December 9, 1998
RT0288
Human-Computer Interaction 19 pages

Research Report

A word-based Japanese language model

N. Itoh, M. Nishimura, S. Ogino, and K. Yamasaki

IBM Research, Tokyo Research Laboratory
IBM Japan, Ltd.
1623-14 Shimotsuruma, Yamato
Kanagawa 242-8502, Japan

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

自然言語処理　Vol. 6　No. 1　Jan. 1999

単語単位による日本語言語モデルの検討

伊東　伸泰[†]　西村　雅史[†]
荻野　紫穂[†]　山崎　一孝[†]

日本語では単語の境界があいまいで、活用等のルールに基づいて定義された単位である形態素は必ずしも人が認知している単語単位や発声単位と一致しない。本研究では音声認識への応用を目的として人が潜在意識的にもつ単語単位への分割モデルとその単位を用いた日本語の言語 (N-gram) モデルについて考察した。本研究で用いた単語分割モデルは分割確率が 2 形態素の遷移で決定されるという仮定を置いたモデルで、人が単語境界と考える点で分割した比較的少量のテキストデータと形態素解析による分割結果とを照合することにより、パラメータの推定を行った。そして多量のテキストを同モデルにしたがって分割し、単語単位のセット（語彙）と言語モデルを構築した。新聞 3 誌とパソコン通信の投稿テキストを用いた実験によれば約 44,000 語で、出現した単位ののべ 94-98%がカバーでき、1 文あたりの単位数は形態素に比べて 12%から 19% 少なくなった。一方、新聞とパソコン通信ではモデルに差があるもののその差は単語分割モデル、言語モデル双方とも事象の異なりとして現れ、同一事象に対する確率の差は小さい。このため、新聞・電子会議室の両データから作成した言語モデルはその双方のタスクに対応可能であった。

キーワード:　音声認識、ディクテーション、*N-gram* モデル、形態素解析

A Word-based Japanese Language Model

Nobuyasu Itoh [†] , Masafumi Nishimura [†] , Shiho Ogino [†]
and Kazutaka Yamasaki [†]

This paper deals with a word-based language model of Japanese. In Japanese, word boundaries are not stable and grammatical units do not necessarily coincide with human intuition. For accurate segmentation it is therefore necessary to create a vocabulary set that covers human utterance units. In our word-segmentation method, a model of word boundary is described by morphological parameters (i.e. part of speech), which are learned by comparing results of human segmentation with those of Japanese morphological analyzer. Then by using pseudo-random number and the model, it is determined whether each morpheme transition is a word boundary. As a result, we obtain a vocabulary set and learning data for Japanese language model automatically. According to our experiments using articles from three newspaper and appended texts in network-based forums, about 44,000 words cover 94-98% of all words in the test data, and the average numbers of words per sentence are 12-19% smaller than those of morphemes. The parameters of word segmentation model and language model are quite different in newspaper articles and forum's texts. However, the difference does not exist in the probabilities of common events, but in the kinds of events. Therefore the language model, which was created from newspaper articles and forum's text, gave the satisfactory results for both test set.

KeyWords: *Speech recognition, Dictation, N-gram model, Morphological analysis*

1 はじめに

音声認識技術はその発達にともなって、その適用分野を広げ、日本語においても新聞など一般の文章を認識対象とした研究が行なわれるようになった (松岡達雄, 大附克年, 森岳至, 古井貞熙, 白井克彦 1996; 西村雅史, 伊東伸泰, 山崎一孝, 荻野紫穂 1998b)。この要因として、音素環境依存型 HMM による音響モデルの高精度化に加え、多量の言語コーパスが入手可能になった結果、文の出現確率を単語 N 個組の生起確率から推定する N-gram モデルが実現できるようになったことが挙げられる。日本語をはじめとして単語の概念が明確ではない言語における音声認識を実現する場合、どのような単位を認識単位として採用するかが大きな問題の 1 つとなる。この問題はユーザーの発声単位に制約を課す離散発声の認識システムの場合に限らない。連続音声の認識においても、ユーザーが適時ポーズを置くことを許容しなければならないため、やはり発声単位を考慮して認識単位を決める必要がある。従来日本語を対象とした自然言語処理では形態素単位に分割することが一般的であり、またその解析ツールが比較的よく整備されていたことから N-gram モデル作成においても「形態素」を単位として採用したものがほとんどである (松岡達雄他 1996; 伊藤克亘, 松岡達雄, 竹澤寿幸, 武田一哉, 鹿野清宏 1996)。しかしながら、音声認識という立場からあらためてその処理単位に要請される条件を考えなおしてみると、以下のことが考えられる。

- 認識単位は発声単位と同じか、より細かい単位でなければならない。形態素はその本来の定義から言えば必ずこの条件を満たしているが、実際の形態素解析システムにおいては、複合名詞も 1 つの単位として登録することが普通であるし、解析上の都合から連続した付属語列のような長い単位も採用している場合があるためこの要請が満たされているとは限らない。
- 長い認識単位を採用する方が、音響上の識別能力という観点からは望ましい。つまり連続して発声される可能性が高い部分については、それ自身を認識単位としてもっておく方がよい。
- 言語モデルを構築するためには、多量のテキストを認識単位に分割する必要があり、処理の多くが自動化できなければ実用的ではない。

これらは、言い換えれば人間が発声のさいに分割する (可能性がある) 単位の Minimum Cover Set を求めることに帰着する。人が感覚的にある単位だと判断する日本語トークンについて考察した研究は過去にも存在する。原田 (原田悦子 1989) は人が文節という単位について一貫した概念を持っているかについて調査し、区切られた箇所の平均一致率が 76%であり付属語については多くの揺れがあったと報告している。また横田、藤崎 (横田和章 藤崎博也 1996) は人が

† 日本アイ・ビー・エム 東京基礎研究所, Tokyo Research Laboratory, IBM Japan, Ltd.

