日本語における単語の造語モデルとその評価

永井秀利† 日高達††

べた書き日本語文の形態素解析では多くの曖昧さと未登録語の存在が大きな問題となる．これらの問題に対処するものとして，本論文では日本語における単語の造語モデルを示す．このモデルは，日本語における単語の造語は漢字の意味的，音韻的結合によってなされると仮定し，確率モデル化を行ったものである．曖昧さの絞り込みを行う方法の一つとしては，単語の生起確率の利用が考えられるが，分野を限定せずに単語の生起確率を統計的に獲得することは極めて困難である．造語モデルでは通常の統計的手法によらず，単語辞書の見出しから機械的処理により漢字の結合傾向をとらえ，これに基づいて単語の生起確率を推定する．統計的情報を有する場合にはそれをモデルに反映させることも可能である．造語モデルの利用法は生起確率推定だけにとどまらない．このモデルでは未知語の造語も可能であり，造語できる語は単語らしいと見なすことができる．これは，推定された生起確率を文字列の単語らしさの評価値として利用できることを示している．造語モデルによって推定された生起確率が妥当なものであるかの評価は非常に困難である．本論文ではモデルに求められる性質をいくつかあげ，それに基づいてモデルの評価を試みる．単語辞書の見出しという乏しい情報から出発したにしては良好な結果を得ている．

Japanese Word Formation Model and Its Evaluation

Hidetoshi Nagai† and Toru Hitaka††

We propose Japanese Word Formation Model. This model is a stochastic model of Japanese word formation, based on tendencies in semantic and phonologic combinations of Chinese characters. To construct the model, we don't need statistical data, but use entries of Japanese word dictionary. We obtain the tendencies from the entries. If we have statistical data, we can use the data to reflect on the model. Japanese Word Formation Model does not only estimate the probabilities of word occurrence which are useful in disambiguation, but difficult for statistical acquisition. The model can generate unknown words. We can be regard these unknown words as being word-like both semantically and phonologically. Therefore we can use the probabilities estimated by the model to estimate how word-like the string is. It is very difficult to judge whether the probability of each word estimated by Japanese Word Formation Model has adequate accuracy. So we propose and evaluate some properties which the model should have, and obtain satisfactory results. Furthermore we evaluate whether the model is able to reflect importance rank of each registered word, which is wrote on a word dictionary, as rough statistical data. This experimentation also gives us satisfactory results.

1. はじめに

形態素解析はさまざまな自然言語処理の最初のステップであり，ここで生じた曖昧さは後の処理に大きな影響を及ぼす．膠着語である日本語では，文は一般に語と語の間に空白を入れずに書かれるが，このようなべた書きされた日本語文の形態素解析においては，英語などのように単語単位の分かち書きがなされる言語とは異なり，その曖昧さは品詞的なものにとどまらない．どの部分列が単語を構成しているかの決定においても極めて多くの曖昧さを生じる．

曖昧さへの対処方法の一つとしては，単語の生起確率を利用して，解析結果に対する優先順位付けを行うことが考えられる．ところがこの単語の生起確率を統計的に得ることは極めて困難である．したがって何らかの妥当性のある推定法が必要とされる．

べた書き日本語文の形態素解析における曖昧さの問題は，解析対象文中に単語辞書未登録の語が存在した場合，より顕著になる．しかも分野を極めて限定しない限り，文中に出現するすべての単語を辞書に登録しておくことは困難であり，未登録語処理は避けて通れない．英語のように単語単位に分かち書きされる場合は未登録語の存在自体が形態素解析にもたらす困難

† 九州工業大学情報工学部知能情報工学科
Department of Artificial Intelligence, Kyushu Institute of Technology
†† 九州大学工学部情報工学科
Department of Computer Science and Communication Engineering, Kyushu University

さはそう大きなものではないが，日本語のように分かち書きをしない言語の場合，その影響は極めて大きい．

従来，未登録語に対しては，例えば次に述べるような処理が用いられていた．

- 字種の利用（平仮名→漢字のように変化する部分は単語の切れ目である可能性が高いなど）．
- 付属語(列)などにより文節を推定し，そこから付属語(列)を取り除いた部分列を未登録語と見なす．

しかし，これらの手法は未登録語と見なす文字列の単語らしさを十分に評価しておらず，不十分なものといわざるをえない．前者は字種情報に基づき単語らしさを評価しているともいえるが，字種だけでは評価能力が非常に低い．未登録語自体の単語らしさを評価する試みを行っている研究[6),7)]も見受けられるが，対象が限定されていたり，記述に極めて多量の労力を必要としたりするなど，まだ十分に確立されているとはいえない．形態素解析と構文解析，意味解析を融合したシステムで，未登録語を含む文の解析を行う研究[8)~10)]もあるが，未登録語と見なした文字列自体の単語らしさを扱うには至っていない．そこで，単語らしさ評価を行うことができ，かつ実現性の高い処理法が望まれる．

一般に未登録語処理は，登録語だけから構成される構造での解析に失敗して初めて起動される．このことは通常は全く問題にならないが，新しい概念を記述した文章のように未登録語の出現可能性が高い場合には，未登録語処理の起動までに非常に多くの処理量を必要とすることが少々問題となってくる．特に表記の揺れを許す場合のように解析対象文の記述の自由度が上がれば登録語だけで構成できる構造は非常に多くなり，極めて多くの処理量を必要とする．そのため，登録語の可能性と未登録語の可能性とを同時に比較，検討することができればより望ましいといえよう．

ところで日本語は漢字を用いる言語である．漢字は表意文字であるため，漢字の結合は意味的結合を表し，この結合により単語の意味が構成されると考えることができる．言語の伝達には音声も用いられるため，結合に際しては発音しやすいように読みが選択されたり，音の変化を生じたりすると思われる．また，漢字がどの意味で用いられているかに依存した読みの違いも考えられるであろう．現実に使用されている単語には，これらの意味的および音韻的結合に関して何らかの傾向が存在すると思われる．単語辞書には生起確率の高い語や重要語が登録されており，それゆえこの結合の傾向は，単語辞書に登録されている単語から得ることができると考えられる．

