べる手段はない．複数の読み方を持つ単語間の相対的確率について調べてみると，不適当なものがある．それらは２つのグループに分けられる．

(1)辞書に存在する単語間での
　相対的確率が不適当であるもの
　言語：　①ごんご　　②げんご
　音声：　①おんじょう　②おんせい
　化学：　①ばけがく　　②かがく

(2)辞書に存在しない未知語の確率が
　大きくなっているもの
　高調波：　たかちょうは

グループ(1)については，初期値を加減することによって対処できると考える．しかし，グループ(2)については，現在のところ良い解決策がない．

８．おわりに

今回，繰り返し計算によって，一応単語の生起確率の推定値を得たが，まだ不十分な部分が多くある．繰り返し計算の結果は，初期値に依存し，よりよい結果を得るためには，良い値を初期値として与えなければならない．従って，ある程度人手による修正も必要となる．

造語モデルに読みの概念を導入したことにより，漢字のルビ振りやかな漢字変換等への応用も可能となった．しかし実用のシステムを構成するためには，辞書に登録されていない未知語や接辞等の処理も考慮せねばならず，今後に課題を残している．

参考文献

・松延,日高,吉田:"日本語確率文法における書き換え規則の確率の推定について" 情報処理自然言語研究会研資 56-3,1986