短時間に認識できる文字数とその時間との関係から人の認知単位を求め、その単位を解析にも用いることを提案している。しかしながら、これらの研究はいずれも目的が異なり、音声認識を考慮したものではない。そこで、われわれは、人が潜在意識としてもつ単語単位を形態素レベルのパラメータでモデル化するとともに、そのモデルに基づいて文を分割、N-gram モデルを作成する手法を提案し、認識率の観点からみて有効であることを示した (西村雅史 伊東伸泰 1998a)。本論文では主として言語処理上の観点からこの単語単位 N-gram モデルを考察し、必要な語彙数、コーパスの量とパープレキシティの関係を明らかにする。とくに新聞よりも「話し言葉」に近いと考えられるパソコン通信の電子会議室から収集した文章を対象に加え、新聞との違いについて実験結果を述べる。

2　単語単位への分割

本節ではわれわれが採用した単語単位と、同単位への分割手法について述べる。

日本語を分割して発声する場合、その分割点はきわめて安定している点と、人、または時によって分割されたりされなかったりする不安定な点がある。例として「私は計測器のテストを行っています。」という文を考えよう。これは形態素解析により、たとえば

私 + は + 計測 + 器 + の + テスト +を + 行 + っ + て + い + ます + 。

と分割されるが、動詞の活用語尾である「っ」や接続助詞の「て」はほぼ確実に「行」と結合して「行って」と発声されるのに対し、接辞である「器」は分割される場合もあれば、結合されることもあるだろう。そこで文がある位置で「分割」される確率を形態素のレベルでモデル化することを考える。そして人が分割した学習用テキストと同じテキストを形態素解析により分割した結果を照合し、各形態素の遷移ごとに当該点で分割される確率を得る。その後、より大量のテキストをそのモデルに基づいて分割すれば (このプログラムを以後セグメントシミュレータと呼ぶ)、人が分割した傾向をもったわかち書きテキストを容易に得られる。

「分割」される位置としては、形態素の境界 (形態素単位への分割) とさらに細かく形態素の途中 (文字単位への分割) がある。ここで分割記号として$\sharp$を使用し、「分割」は記号「$\sharp$」が生起し、「結合」は「NULL」が生起すると考えれば、前者はある形態素から別の形態素に遷移したときにその間に「$\sharp$」が生起する確率として

$$P(\sharp_i \mid Morpheme_i \rightarrow Morpheme_{i+1})$$

となる。後者のそれは $Morpheme$ を文字列 $C_1C_2,\ldots,C_n$で表すと、その j 番目の文字の後に$\sharp$が生起する確率と考えれば

$$P(\sharp_j \mid Morpheme,\, C_j \rightarrow C_{j+1})$$

と表現できる。モデルのパラメータ (形態素の属性) としては、品詞情報 (KoW)、連接属性

(Part of Speech: PoS)、、そして表記 ($String$) を採用し、($KoW[PoS], String$) と表現する。ここで品詞、連接属性とはわれわれの用いた形態素解析プログラム (丸山宏 荻野紫穂 1994) の出力として得られるものであり、品詞は 81、連接属性は 119 に分類されている[1]。したがって形態素単位の分割では 6 個、文字単位への分割では 4 個のパラメータで記述されることになるが、そうすると明らかに多量の学習用テキスト (人が分割したもの) が必要となる。そこで頻度が閾値以下であるような場合については、パラメータを特定の順序で縮退させた確率値を用意しセグメントシミュレータの実行時も、確率が記述されているレベルまで同様の順序で縮退し、当該確率値で代用することを考える。縮退の順序にはさまざまなものが考えられるが、モデルのパラメータについてその種類数を考えると表記、連接属性、品詞の順に少なくなることは明らかであり、縮退もそれにしたがうのが妥当であろう。また基本的にはある出現回数を閾値としたときより多くの種類の遷移確率が得られることが望ましい。このような観点からいくつかの予備実験を行い経験的に縮退順序を決定した。この順序と参照される確率値を木構造で表現したのが図 1である。各ノードには形態素の属性とその属性が満たされた場合に分割される確率が対応する。たとえば図 1中

$$P(\sharp \mid V.infl.[29] \rightarrow Conj.p.p.[69],\ \text{て})$$

は形態素単位への分割に対する記述例で、形態素の属性が動詞活用語尾 [29] から接続助詞 [69]「て」へ遷移したときに、その間で分割される確率を意味する[2]。1 つ上のレベルでは、表記（ここでは「て」）が省略される。ただし品詞が名詞の場合には文字数が分割確率を記述するパラメータとして有効と考えられるので[3]、表記を省略した場合、文字数をパラメータとして残した。さらに上位レベルでは、連接属性番号も省略し、品詞 $V.$ *infl.* から $Conj.$ $p.p.$ への遷移に対して、人が分割する確率を記述する。たとえば、「積んで」という文節を形態素に分割すると

$$\text{積}\,(Verb[8])\ +\ \text{ん}\,(V.infl[30])\ +\ \text{で}\,(Conj.p.p.[69])$$

となるが、その中に現れる「ん」と「で」の間で分割されたカウント等もマージした上で算出された確率となる。このように木はリーフから上位のノードに行くにしたがって縮退されたパラメータ、言い換えればより大まかなパラメータとなる。

一方、前節で述べたように人は形態素として定義されたトークンをさらに文字単位で分割する場合もある。これは形態素解析の都合上連続した付属語列を 1 つの形態素としてとり扱うことが行なわれるためである。たとえばわれわれの用いた形態素解析用文法では「... かどうか」

1 品詞情報は学校文法でいう品詞分類（「動詞」「助動詞」など）に相当するが、解析の都合上一般にその品詞であると認められていない形態素に当該品詞を割り当てている場合がある。その場合は、後の処理のため同じ品詞でも単に「助動詞」とするのではなく「助動詞 A」のように区別しており、結果的に種類が増大している。また連接属性は品詞を活用型などによりさらに詳細分類したもので、たとえば動詞は 17 種類に分類されている。意味からすれば品詞情報を PoS (Part of Speech) とすべきであろうが、ここでは文献 (丸山宏・荻野紫穂 1994) の記法にしたがった。

2 $V.$ *infl.* は Verb inflection、 $Conj.$ $p.p.$ は Conjunctive post-positional particle の略。