造語モデルはこの結合傾向に基づき，単語の造語を確率モデルとして構成したものである．実現性を高めるために，意味的，音韻的結合に関する単語の詳細な分析を行うことを避け，極力人手を必要とせずに機械的処理で構築できるようなモデル化を行っている．造語モデルは単語の生起確率を推定すると同時に，漢字の意味的，音韻的結合傾向に基づいた文字列の単語らしさ評価にも利用できる．未登録語の単語らしさの評価という点では，その未登録語の文字列自体を評価する点で，従来の手法よりも妥当性が高いと思われる．このモデルを通常の単語辞書検索の代用として解析処理に組み入れた場合，登録語の可能性と未登録語の可能性とを同時に検索できるという利点を持つ．

造語モデルはほかにも利用可能な特徴を持ち，これを利用した研究[11)~14)]も行われている．しかし，未登録語評価値としての有効性を含めて，造語モデルが出力する確率が妥当なものであるかを評価することは困難である．そこで本論文では，モデルが持つべき性質を想定し，これを満足するかどうかについて実験を行った．また，実際の解析において問題となりやすい同表記語についても，登録語，未登録語を含めて，造語実験により確率大小関係の調査を行った．さらに，単語辞書中に含まれる重要語に関する情報を活用することでモデルにより推定される確率をより妥当なものにする試みについて，実験および考察を行っている．本論文では，日本語における単語の造語モデルについて示すと同時に，これら実験結果について報告する．

2. 単語の造語モデル

本章では，日本語における単語の造語モデルについて，その基本概念や確率値推定法を述べる．

2.1 造語単位

本論文では，日本語において一般に用いられているものを‘単語’とし，そのうち単語辞書に登録されているものを‘登録語’，登録されていないものを‘未登録語’と呼ぶ．また造語モデルによって造語されるものを‘語’と呼び，そのうち登録語ではないものを‘未知語’と呼ぶ*．

* いかなるものが‘単語’であるのかは難しい問題である．ここではこれについては議論しない．造語モデルを作成する際には，単語辞書に登録されているものはすべて‘単語’として扱う．

1章で述べたように，日本語における単語は，基本的には表意文字である漢字の意味的，音韻的結合によって構成されると考えることができる．そこで日本語における単語の造語に関して，次の仮定を置く．

【仮定1】 日本語における単語の造語は意味的，音韻的最小単位が結合することによって行われる．

この意味的，音韻的最小単位を‘造語単位’と呼ぶ．すなわち，単語は造語単位の意味的，音韻的結合によって構成されると仮定する．造語単位には基本的には表意文字である漢字1文字が相当するが，日本語では平仮名などの表音文字も使用される．表音文字はそれ1文字では音を表すだけである．また，いわゆる当て字に関しては漢字単位での音の分離は不可能である．これらを考慮して，造語単位を次のように定義する．

【定義1】 綴りと読みとの組から構成される次のものを造語単位とする．

(1) 単語中において読みの最小単位を持つ漢字列とそれに続く一連の平仮名列

(2) (1)以外で単語中に存在する一連の同一文字種の非漢字列

上記定義(1)において，漢字に続く平仮名列までを含めて造語単位としたのは，漢字の送り仮名を考慮したためである．送り仮名は独立した意味を表さず，音韻を整えるだけであるため，意味的にも音韻的にも漢字と切り離して考えるべきではない．この定義は単語の詳細な分析に基づくものではないため，不適切な造語単位を切り出す可能性もあるが，このように定義することにより，単語の漢字表記と読みとの対応情報から，機械的に造語単位を切り出すことが可能となる．上記定義に基づく造語単位の例を次に示す．

【造語単位の例 (〔綴り〕/〔読み〕)**】**

国/こく 文/ぶん 法/ぽう：国文法

梅雨/つゆ 入り/いり ：梅雨入り

アーク/あーく 灯/とう ：アーク灯

日本語の単語として用いられるものの中には，複合語の省略語として，複合語を構成する各単語の一部の文字をつなぎあわせて造語されているものもある．このようにして造語される際に選び出される文字が，各単語を意味的に代表するものであれば，このモデルで扱うことも可能である．しかし実際には，単なる先頭文字や，第1アクセントを有する文字など，必ずしも各単語を意味的に代表する文字が選ばれるわけではない．このような造語はここで扱うものは異なる造語であるため，本論文におけるモデルの対象外とする．同様に人名や地名，片仮名表記される外来語に関しても対象外とする．このような単語に対しては全く異なる造語のモデル化が必要となるが，この種の造語を形式的に扱うのは困難であると思われる．

2.2 造語過程

造語モデルを構築するにあたり，単語の造語に関して次の仮定をおく．

【仮定2】 造語単位の結合による単語の造語過程はマルコフ過程である．

造語過程のモデル化の場合，状態を定める基礎となる造語単位の種類が多いため，2重以上のモデルでは状態数が膨大となり，実用的ではない．そこで造語モデルでは，造語単位を一つの内部状態とする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だけを考慮したモデルといえる．

音韻的な面から見た場合，常に直前，直後のつながりだけが影響を及ぼしあう．そのため音韻面での結合傾向は，1重マルコフモデルで十分にとらえることができる．意味的な面から見た場合，2造語単位以下から構成される単語については結合が存在しないか，あるいは隣接する関係しか存在せず，1重マルコフモデルで妥当である．問題となるのは，3造語単位以上から構成される単語である．

例えば“最大値”(最/さい＋大/だい＋値/ち)という単語のように造語単位が左から順に結合し，部分構造の最右端の造語単位がその部分構造を意味的に代表しているような構造(**図1**)であれば，1重マルコフモデルでとらえることができる．しかし，“高品質”(高/こう＋品/ひん＋質/しつ)という単語のように意味的に結合しているものが隣接関係を持たない構造(**図2**)の場合はうまく扱うことができない．

日本語の単語の場合，後方に位置する漢字(造語単位)が単語の意味を代表することが多い．そのような

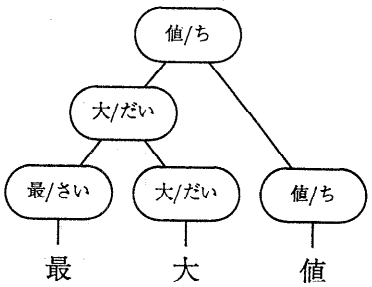

図1 1重マルコフモデルでうまく扱える構造
Fig. 1 Word modeled by Markov model.