3 「誤認識」が「誤」「認識」と分割されるよりは「音声認識」が「音声」「認識」となりやすいなど。

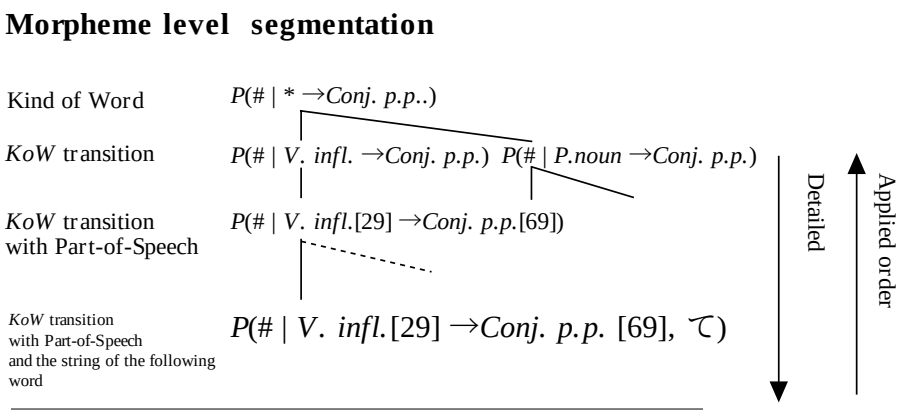

図 1　セグメントシミュレータにおけるパラメータ縮退の順序

という付属語列が助詞として扱われているが「か」＋「どうか」と分割されることもある。そこで形態素レベルの分割よりもさらに詳細なレベルとして、文字レベルの分割をモデル化した。このような確率木はつぎのように構成することができる。つまりもっとも細かい分類における各パラメータについて、人が分割した結果と形態素解析の結果を照合してカウントし、その値をリーフから上位ノードに伝搬させた後、確率値に正規化すればよい。全カウント数が少ないと当該確率 (推定値) の信頼性が低いので、カウント、マージ作業を行なって、頻度がある閾値以上のノードを最終的なノードとして採用することにする。

このモデル化では学習データの量に応じて、そのデータから得られる情報を最大限に利用することができる。たとえば、2 文字漢語から接尾辞への遷移には、非常に多くのものがあるが、その分割されやすさは接尾辞の種類によって異り、それらを捨象してモデル化したのでは、あいまいさが大きくなってしまう。しかし逆にそのすべてを細分化したのでは、頻度が低い接尾辞に対するルールが得られないか、または信頼性の低い確率推定値となってしまう。本手法によれば学習データ中に頻度が高いものについてはより細かい分類でモデル化され、頻度が下るにしたがって統計として信頼にたる単位まで縮退されたパラメータによる確率値が得られることになる。

3　形態素解析プログラムの変更

3.1　現代語書き言葉以外の表現への文法の対応

形態素解析システムは、一般に新聞記事に代表される現代語書き言葉を処理できるように開発されてきた。しかし近年、データとして使用されるコーパスの大規模化に伴い、現代語書き言葉以外の表現、特に、会話風の表現（以下、口語体と示す）を扱う試みが増加してきた (黒橋禎夫, 坂口昌子, 長尾真 1996)。われわれが従来使用してきた形態素解析の文法規則 (丸山宏・荻野紫穂 1994) も、原則として現代語書き言葉に対応したもので、口語体への対応は十分ではない。一方本研究で用いる学習用テキストは新聞に限らず、パソコン通信の投稿テキストが含まれており、口語体への対応なくしては充分な精度の解析結果を得ることができない。以下の点を考慮して、より多様な文に対応できるよう形態素解析の文法を記述した。

- 元の文法に対する変更を少なくして派生的な影響を抑える。
  口語体によく現れる縮退形で、五段活用連用形に接続する「ちゃ」には、接続助詞「て」および係助詞「は」の連なり「ては」の縮退と（例: 書い**ちゃ**いけない）と、接続助詞「て」および補助動詞「しまう」の語幹の連なり「てしま」の縮退（例: 書い**ちゃ**う）とがある。前者は直後で文節を切ることができる非活用語、後者はワア行五段活用をするので、ワア行五段活用語尾が接続し、かつ直後で文節末に遷移できる「ちゃ」という形態素の規則を作成すれば形態素解析処理を行うことができる (黒橋禎夫他 1996)。しかし、品詞や活用形を単語分割モデルで利用すると、「ちゃ」に品詞として接続助詞を付与すれば「接続助詞にワア行五段活用語尾が接続する」という一般化が、また動詞を付与すれば「五段動詞語幹が文節末に遷移する」という一般化が行なわれかねない。これを避けるには、「ちゃ」に新たな品詞を付与するか、または「ちゃ」に二種類あるとするという対応が考えられるがわれわれは後者の方法を採った。形態素解析としては前者が望ましいと思われるが、後の単語分割モデルに影響を及ぼす可能性がある場合は、元の文法規則への影響がより少ないものを採用した。また、文語活用の残存形などで、現代語活用に全く同じ形があるものについては、現代語活用の形態素に接続条件を加えて対処した。
- 縮退形の品詞付与では元の形態素列のうち活用語尾や自立語がもつ品詞を優先する。
  形容詞仮定形活用語尾「けれ」および接続助詞「ば」の連なりの縮退である「きゃ」「けりゃ」の前連接属性は「けれ」、後連接属性は「ば」にほぼ等しい。こうした縮退形の品詞は、元の形態素列のもつ連接属性のうち活用語尾や自立語のものを優先して付与した (荻野紫穂 1998)。
- 省略による空文字列は次形態素への遷移を追加して対処する。
  「勉強しよ」「読も」などのように形態素末が落ちる縮退の場合、前者は助動詞「よう」の縮退「よ」を定義すればよいが、後者は助動詞「う」そのものが脱落しているので、動

詞未然形から「う」の次の形態素への遷移を追加して対処する。

3.2　複合名詞の分割

形態素解析の辞書には、現在までの使用目的に応じて複合語が一語扱いで登録されていることが多いが、単語分割モデル構築のための形態素解析としては短単位に分割されていた方が都合がよい。そこで、複合語の中でも特に多い複合名詞を分割対象として、分割データベースとヒューリスティック規則により、形態素解析で複合名詞分割を行なうことにした。複合名詞の分割データベースは、2 カ月分の新聞記事（産経新聞）を形態素解析してその結果から一定以上の頻度で出現する 3 文字以上の名詞を抜き出した後、人手で、分割する位置の情報を付与することにより作成した。このデータベースには約 25,000 語の複合名詞が含まれている。ヒューリスティック規則は、以下の条件を満たすように作成した。

- 1 語の名詞よりも 2 語以上の名詞連続のコストが小さい。
  名詞連続中では、2 語のコストがもっとも小さく、次第にコストが増大するように設定する。これは複合名詞を分割する際、あまり細かく切り過ぎないようにするためである。
- 1 文字名詞は他の名詞に比べてコストが大きい。
  上記と同様、過分割を防ぐためである。
- 分割対象は 3 文字以上の複合名詞とする。
  1 文字ずつに過分割しないためである。
- 未知語のコストは 1 語の名詞より大きい。

また、分割の結果に 3 文字以上の名詞が含まれている場合は、再帰的にそれを分割し、分割が不可能になるまで繰り返す。

4　分割モデルの作成と分割過程

4.1　分割確率の推定

分割ルールとその確率を推定するため、計 17 人の被験者により、新聞 5 カ月分（日経新聞 3 カ月および産経新聞 2 カ月）、日本語用例集 (合計約 26,000 文)、そしてパソコン通信「ピープル」の電子会議室（以下電子会議室）から採取した文章 (約 9,500 文）を分割する作業を行った[4]。

新聞や日本語用例集はいわゆる「書き言葉」のスタイルであるのに比較して電子会議室の文章はより口語体に近く、これらは分割モデルにも影響を与える可能性がある。そこで両者のデータは別々に取り扱って分割モデル (確率木) を構成した。その結果前者は 2,829 個、後者は

4 文選択は文の長さが一定の範囲に入っていることを除けば無作為に行なった。また被験者には 1. 不自然にならない限り、より細かく分割すること 2. 書かれた文章ではなく発声する場合の分割点を回答することという指示を与えた。

表 1　生成された木に記述された分割確率の例（新聞データから得られたもの）

パラメータ値	分割確率
名詞 [19] → 名詞 [19],「者」	0.33
名詞 [19] → 名詞 [19],「人」	0.71
名詞 [19] → 形容動詞 [18],「的」	0.36
動詞活用語尾 [29] → 接続助詞 [69],「て」	0.03
名詞 [19] → 格助詞 [77],「を」	1.0

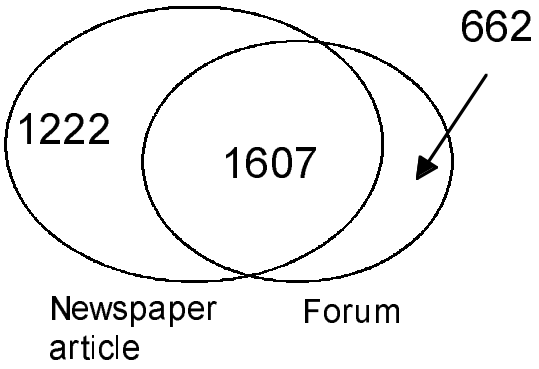

図 2　確率木のノード数

2,269 個のノードからなる木が得られた。表 1に一例を示す。ただしノードとして採用するか否かの閾値には当該ノードの出現回数（カウント）を用い、その値は学習データ中の単語数に比例させた [5]。 2 つの確率木について得られたノードをいずれに含まれるかで分類し数を示したものが図 2である。得られたノードは、かなりの異なりがあることがわかる。たとえば電子会議室データから得られた確率木にのみ存在するノードの中で出現回数の多いものから上位 3 個[6]をあげると以下のようになる。

1. 接続助詞 [69]　→　活用語尾 [31]「る」
2. 助動詞 [62]　→　接続助詞 [73]「が」
3. 助動詞 [48]　→　接続助詞 [73]「けど」

これらの遷移を含む例文を上げると 1. 読ん+で+**る**, 2. ... です+**が**, 3. ... だ+**けど**などであ

5 新聞データの場合で 50 である。
6 遷移後の表記 $String$が縮退していないレベルのものに限った。

り、明らかに口語体特有の言い回しに伴う遷移が抽出されている。一方新聞データから学習した確率木にのみ存在するノードをみると体言止めに伴う遷移 (サ変動名詞 [13] → 句点 [100]「。」i.e.「... を議論 + 。」) や漢語の接辞 (名詞 [19] → 接辞 [19],「会」) など直感的にも電子会議室等の文章では比較的頻度が低いと考えられるものが多かった。また両方の確率木に共通して出現しているノード 1,607 個について分割確率の相関係数を求めたところ 0.980 となりきわめて高い。したがって共通するノードについてはほとんど違いはなく、2つの確率木の違いはノードつまりルールそのものに現れていることがわかった。

これらのモデルに基づいて以下のように多量の (形態素解析された) テキストを分割・統合する。

(1) 各形態素およびその遷移について、連接属性番号、品詞、形態素の表記を得て、確率木のリーフに記述があるかどうかを調べる。

(2) なければ、木作成の説明で述べた順にパラメータ値を縮退させ、確率木に記述があるかどうかを調べる。

- 記述があれば、0 から 1 の範囲の乱数を発生し、その値がノードに付随する確率以下であれば当該位置で分割し、そうでない場合は分割しない。
- 記述がなければ、縮退を繰り返す。

(3) もっとも上位のノードにも該当しない場合、形態素の分割点であれば当該位置で分割し、それ以外は分割しない。

なお N-gram モデル作成には、乱数による分割処理 (セグメントシミュレータ) は必ずしも必要ではなく、形態素解析の結果と分割確率を使って直接各 N-gram の生起確率を推定することも可能である。

4.2 単語カバレージ

われわれの提案した単語単位に基づく語彙を作成するための予備実験として日経新聞 3 カ月分 (合計 446,079 文) を用い、前節の手続きを適用して分割、連結を行う実験を行った。西村らの報告 (西村雅史, 大嶋, 野崎 1995) によれば形態素を単位とした場合、約 97%はおよそ 3 カ月分のテキストで収集できる（言い換えれば飽和する）ことがわかっている。その結果を図 3に示す。単語は合計で約 10^7 個、のべ 216,904 種類の単語が生成された。図はそれらを頻度の高いのものから順にとった場合のカバレージを示している。ただし数字表現、姓名はカウントから除いている。一方同じテキストから形態素は 132,164 個が生成された。これによれば単語単位を採用すると、形態素よりはより多くの種類が必要ではあるものの、決して発散するものではなく、たとえば上位約 25,000 個 (種類) の単語で全トークンの約 95%がカバーでき、取り扱いが可能な語彙数であることがわかる。