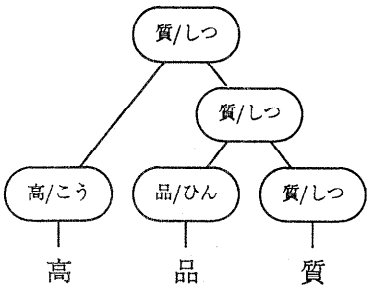

図 2 1重マルコフモデルではうまく扱えない構造
Fig. 2 Word not modeled by Markov model.

単語は図1のような構造を持ち，1重マルコフモデルでうまく扱うことができる．

このように3造語単位以上から構成される単語には1重マルコフモデルではうまく扱えないものも存在する．これに対し，単語の構造を木構造で見てきたことから見てもわかるように，造語を CFG で見ればより精密にとらえられるであろうと考えるのは当然である．しかし何万もの単語に対してその構造を分析するには非常に多くの労力を必要とする．しかも構造の決定が困難であるような単語も多い．そのため，実現性が非常に低くなる．

日本語の単語は2造語単位以下から構成されるものが多く（実験で用いた単語辞書では約68%が2造語単位以下），4造語単位以上から構成される単語は極端に数が少なくなる（実験で用いた単語辞書では約2%）．そこで，うまく扱うことができない構造も存在するが，数量的に見てほぼ妥当なモデルであると考えて，1重マルコフモデルを採用する．

2.3 造語単位の細分化

単純な1重マルコフモデルによるモデル化では造語能力（モデルが造語できる語の総数）が大きくなりすぎるという問題がある．すなわち，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，解析のモデルとして用いた場合，長い文字列を単語としてしまう傾向がある．

単語辞書の性質を考えた場合，漢字2文字以下の単語はほとんど網羅されていると考えることができ，モデルの対象外である省略語などを除くと，漢字2文字以下の単語が新たに造語されることは極めてまれである．これを考えると造語モデルでは漢字2文字以下の未知語を造語しない方が望ましいが，単純なモデルではそのような未知語を非常に多く造語する可能性がある．そこで，登録語はすべて造語でき，未知語（特に2造語単位以下）の造語を抑制する機構を導入する．

造語単位の中には，1造語単位で単語を構成するものや，前後のいずれか，あるいは両方でほかの造語単位と結合して単語を構成する傾向を持つものがある．日本語では基本的に前方から修飾を行う構文構造を持つと考えられるが，造語単位の意味的結合においてもこの傾向は強く，造語単位が結合する場合，前方に位置する造語単位が後方に位置する造語単位を修飾し，後方に位置する造語単位が全体を意味的に代表することが多い．したがって，造語単位の位置的傾向がその機能的傾向をも表すと考えることができる．

音韻的に見た場合，前方にほかの造語単位を必要とするものは，連濁のように音韻変化を生じているものもあり，独立性が少々弱い．逆に後方にほかの造語単位を必要とするものは，音韻的終結性が弱い．

この造語単位の位置的傾向（接続傾向）に着目し，造語単位を次の四つの型に細分する．

Type 1 「α」: 前後ともほかの造語単位と結合しない
Type 2 「α 」: 後方にほかの造語単位を要する
Type 3 α」: 前方にほかの造語単位を要する
Type 4 α : 前後共にほかの造語単位を要する

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

少数単語世界において造語モデルの例を示すと同時に，位置ラベルによる造語能力抑制の効果をみる．

単語集合 Ws= {文/ぶん，文法/ぶんぽう，学問/がくもん，国語/こくご，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}（〔綴り〕/〔読み〕）とする．Ws から構成される単純な1重マルコフモデルは**図3**のように，位置ラベル付きのモデルは**図4**のようになる．ただし _I, _F は，それぞれ初期状態，最終状態を表す．

どちらのモデルも Ws の単語はすべて造語できる．未知語については，図3のモデルが15個｛問，調，文語，英語，文語圏，国語圏，言語圏，英語調，国語

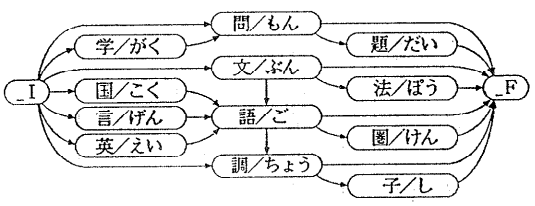

図 3 単純な1重マルコフモデルの例
Fig. 3 An example of simple Markov model.

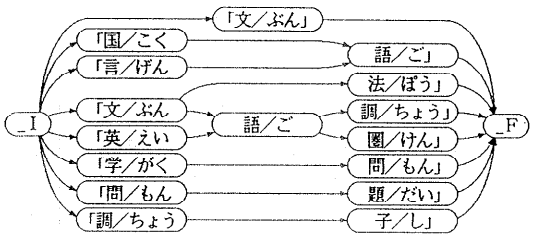

図 4 位置ラベル付き1重マルコフモデルの例
Fig. 4 An example of Markov model with position labels.

調，言語調，学問題，文語調子，英語調子，国語調子，言語調子} を造語するのに対し，図4のモデルは2個 {文語圏，英語調} を造語するにすぎない（読みは省略）．

位置ラベル付きモデルでは2造語単位以下から構成される未知語は造語されなくなる．モデルの状態数は，単純な1重マルコフモデルに比べ高々4倍であり，実現上の支障はない．以下，単に造語単位という場合は，この位置ラベル付けされた造語単位を指すものとする．

2.4 未登録語の単語らしさ評価能力

未登録語の存在が否定されない場合，自然言語処理には非常に多くの困難が伴う．

しかしながら，対象分野を極めて限定しない限り，入力文中に出現するすべての単語を辞書に登録しておくことは非常に困難である．極めて大きな辞書を用いることにより，未登録語の出現をまれなものとするというアプローチもあるが，いかに巨大な辞書を作ろうとも未登録語の存在を完全に否定することはできず，一般的には未登録語処理を避けて通ることはできない．それどころか，新しい概念が記述されたテキストからその概念を学習することを考えた場合，通常は未登録語が存在すると考えるべきであろう．