このとき確率木の各ノード（ルール）がどのような割合で使われたかを示したのが表 2であ

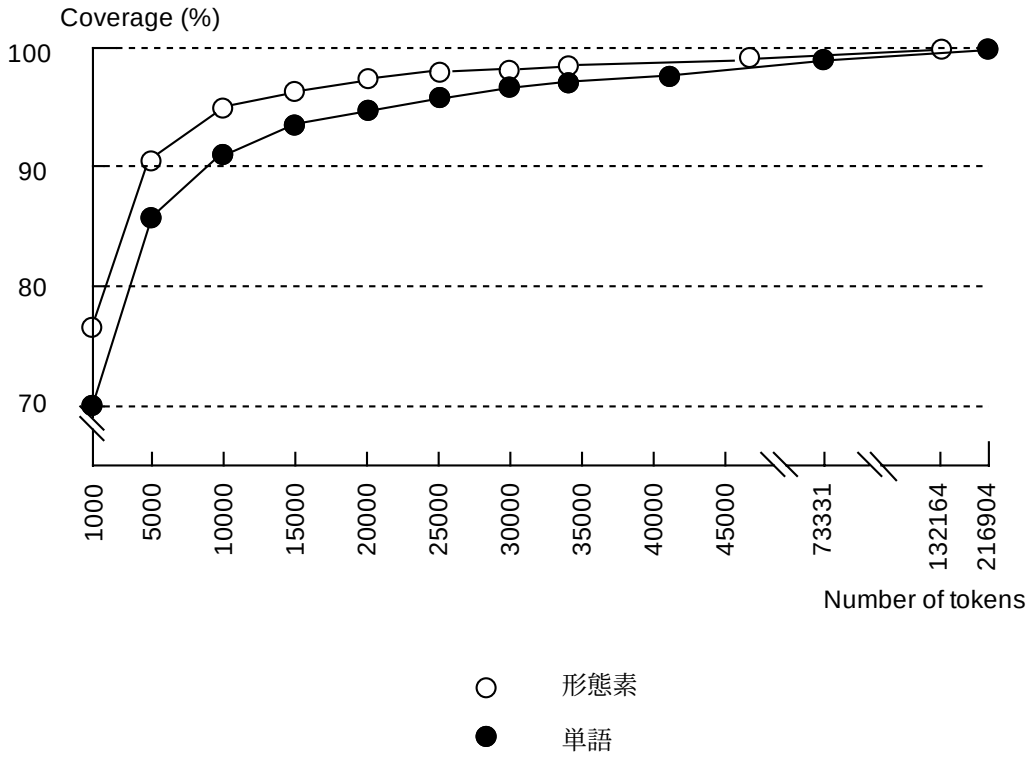

図 3　日経新聞 3 ヵ月のテキストに対する単語数とカバレージ

る。表から明らかなように全体の約 60% の場合には、一番詳細なレベルのルールが適用されていることがわかる。

表 2　適用されたノードの比率 (階層別)

パラメータ値	比率 (%)
$P(\sharp \mid KoW_1[PoS_1] \rightarrow KoW_2[PoS_2], String)$	59.6
$P(\sharp \mid KoW_1[PoS_1] \rightarrow KoW_2[PoS_2])$	29.2
$P(\sharp \mid KoW_1 \rightarrow KoW_2)$	3.9
$P(\sharp \mid KoW_2)$	6.6
該当なし	0.7

5　語彙とコーパス

5.1　コーパスの前処理

用意したコーパスのソースは日経新聞（93 年から 96 年）、産経新聞（92 年 10 月から 97 年)、毎日新聞（91 年と 92 年)、EDR コーパス (EDR 1995)、そしてパソコン通信「ピープル」に投稿された電子会議室の記事である。ただし日経、産経の両紙は示した期間のすべてではなく、月単位で時期が重複しないように選択したサブセットである。新聞についてはその本文を句点単位で文として取り出し、前節で述べた処理を行った。ただし数字については形態素解析で１単語（品詞「数字」）として扱われてしまうので当該トークンをすべて桁付きの漢数字に変換した後、西村 (西村雅史・伊東伸泰 1998a) に記載された数字の読み上げ単位に合わせて分割した。すなわち整数については「十、百、千、万、億」を位と定義し先行する数字と位で１つの単位として取り扱い、小数点以下の位については１桁づつに分割する。たとえば１２３４．５６は「千」「二百」「三十」「四」「・」「五」「六」と変換・分割されることになる[7]。

一方ディクテーションのアプリケーションや一般ユーザーが入力するであろう文、言い回しを考えると新聞だけでは明らかに不足である。そこでより口語体に近いデータとしてパソコン通信「ピープル」から約 90 の電子会議室に投稿されたテキストを用意した。会議室・話題の種類そして投稿時期について特に恣意的な選択は行っていないが、結果としてはパソコン関連の話題が多く、テキスト量でみて約半分を占めている。電子会議室の投稿文は文ばかりではなく、文字を利用した表、絵などが多数含まれている他、他人の記述を引用する場合が多く、これらを含めてしまったのでは学習用コーパスとして不適切であることは明らかである。そこでルールベースでこれらをとり除くフィルターを作成した。主なルールとしては以下のようなものがある。

- 引用記号（「＞＞」など）をもとに引用部分だと判断した行は除く。
- 記号文字 (「ー」「＊」など）の一定以上の繰り返しを含む行は除く。
- フェースマーク（「:-)」など）のリストを作成し、それにマッチした箇所は特別な１個の記号に置き換え、未知語の扱いとする。

このフィルターを通した後、句点に加え空白行、一定数以上の連続した空白を手がかりとして文を取り出し、形態素解析、セグメントシミュレータの処理を行った。

5.2　語彙の作成

以上の分割済みテキストの内、日経新聞、産経新聞、EDR、そして電子会議室について、95％以上のカバレージをもつ語彙を作成したところ、約 44,000 語の単語からなるセット (44K

7 電子会議室の文章では電話番号や ID 番号にともなう数字があり、これらは位付きで読むことに適さない。そこでルールでそれらに該当すると判断した場合は１桁づつに分割した。

語彙) が得られた。このようにして得られた語彙は、人が日本語について単語単位だと感覚的に思うセットを示していると考えられる。たとえば「行う」という動詞とその後続の付属語列からは

行い	行いたい	行う
行うべき	行え	行えば
行える	行った	行ったら
行って	行っても	

の計 11 単語が生成された。また「たい」や「べき」といった単語も生成されており、分割に揺れがある部分では複数の分割に対応した単語が得られることがわかる。

5.3　学習コーパス文の選択

前節の結果得られた各文は局所的に見ると記号ばかりであったり、姓名の列挙部分であったりして学習コーパスには適さないものが含まれている。また電子会議室のテキストはフィルターのルールでカバーしきれなかった部分で単語ではないトークンが無視できない程度に生じていた。このような文については人手で採用するかどうかを決める、あるいは当該部分を除くことが望ましいが、多量のコーパスについてそのような作業を行うのは不可能なため、ここでは以下の条件のいずれかに当てはまる文は採用しないことにした。

- ２単語以下から構成される文
- 文の単語数に対する記号の数が一定以上の文
- 44K 語彙に対して未知語の数が一定以上の割合で含まれる文

音声認識用のコーパスにおいて句読点や括弧表現をどのように取り扱うべきかについてはさまざまな議論がある。松岡ら (松岡達雄他 1996) はカギ括弧以外の括弧（( )【】など）について内容ごと削除しており、伊藤ら (伊藤克亘他 1996) は括弧の用いられ方（引用、強調など）に応じて削除すべきかどうかを自動判別している。括弧による表現には確かに読み上げに適さないものも含まれているが、本研究では文章入力手段としての音声認識システムの構築を重視し、これらを削除しないことにした。また同じ理由で句読点も削除していない。その結果得られた文の数をソース別に示す（表 3)。