造語モデルにおいては，単語を構成する造語単位の結合傾向に基づき，その単語の生起確率を推定している．したがって，造語モデルにより造語できる語は単語らしいということができ，ある文字列が語として生起する確率をモデルにより求めれば，その確率値はその文字列の単語らしさの評価値として利用することができる．2章で述べたことからわかるように，意味的，音韻的最小単位である造語単位と表記文字列とは密接な関係を持つ．そのため造語モデルによる評価は，単なる文字の羅列としてではなく，意味的，音韻的に単語らしいかということを評価しているといえる．

従来の方法と比較した場合，文字列の単語らしさを評価している点で，未登録語処理としての妥当性が高いと考えられる．造語モデルの上では登録語であるか否かは造語に際して区別されない．そのため，例えば単語辞書の代用としてモデルを解析システム中に組み込めば，登録語の可能性と同時に未登録語の可能性も調べることができるという利点を持つ．ただし，省略語や人名など，モデルの対象外の未知語に対しては，従来と同様に未登録語処理を別処理として用意せざるをえない．しかし，別処理で扱わねばならない未知語の種類が限られるため，すべての種類の未知語を別処理で行う場合と比べて，負担は軽減されると思われる．

2.5 単語の生起確率の推定

造語モデルによると，単語 w が $\alpha_1\alpha_2\cdots\alpha_n$ という造語単位列から構成されているとき，w の生起確率は

$$P(w)=P(\alpha_1|\_I)\cdot P(\alpha_2|\alpha_1)\cdots P(\alpha_n|\alpha_{n-1})\cdot P(\_F|\alpha_n) \quad (1)$$

で計算される．ただし，$P(\beta|\alpha)$ は造語単位 α で表される状態から造語単位 β で表される状態への遷移確率値であり，_I，_F はそれぞれ初期状態，最終状態を表す．

遷移確率が与えられれば，単語の生起確率は式(1)から容易に計算できる．しかし今は正確な値は得られていない．したがって何らかの方法でこの確率値を推定する必要がある．

ここで，次の仮定を置く．

【仮定3】 単語辞書には生起確率の高い単語が登録されている．

単語辞書は生起確率の情報を持たない．しかし，生起確率は低くても概念的に重要であるために登録された一部の例外を除いては，高い生起確率を持つ単語から優先的に登録されていると考えることができる．

マルコフモデルにおける遷移確率推定法としては反復計算によるものがよく知られている[15),16)]が，この方法ではサンプル集合の生起確率の総和を極大にすることを保証しており，仮定3を満足できる保証はない．そこで，ここではすべての登録語の生起確率の積を最大にすることを目指す．つまり，Dを登録語の集合，$P(w)$ を単語 w の生起確率とするとき，

$$\prod_{w\in D} P(w) \quad (2)$$

を最大にするように，状態遷移確率を定める．

式(1)を利用して式(2)を最大にする状態遷移確率を求めると，

$$P(\beta|\alpha)=\frac{\sum_{w\in D}N(\alpha\to\beta,w)}{\sum_{\gamma\in U}\sum_{w\in D}N(\alpha\to\gamma,w)} \quad (3)$$

となる．ただし，Uは造語単位の全体集合，$N(\alpha\to\beta, w)$ は単語wにおける $\alpha\beta$ という造語単位の連接（α からβへの状態遷移）の出現回数である．

ここで全登録語の生起確率の単純な積を扱ったのは，各登録語の出現頻度に関する情報が全くないためであるが，もし何らかの形で各登録語に対する重み付けの情報を有しているならば，それを反映させる手段を用意しておくべきである．そこで重み付けを行った登録語生起確率の積を最大にすることを考える．ここでは登録語wに与えられる重み付けを $I(w)$ として，

$$\prod_{w\in D}P(w)^{I(w)} \quad (4)$$

を最大にする．これを最大にする各状態遷移確率は，

$$P(\beta|\alpha)=\frac{\sum_{w\in D}I(w)\cdot N(\alpha\to\beta,w)}{\sum_{\gamma\in U}\sum_{w\in D}I(w)\cdot N(\alpha\to\gamma,w)} \quad (5)$$

によって与えられる．

ところが位置ラベル付きモデルにおいてこの計算を行った場合，1ないし2造語単位から構成される語の間での生起確率の比率は，初期値として与えた比率（式(3)では同一，式(4)では重み付けの比率）から変化しないという問題が生じる☆．

そこでまず式(3)または(4)により位置ラベルなしの単純なモデルを構築し，これから得られた各登録語の生起確率値を重み付けとして，式(4)により位置ラベル付き造語モデルを構築する．

3. 評価実験とその結果

2.2 節で造語モデルを1重マルコフモデルでモデル化することの妥当性については述べた．しかし，任意に選出した二つの単語のどちらが頻出語かを決定することが非常に困難であることからもわかるように，造語モデルから得られた個々の単語の生起確率が妥当なものであるかを客観的に評価することは極めて難しい．そこで今回の実験では造語モデルに期待される性質を上げ，これを基準に評価を試みる．

3.1 造語モデルの数値的データ

今回は，九州芸工大と九大で開発された日本語単語機械辞書[18]から抽出し，若干の手を加えた名詞79,026語に基づき，モデルの作成および評価実験を行った．**表1**に構成造語単位数ごとの登録語数，**表2**にタイプごとの造語単位数，**表3**にモデルの状態遷移数を示す．

表 1 構成造語単位毎の登録語数
Table 1 Distribution of registered words.

構成造語単位数	登録語数
1	9,157
2	44,476
3	23,789
4	1,543
5	59
6	2

表 2 造語単位数
Table 2 Distribution of word formative unit.

ラベルなし	16,325
「α」	9,157
「α	7,240
α」	5,661
α	3,434
ラベル付き総計	25,492

表 3 造語単位間遷移数
Table 3 Distribution of state transitions.

遷移種別	遷移数
_I → 「α」	9,157
_I → 「α	7,240
「α → β	14,492
「α → β」	44,476
α → β	1,262
α → β」	17,733
小計：	94,360
「α」→ _F	9,157
α」→ _F	5,662
総計：	109,179

_I：初期状態，_F：最終状態．
「α」，β」など：各位置ラベルタイプの造語単位で表される状態．

位置ラベル付きモデルでの状態遷移の総数は109,179である．ただし，'」'の位置ラベルを持つ造語単位からの遷移先は最終状態しか存在しないため，実質的には α」→_F や 「α」→_F といった遷移は無視することができる．これらを除いた場合の遷移数は94,360である．これは登録語数に対し 19.4% の増加にすぎず，実現には支障がないことがわかる．

モデルの造語能力を見るために10造語単位から構

☆ これは以下の関係式から証明できる．
ただし $w_1=$「α_1」，$w_2=$「$\alpha_2\beta_2$」という造語単位構成を持つ単語とする．
- $\sum_{w\in D}I(w)\cdot N(\_I\to$「$\alpha_1$」$, w)=I(w_1)$
- $\sum_{w\in D}I(w)\cdot N($「$\alpha_2\to\beta_2$」$, w)=I(w_2)$
- $\sum_{w\in D}I(w)\cdot N(\_I\to$「$\alpha_2, w)$
$=\sum_{\gamma\in U}\sum_{w\in D}I(w)\cdot N($「$\alpha_2\to\gamma, w)$

表 4 造語モデルの造語総数
Table 4 Word-forming ability of word model.

構成造語単位数	ラベルなし		ラベル付き		登録語
	登録語	未知語	登録語	未知語	
1	9,157	2,139	9,157	0	9,157
2	44,476	6,436	44,476	0	44,476
3	23,789	1,514,117	23,789	202,235	23,789
4	1,543	49,155,539	1,543	293,777	1,543
5	59	1,634,781,677	59	406,694	59
6	2	5.51081×10^{10}	2	598,791	2
7	0	1.86764×10^{12}	0	1,016,139	0
8	0	6.34222×10^{13}	0	1,957,837	0
9	0	2.15544×10^{15}	0	3,765,738	0
10	0	7.32764×10^{16}	0	7,215,028	0

成される語まで造語させてみた結果を**表 4** に示す．

登録語はすべて造語でき，特に位置ラベル付きモデルでは構成造語単位数が1ないし2の未知語は造語しない．この結果から，位置ラベルにより未知語の造語が非常に抑えられていることや，位置ラベルを用いてもまだかなりの数の未知語を造語することがわかる．

3.2 造語モデルの全体的性質評価

造語モデルの全体的性質として要請されるものには，以下のものが上げられる．

(1) 登録語はすべて造語できる．
(2) 2造語単位以下からなる未知語は造語されない．
(3) 未知語の生起確率は登録語よりもおおむね低い．
(4) 長い単語の生起確率は小さい．

これまでに述べたように，位置ラベル付き造語モデルでは上記要請(1)，(2)は満足される．そこで要請(3)，(4)について見るために，生起確率の高い順に200万語を造語させた際の結果を**図 5～7** に示す．

図5に生起確率の高い順にある順位の語が造語されるまでに，登録語のうち何%が造語されるかを構成造語単位数ごとに示している．これによれば累積50万語までにはほとんどの登録語（特に3造語単位以下から構成される登録語についてはほぼすべて）が造語され，登録語については要請(4)の傾向を持つことがわかる．

図6は造語された登録語，未知語の分布を生起確率値でみたものであり，10^{x+1}～10^{x} の範囲にそれぞれ何語が存在しているかを示す．これによりモデルが要請(3)の傾向を持つことがわかる．

図7は確率の高い順に1万語単位で，3～8造語単位からなる未知語が各何語含まれるかを示す．これにより未知語に関しても要請(4)の傾向を持つといえる．

以上により，造語モデルは要請(1)～(4)を満足し

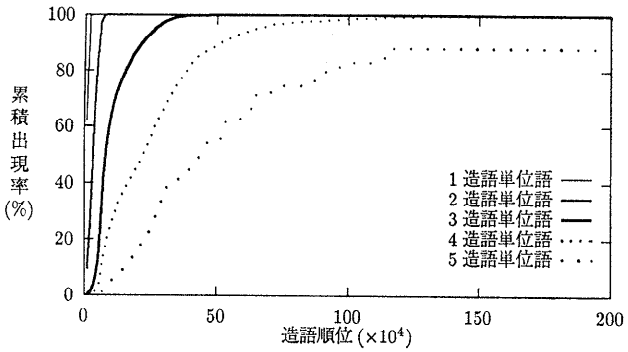

図 5 構成造語単位数別登録語累積出現率
Fig. 5 Cumulative frequency of registered words classified by the number of formative units.

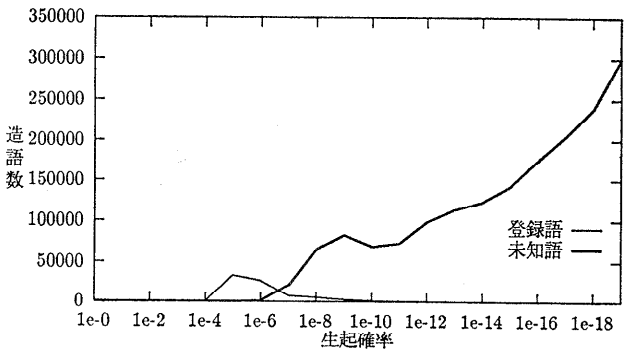

図 6 登録語・未知語分布
Fig. 6 Frequency distribution of generated words.

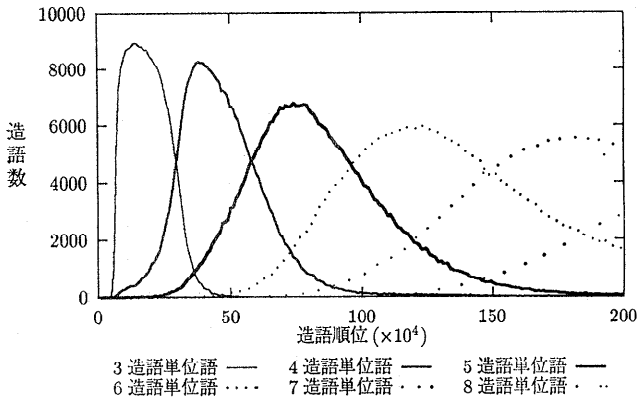

図 7 構成造語単位数別未知語分布
Fig. 7 Frequency of unregistered generated words classified by the number of formative units.

表 5 登録語カテゴリー
Table 5 Distribution of the rank of importance.