6　単語単位による言語モデル

前節にしたがって単語単位に分割されたテキストを学習データとして N-gram モデルを学習するわけであるが、生起確率の計算上考慮すべきこととして数字、時刻などとくに各単語に確率上の差をつけるべき理由がないもの、および意味がまったく同じでありながら表記の異なる

表 3　ソース別のテキストサイズ（文と形態素の数、K は 1,000、M は 100 万を意味する）

ソース	文数 (K)	形態素数 (M)
日経新聞	715	20.9
産経新聞	1,837	49.4
毎日新聞	1,401	41.4
EDR	169	4.4
電子会議室	1,565	33.6

表 4　テストデータにおける諸元 (イタリックは 1 文あたりの平均数)

	文数	形態素数	単語数	カバレージ (%)
日経新聞	600	21,378	18,725	98.3
		35.6	*31.2*	
毎日新聞	725	22,051	18,608	96.1
		30.4	*25.7*	
産経新聞	775	21,702	17,751	96.0
		28.0	*22.9*	
電子会議室	1,381	29,979	24,204	94.4
		21.7	*17.5*	

揺らぎが生じているものの取り扱いがある。前者については各単語をクラスにまとめて確率を計算することにし、合計 36 クラス作成した。後者は新聞の場合、用語統一がなされているため影響は少ないと考えられるが、電子会議室のテキストでは「コンピュータ」と「コンピューター」、「組み合わせ」と「組合せ」といった単語は両者とも多数含まれており明らかに無視できない。そこで 44K 語彙について「読み」をもとに同義語の候補を抽出した上でチェックを行い、約 1,800 エントリの別名リストを作成した。*N*-gram をカウントするさい、このリストを参照して 1 つの表記に統一した上で学習を行っている。

一方、テストデータとして新聞 3 種類、電子会議室のテキストを別に用意し、被験者 (単語分割モデルの学習データを作成した被験者とは異なる）により分割を行なった。テストデータのそれぞれについて文数、形態素数、単語数、そして 44K 語彙のカバレージを表 4に示す。この表から 1 文あたりの単語数は形態素数に比較して 12-19%程度少なくなることがわかる。本実験の目的は

- 単語を単位とした *N*-gram モデルの有効性、コーパスの必要量を評価する。
- 新聞と電子会議室において単語 *N*-gram モデルから見た違いを明らかにする。

の 2 点である。そこで新聞、電子会議室のそれぞれについてその種類、時期の違いを捨象するため、全学習データを文単位でシャッフルした上で 8 個に分割したサブセット (新聞: N-1,..,8、電子会議室 F-1,...,8）を作成した。そして各サブセットをさらに 95%と 5%の比率で分割し前者を *N*-gram カウント、後者を Held-out 補間のパラメータ学習用に用いた。

まず新聞について学習データ (N-1,..,8) を順に増加させながら言語モデルを作成し、各モデ

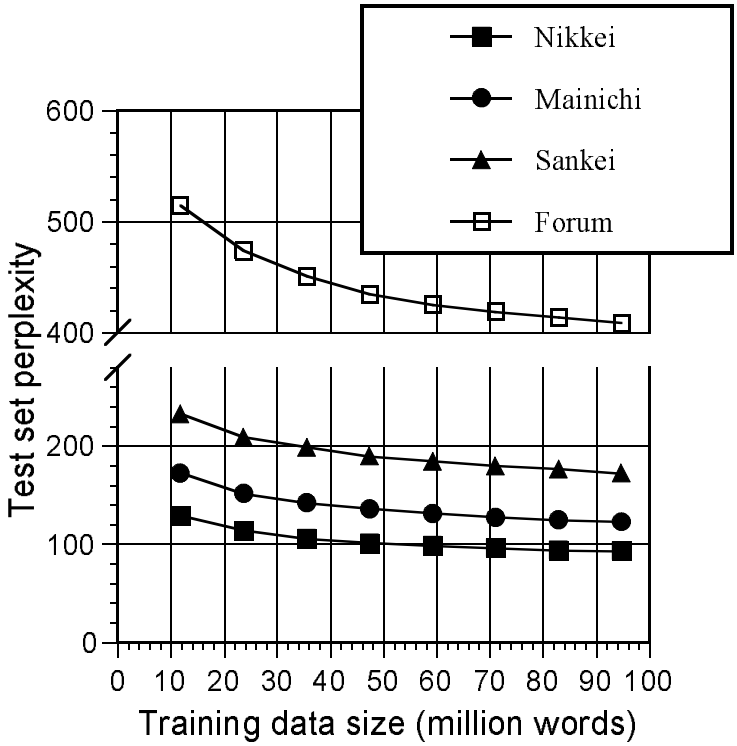

図 4　新聞データから学習したモデルのテストセットパープレキシティ

ルをテストセットパープレキシティで評価した。ただし学習データに１回でも出現した N-gram (trigram まで) はすべて使用しており、また未知語部分については予測を行っていない。結果を図 4に示す（電子会議室に関するデータは「Forum」と表記している)。予想されるようにいずれのテストデータでも学習コーパスの増加にともなってパープレキシティは緩やかに改善されるが次第に飽和する傾向がみてとれ、いずれの場合も学習データセットを 7 個から 8 個に増やしたときのパープレキシティの改善率は 1-2%程度でしかない。パープレキシティの絶対値には相当の差があり、新聞といってもひとくくりにできないことは明らかだが[8]、その値 (100-170) は音響識別上対応可能な値であると考えられる (西村雅史他 1998b)。一方電子会議室のテストデータはもっとも良いケースでも 400 以上のパープレキシティを示しており新聞の学習データだけでは対応できていないことがわかる。