構成造語単位数	ランク			
	0	1	2	3
1	5,940	2,588	425	204
2	28,848	14,076	1,149	403
3	22,105	1,658	26	0
4	1,502	41	0	0
5	57	2	0	0
6	2	0	0	0
合 計	58,454	18,365	1,600	607

ていると考えることができる.

次に登録語の初期確率に重み付けを行った場合の効果について見る.

重み付けの情報としては次のものを用いた.

(1) 造語モデル作成の原データとなった日本語単語機械辞書が持つ重要語情報

(2) 新明解国語辞典[19]の重要語および最重要語情報

これにより登録語を重要語ランク 0～3 の 4 カテゴリーに分けた. ランク 0 が非重要語, ランク 3 が最重要語である. 各カテゴリーに属する登録語数を**表 5** に示す.

この分類に基づき, 重み付けをした初期確率値を次のように与えた.

(1) 同一カテゴリーに属する登録語は等確率.

(2) カテゴリーごとの登録語生起確率の総和はすべてのカテゴリーで等しい.

これから構築された造語モデルについて, 要請(1)～(4)を満足するかを調べるために, 図 5～図 7 と同様のグラフを作成した. 結果は, 重み付けを与えない場合と比べて, グラフ上での顕著な差は見いだせなかった. そのため紙面の都合上省略する. すなわち, 重み付けを行っても, 登録語と未知語との関係に関する要請(1)～(4)を傾向として満足している.

次に各カテゴリーごとにモデルの性質を見る.

カテゴリーについて考えた場合, カテゴリー内においても要請(4)を満たすことが望まれるのに加えて,

(5) 高い初期確率を与えられたカテゴリーの登録語は低い初期確率を与えられたカテゴリーの登録語よりも高い確率を持つ傾向にある.

という性質が要請される.

カテゴリー間の完全な順序関係を要請しないのは, カテゴリーの境界は曖昧であり, カテゴリー分けが必ずしも正確なものではないと考えられるためである.

重み付けを与えた場合の, 造語実験の結果を**図 8～図 12** に示す. これらは生起確率の高い順にある順位の語が造語されるまでに各カテゴリーごと, あるいは各カテゴリーに属する登録語の構成造語単位数ごとにその何%が造語されたかを示している. **図 13** は, 比較のために重み付けを与えない場合について図 8 と同様のグラフを作成したものである.

これらのグラフから以下のことを見ることができる.

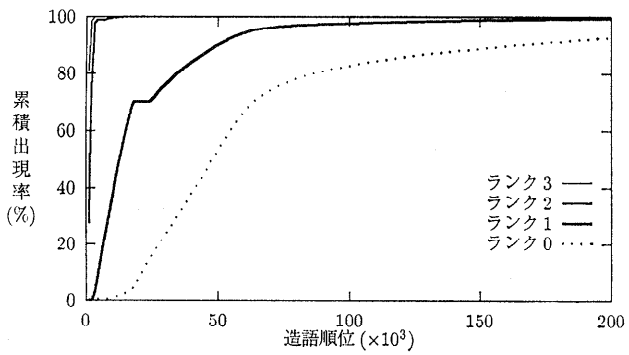

図 8 重要語ランク別出現率
Fig. 8 Cumulative frequency of words classified by the rank of importance.

(1) 重み付けの有無にかかわらず，各カテゴリー内においても要請(4)の傾向を持つ．

(2) 重み付けを行った場合，要請(5)の傾向を持つ．

(3) 重み付けを与えなくとも，重要語の方が高い生起確率を持つ傾向がわずかながら見受けられる．

これらの結果から，重要語情報などの情報も，重み付けとして初期確率に反映させることにより，造語モデルで十分に活用可能であるということができる．特に上記(3)はモデルの妥当性を補強する結果といえよう．

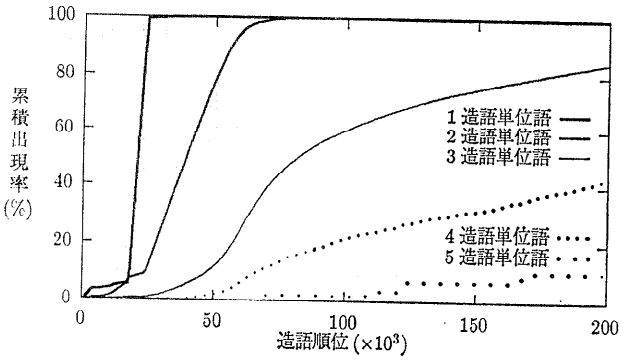

図 9 構成造語単位数別累積出現率（重要語ランク0）
Fig. 9 Cumulative frequency of importance rank 0 words classified by the number of formative units.

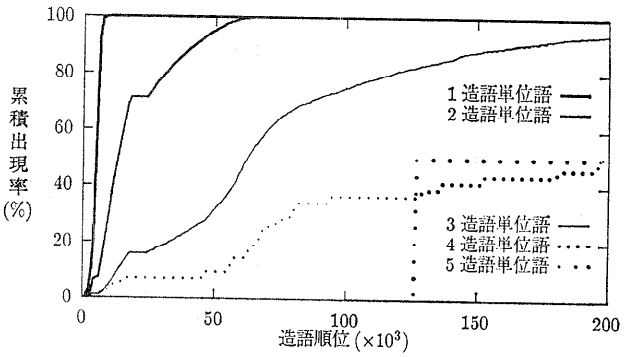

図 10 構成造語単位数別累積出現率（重要語ランク1）
Fig. 10 Cumulative frequency of importance rank 1 words classified by the number of formative units.

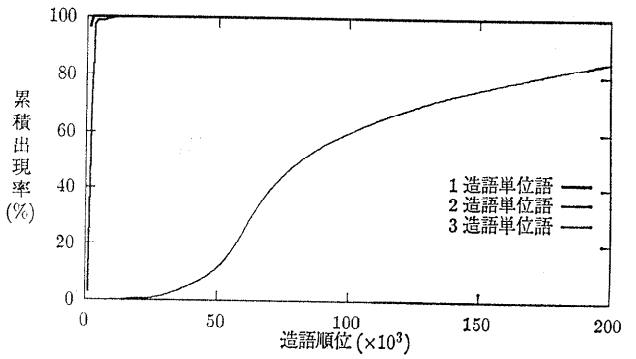

図 11 構成造語単位数別累積出現率（重要語ランク2）
Fig. 11 Cumulative frequency of importance rank 2 words classified by the number of formative units.