われわれの目的は新聞にとどまらず、より口語体に近い電子会議室に投稿される文にも対

8 コーパスの量では産経新聞が一番多く、学習データ量でとくに不利に扱われたとは考えにくい。

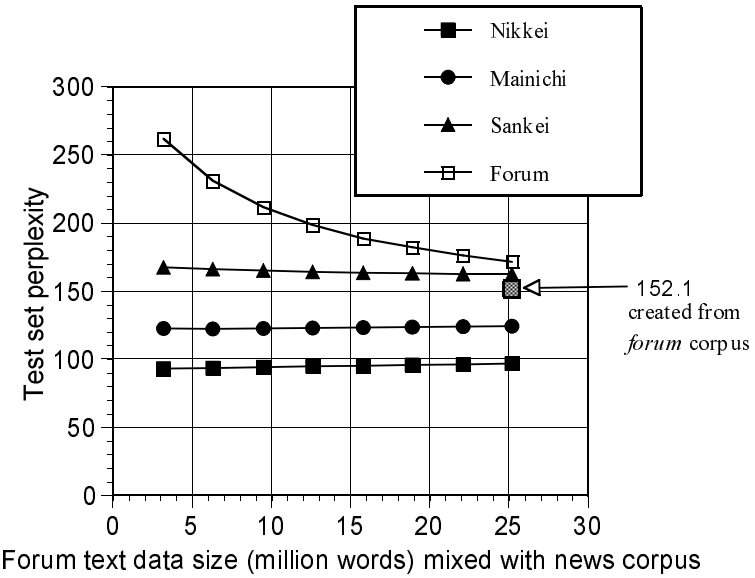

図 5　混合した電子会議室データのサイズとテストセットパープレキシティ

応できる言語モデルを作成することである。そこで新聞データすべてを使用した言語モデルをベースとし、電子会議室の学習データ (F-1,..,8) を加えていくことにより各テストデータのパープレキシティがどのようになるかを評価した。結果を図 5に示す。この結果、電子会議室についてそのパープレキシティは改善される一方、使用したデータ量の範囲 (約 25M 単語) では、新聞に対する影響はほとんどなかった[9]。一方電子会議室のみから作成した言語モデルで（電子会議室の）テストデータを評価すると 152.1 であり、若干の差は見られるものの、混合学習データから作成した言語モデルは新聞・電子会議室の双方に対応できることがわかる。これは双方の統計的異なりが共通している N-gram の確率が相違しているというよりも、N-gram の種類に、より大きく現れていることを示唆している。

一方コーパスのサイズと結果として得られたモデルのサイズ、すなわち N-gram の異なり数の関係を見たのが図 6である。これは新聞データ (N-1,..,8) の場合であるが、bigram、trigram とも飽和する傾向は見てとれない。電子会議室テキストを加えた場合も同様で N-1,...,8、F-1,...,8 すべてを学習データに使用した場合の N-gram 数は trigram が 31M 個、bigram が 5.6M 個に達した。とくに trigram は学習データサイズの増分に対しほとんど比例して増加している。今後主記憶、外部記憶の容量がさらに増加するとしてもこの N-gram 数（異なり）のままでは、実装することが難しい。そこで N-gram の中で低頻度のものを除くことが、パープレキシティに

9 細かく見れば、産経新聞はさらに改善されるのに対し、日経新聞はわずかながら悪くなる傾向があり、新聞間の差を示唆している。

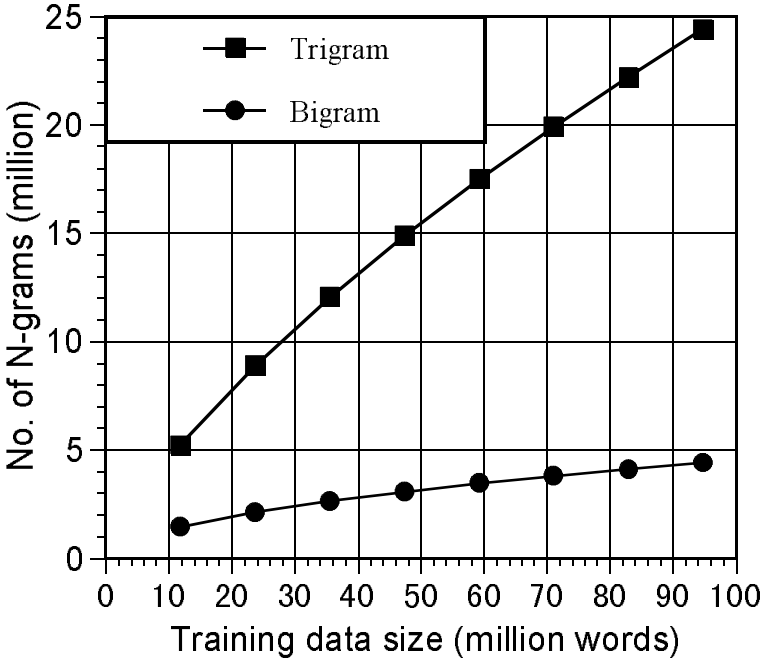

図 6　学習データサイズと *N*-gram の異なり数 (新聞)

どのような影響を与えるかを検証する実験を行った。結果を図 7に示す。*N*-gram の異なりの多くを占めるのは trigram なので、学習データは N-1,...,8、F-1,...,8 すべてを使用した上で、言語モデルを作成するとき trigram の最低出現回数を設定することにより、モデルのサイズを変更している。図から trigram の異なり数が 5M 個以下になるとパープレキシティが急速に悪くなる傾向が見てとれるが、一方モデルサイズを 1/3〜1/5 にした程度ではパープレキシティの差は小さいことがわかる[10]。

7　おわりに

このように、本研究では比較的少量の人による分割データから揺らぎを含めた分割傾向を推定する手法について述べ、新聞およびパソコン通信の電子会議室を学習データとして、そのモデルからつくられた単語の集合と言語モデルについて考察した。結果として、人が単語と意識する単位はその揺らぎを含めても発散することはなく、約 44K で 94-98%程度のカバレージが

10 ここでは、言語モデルに含める最低出現回数を 1,...,8 に設定している。グラフから出現回数 1 のものを除くだけで trigram の異なり数は 31M 個から 9M 個に減少することがわかる。

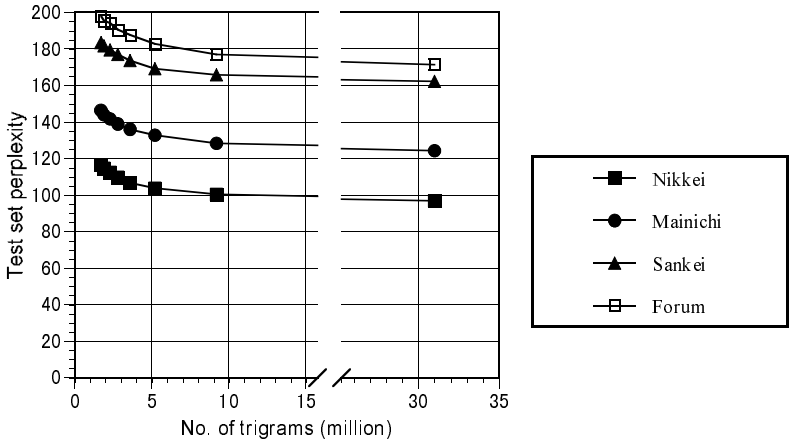

図 7　trigram の異なり数とテストセットパープレキシティ

得られること、形態素に比較して１文あたりの要素数が 12-19%程度減少すること、電子会議室と新聞では、*N*-gram モデルからみた統計量に相当の差があり、予想されたように新聞単体では十分に対応できないものの、新聞をベースとして電子会議室のテキストを混合させたデータから作成した言語モデルは新聞のテストデータに対するパープレキシティを増大させることはほとんどなくその双方に対応可能であることがわかった。分割モデル、*N*-gram モデルのいずれも、データの種類（新聞、パソコン通信）に依存している。これ自体は容易に予想できることであるが、その異なりが共通する事象の確率が異なるというよりも事象自体の異なりにより大きく現れていることは興味深い。

形態素との効率比較という意味では、同一学習データから作成した言語モデルを用いて単位長さ（たとえば文）あたりのパープレキシティを比較する必要がある。これについて学習、テストデータ量は少ないものの、すでに報告を行っており、文あたりパープレキシティがほぼ等しく、したがって単位長が長い分、より有利な単位となっていることを確認している（西村雅史他 1998b）。

コーパス量とパープレキシティの関係について、とくに日本語に関して報告された例はほとんどないため、他の研究と比較して議論することが難しい。本研究の実験からは 400 万文強のデータではまだパープレキシティが減少するが、その改善率は低く数倍以上のデータがないと意味のある改善が難しいことを示唆している。

人が感覚的にある単位だと判断する日本語トークンについて考察した他の研究との関連についても述べておきたい。原田（原田悦子 1989）は人のもつ文節単位の概念に関する調査結果か

ら、「文字列またはモーラ長が一定以上になると分割しようとする動機がたかまる」という仮説を提起している。われわれの分割モデルでは分割が 2 形態素の遷移情報のみで独立に起こることを仮定しているが、この独立性については検討が必要であろう。横田、藤崎 (横田和章・藤崎博也 1996) が短時間に認識できる文字数とその時間との関係から求めた認知単位は、とくに平均長は述べられていないものの、例をみる限りわれわれの単位より明らかに長い。同論文では「人は文を文字単位で処理しているのではない」と結論しているが、加えて、分割できる最小単位の列として知覚されているのでもないということになる。

今後は、コーパスサイズをより大きくするとともに句読点を削除した場合との比較・考察や、単語分割モデルの分割確率とポーズ位置との関係 (竹澤寿幸 森元逞 1996)、さらに上記で述べた分割の独立性について検討したいと考える。

本研究にテキストデータ使用を許諾していただいた、産経新聞社、日本経済新聞社、毎日新聞社（CD-毎日新聞 91-95）、そして (株) ピープルワールドカンパニーに感謝いたします。

参考文献

EDR (1995). 電子化辞書仕様説明書. (株) 日本電子化辞書研究所.

原田悦子 (1989). “日本語テキストにおける認知的単位.” 情処文書処理研究会, DPHI22-3.

伊藤克亘, 松岡達雄, 竹澤寿幸, 武田一哉, 鹿野清宏 (1996). “大語彙連続音声認識研究のためのテキストデータ処理.” 音響学会講演論文集, 3-3-10, pp. 105–106.

黒橋禎夫, 坂口昌子, 長尾真 (1996). “京都大学におけるテキストコーパスの作成.” 情報処理学会「大規模テキストコーパスの作成及び共有の問題」シンポジウム論文集, pp. 19–26.

丸山宏 荻野紫穂 (1994). “正規文法に基づく日本語形態素解析.” 情報処理学会論文誌, **35** (7), 1293–1299.

松岡達雄, 大附克年, 森岳至, 古井貞熙, 白井克彦 (1996). “新聞記事データベースを用いた大語い連続音声認識.” 電子情報通信学会論文集, **J79-D-II** (12), 2125–2131.

西村雅史, 大嶋良明, 野崎広志 (1995). “単語を認識単位とした日本語ディクテーションシステム.” 情処 51 全国大会, 3R-7, pp. 117–118.

西村雅史 伊東伸泰 (1998a). “単語を認識単位とした日本語ディクテーションシステム.” 電子情報通信学会論文集, **J81-D-II** (1), 10–17.

西村雅史, 伊東伸泰, 山崎一孝, 荻野紫穂 (1998b). “単語を認識単位とした日本語の大語彙連続音声認識.” 情処音声言語処理研究会, SLP20-3, pp. 17–24.

荻野紫穂 (1998). “毎日新聞テキストデータベース.” 技術研究組合新情報処理開発機構テキストサブワーキンググループ「研究開発用知的資源タグ付きテキストコーパス」報告書, pp. 122–135.

竹澤寿幸 森元逞 (1996). “部分木に基づく構文規則と前終端記号バイグラムを併用する対話音声認識手法.” 電子情報通信学会論文誌, **J79-DII** (12), 2078–2085.

横田和章 藤崎博也 (1996). “認知単位の bigram を用いた日本語文解析の一方法.” 自然言語処理, **3** (4), 129–139.

略歴

伊東 伸泰: 1982 年大阪大学基礎工学部生物工学科卒業．1984 年同大学院博士前期課程修了．同年，日本アイ・ビー・エム（株）入社．東京基礎研究所において文字認識、音声認識の研究に従事．情報処理学会会員．

西村 雅史: 1981 年 3 月大阪大学基礎工学部生物工学科卒業．1983 年 3 月同大学院物理系博士前期課程修了．同年，日本アイ・ビー・エム（株）入社．以来，同社東京基礎研究所において，音声認識などの音声言語情報処理の研究に従事．工学博士．平成 10 年情報処理学会山下記念研究賞受賞．情報処理学会，日本音響学会，電子情報通信学会各会員．

荻野 紫穂: 1986 年東京女子大学文理学部日本文学科卒業．1988 年同大学院文学研究科修士課程修了．同年，日本アイ・ビー・エム（株）入社．東京基礎研究所に勤務．現在，音声認識システムの研究開発に従事．情報処理学会，人工知能学会，計量国語学会各会員．

山崎 一孝: 1988 年東京工業大学工学部情報工学科卒業．1990 年同大大学院総合理工学研究科システム科学専攻修士課程修了．1993 年同大大学院理工学研究科情報工学専攻博士課程修了．工学博士．同年，日本アイ・ビー・エム（株）入社．東京基礎研究所に勤務．文字認識，音声認識の研究および製品開発に従事．電子情報通信学会会員．

(1998 年 4 月 1 日 受付)
(1998 年 7 月　日 再受付)
(1998 年 8 月　日 採録)