3.3 造語モデルの部分的性質評価

原データとなった単語辞書では，同音語についての優先順位が与えられている．その信頼性については情報がなく，どこまで信頼してよいのか判断できないという問題もあるが（事実，個人的には納得しかねる優先順位が与えられているものもある），造語モデル評価の参考にはなるであろう．そこで同音語について確率の大小関係を調査したものが**表6**である．これは二つの同音語のすべての組について，確率大小関係が日本語単語機械辞書の優先順位と一致しているかを調べたものである．この際，造語モデルにおいて等確率となってしまったものも誤りと見なしている．

調査結果では重み付けの効果が明確に現れている．重み付けだけで正答率が定まっているのではないことは，重み付けにより与えられた初期確率値での調査における正答率との比較から理解することができる．

次に，登録語の取りうる全表記を生成し，同一表記となるものについて生起確率を比較した．この場合，優先順位情報が存在しないため，3.2 節で述べたカテゴリー間での比較を行った．結果を**表7**に示す．

重み付けなしの場合に正答率が低いのは，もともと情報がないため，ある程度はやむをえないといえよう．その代わり，重み付けの情報が与えられた場合はかなり高い正答率となっている．

これらの結果は，重み付けの情報を与えた場合には，たとえそれがかなり粗いものであっても，まずまず満足できるモデルを構築できることを示している．

次に登録語と未知語との間で同表記となりうるものに関して，生起確率を比較し

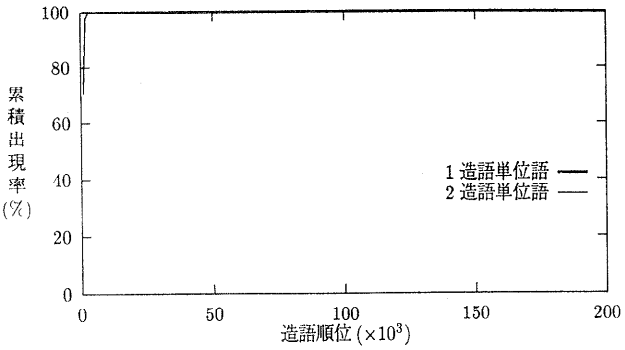

図 12 構成造語単位数別累積出現率（重要語ランク 3）
Fig. 12 Cumulative frequency of importance rank 3 words classified by the number of formative units.

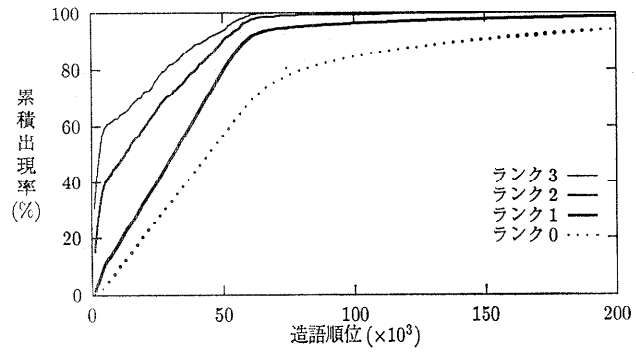

図 13 重み付け非付与時重要語ランク別出現率
Fig. 13 Cumulative frequency of words classified by the rank of importance, but with no difference of initial probabilities by rank.

表 6 同音語優先順位比較実験結果
Table 6 Comparison between homonyms.

同音語が存在する読みの総数		7,206 種
同音語組の総数		26,880 組
	同音語組正答率	最優先語正答率
重み付けなし	55.83%	49.82%
重み付けあり	63.34%	61.25%
(参考)		
重み付け初期確率	32.83%	39.95%

表 7 登録語カテゴリー間優先順位比較実験結果
Table 7 Comparison between words with same notation.

同表記語総組数	40,799 組
内，異カテゴリー間での組数	15,147 組
	異カテゴリー間正答率
重み付けなし	58.46%
重み付けあり	88.76%

た実験を示す．比較は漢字表記(正書法)，かな表記（読み），全表記（漢字かな混じりで記述可能な表記すべて）について行った．結果を**表 8** に示す．

これは登録語と未知語とで同表記となる組をすべて求め，登録語の生起確率が未知語の生起確率よりも高くなっているものを正答として数え上げたものである．各組において登録語の方が真に生起確率が高いかは評価していないため，実際には逆の確率大小関係の方がもっともらしいと思える組も存在する．

表 8 から重み付けを行わない方が高い正答率となっていることがわかる．重み付けを行った場合に正答率が低下するのは，重み付けにより強い結合と見なされた造語単位間結合の影響による．そのような結合を含む未知語は比較的高い生起確率を持ちやすくなり，低い重み付けをなされた登録語よりも生起確率が大きくなることが増えるためである．各登録語に与える重み付けの値の差異を大きくするほど，このような逆転も多くなると予想される．

このことは未知語の単語らしさ評価においてはそれほど問題にはならないが，単語の生起確率値推定としては少々問題となる．しかし生起確率が高ければ必ず単語辞書に登録されるというものでもないこと，重要な語であれば生起確率はかなり低くとも登録されることなどを考えれば，およそ 7 割という正答率は，良くはないが極端に悪いというほどでもないであろう*.

3.4 考　　察

造語モデルの構築は，原データとなる辞書見出しに大きく依存する．今回実験に用いた単語辞書では本章で述べてきたような結果を得たが，原データによっては全く異なる結果を生じる可能性がある．

造語単位の接続傾向をとらえるという観点からすれば，情報源となる登録語は多い方が望ましいのだが，ただ多ければいいというものではない．複合語が数多く登録されているような辞書では，本論文のモデルで

* 実用性は数字ほどは低くないであろうと考える．例えば，同表記の登録語が 3 語，未知語が 2 語存在したとする．このとき確率の高い順に［未，登，登，登，未］と出てきたとすると，この実験での正答率は 50% となるが，実用性はそう低くはないように思える．

表 8 同表記登録語・未知語間比較実験結果
Table 8 Comparison between generated words with same notation.

漢字表記総比較組数		534組
	正答数	正答率 (%)
重み付けなし	374	70.0
重み付けあり	359	67.2
かな表記総比較組数		7,206組
	正答数	正答率 (%)
重み付けなし	5,342	74.1
重み付けあり	5,154	71.5
全表記総比較組数		24,206組
	正答数	正答率 (%)
重み付けなし	17,204	71.1
重み付けあり	16,635	68.7

うまく扱えない構造を多く含む可能性が高い．このような構造に由来する不適切な状態遷移が多くモデル中に紛れ込めばそれだけ多くの不自然な造語を行うことになる．現在の一般的な形態素解析技法では，複合語のような長い単語も数多く辞書に登録しておいた方が都合がよいが，そのような辞書を用いて造語モデルを構築すると良い結果は得にくいであろう．

また，今回の実験では表に出なかったが，造語モデルを実際に解析に用いることを考えた場合，1造語単位から構成される語の生起確率が高すぎないかという問題も残している．複合語を生成する文法規則とも絡んでくるが，このように短い単語の確率が高すぎると，長い単語を短い単語の複合語として誤認識しやすくなる．この点についての検討は，今後の課題である．

4. おわりに

本論文では，単語の生起確率の推定および未登録語の単語らしさ評価を行うためのモデルとして造語モデルを構成し，その評価実験を行った．このモデルでは，単語の造語をマルコフ過程としてとらえモデル化しているが，その状態遷移確率を通常の統計的手法により得るのではなく，単語辞書の見出しを利用して推定する方法を取っている．このモデルが単語辞書の見出しという乏しい情報から出発して作成したものであることを考慮すると，実験結果は造語モデルが十分成功していることを示していると考える．

謝辞 著者の一人である永井に対し，日頃より助言いただいている九工大の野村浩郷教授，ならびに論文の質の向上のために適切なご指導をいただいた査読者の方々に感謝いたします．また，本研究では，九州芸工大および九大にて開発された「九州大学大型計算機センター公用データベース日本語単語辞書（原辞書：九州芸工大自立語辞書 KID-J 82)」を使用させていただきました．九州芸工大の稲永紘之助教授を始め，辞書作成に携わられた方々に感謝いたします．

参考文献

1) 藤崎哲之助：確率的言語処理へのアプローチ，情報処理学会自然言語処理研究会資料，41-6 (1984).
2) 武田浩一，藤崎哲之助：統計的手法を用いた漢字複合語の短単位分割，情報処理学会自然言語処理研究会資料，48-2 (1985).
3) 松延栄治，日高 達，吉田 将：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，情報処理学会自然言語処理研究会資料，55-4 (1986).
4) 永井秀利，松延栄治，日高 達：未登録語の単語らしさの評価値を計算する単語の造語モデル，九州大学工学集報，Vol. 63, No. 5, pp. 527-533 (1990).
5) 吉田 将：形態素解析，日本語情報処理第4章，pp. 86-113，電子情報通信学会 (1984).
6) 木谷 強：固有名詞の特定機能を有する形態素解析処理，情報処理学会研究報告，92-NL-90, pp. 73-80 (1992).
7) Yoshimura, K. and Shudo, K.: Towards the Intelligent Processing of Unknown Words, *International Symposium on Natural Language Understanding and AI*, pp. 161-166 (1992).
8) 大場健司，元吉文男，井佐原均，横山晶一，石崎俊：未定義語を含む文の多段階解析法，情報処理学会自然言語処理研究会資料，70-4 (1989).
9) 塚田孝則，西野敏行，小柳和子：未登録語を含む文の一解析法，情報処理学会研究報告，89-NL-73, pp. 43-50 (1989).
10) 吉村賢治，武内美津乃，津田健蔵，首藤公昭：未登録語を含む日本語文の形態素解析，情報処理学会論文誌，Vol. 30, No. 3, pp. 294-301 (1989).
11) 中村貞吾，酒井千佳生，永井秀利，日高 達：造語モデルに基づく確率文法の構文解析，第40回電気関係学会九州支部連合大会，812 (1987).
12) 永井秀利，中村貞吾，日高 達：造語モデルに基づく単語表記の扱い，第39回情報処理学会全国大会論文集，1F-3 (1989).
13) Nagai, H. and Hitaka, T.: Japanese Stochastic Grammar with Japanese Word Notation Model, *The 2nd Japan-Australia Joint Symposium on Natural Language Processing*, pp. 139-149 (1991).
14) 永井秀利：文解析における確率の利用，情報処理学会自然言語処理シンポジウム論文集，Vol. 93, No. 1, pp. 33-47 (1993).

15) Baum, L. E. and Eagon, J. A.: An Inequality with applications to Statistical Prediction for Function of Markov Processes and to a Model for Ecology, *Bull. Amer. Math. Soc.*, Vol. 73, pp. 360-363 (1967).
16) Baum, L. E., Petrie, T., Soules, G. and Weiss, N.: A Maximization Techinique Occuring in the Statistical Analysis of Probabilistic Functions of Markov Chains, *The Annals of Mathematical Statistics*, Vol. 41, No. 1, pp. 164-171 (1970).
17) 永井秀利，中村貞吾，日高 達：造語モデルにおける状態遷移確率推定法について，第44回情報処理学会全国大会論文集，4P-1 (1992).
18) 吉田 将，日高 達，稲永紘之，田中武美，吉村賢治：公用データベース日本語単語辞書の使用について，九州大学大型計算機センター広報，Vol. 16, No. 4, p. 27 (1983).
19) 金田一京助，見坊豪紀，金田一春彦，柴田 武，山田忠雄：新明解国語辞典第三版，p. 1328，三省堂 (1981).

(平成5年3月19日受付)
(平成5年6月17日採録)

永井 秀利（正会員）

昭和38年生．昭和61年九州大学工学部電子工学科卒業．昭和63年同大大学院工学研究科電子工学専攻修士課程修了．平成3年同大学院博士課程単位取得退学．工学修士．同年より九州工業大学情報工学部助手となり現在に至る．自然言語処理，特に日本語の計算機処理に興味をもつ．

日高 達（正会員）

昭和14年生．昭和40年九州大学工学部電子工学科卒業．昭和42年同大大学院工学研究科電子工学専攻修士課程修了．昭和44年同大学院博士課程中退．工学博士．同年九州大学工学部助手，昭和48年同講師，昭和55年同助教授，昭和63年同教授，現在に至る．形式言語の方程式論，自然言語処理，手書文字認識の研究を行う．電子情報通信学会会員．