自然言語処理 131－5
音声言語情報処理 26－5
（1999．5．28）

日本語ディクテーションのための言語資源・ツールの整備

伊藤克亘 (電総研)　山田篤 (ASTEM)　天白成一 (アルカディア)　山本俊一郎
踊堂憲道　宇津呂武仁 (奈良先端大)　山本幹雄 (筑波大)　鹿野清宏 (奈良先端大)

我々は、音声認識のための共通のソフトウェアリポジトリを開発する3ヶ年計画 (1997～1999年度) のプロジェクトにおいて、言語モデル構築と、そのための各種言語資源・ツールの整備を行っている。音声認識に用いられる言語モデル構築のためには、対象とする分野の日本語テキストを単語に分割する必要があるが、我々はこの処理のために、日本語形態素解析システム茶筌を採用しその形態素解析辞書の整備を行った。さらに、音声認識の観点から日本語テキストに読みを付与するためのガイドラインを設定し、形態素解析結果に対して正しい読みを付与するツールを作成した。また、JNAS新聞記事読み上げコーパスをもとに、読み表記のガイドラインに基づく読み正解コーパスを作成し、音声コーパスの整備を行った。

Text Processing for Japanese Dictation System: Resource and Tools

Katunobu Itou (ETL)　Atsushi Yamada (ASTEM)　Seiichi Tenpaku (Arcadia)
Shunichiro Yamamoto　Norimichi Yodo　Takehito Utsuro (NAIST)
Mikio Yamamoto (Tsukuba University)　Kiyohiro Shikano (NAIST)

As a part of the research and development activities of three years (1997 ~ 1999) project for developing common software repository for speech recognition, we have been developing various kinds of linguistic resource and tools for constructing language models of Japanese dictation systems. As a tool for segmenting Japanese texts of the target domain into sequence of words, we adopted the Japanese morphological analysis system ChaSen, and furthermore, developed morphological analysis dictionary tailered for our purpose. We also drew up a guideline for annotating pronounciations to Japanese texts from the viewpoint of speech recognition and developed tools for annotating appropriate pronounciatinos to the morphologically analyzed texts. From the JNAS newspaper article read speech corpus, we developed a text corpus annotated with appropriate pronounciations which follow the above guideline.

1　はじめに

大語彙連続音声認識は、音声を利用した様々なアプリケーションの基盤となる技術である。この実現のためには、高精度の音響モデル、高精度の言語モデル、効率のよい認識エンジンが、バランスよく統合化される必要がある。我々は、音声認識のための共通のソフトウェアリポジトリを開発する3ヶ年計画 (1997～1999年度) のプロジェクトにおいて、言語モデル構築と、そのための各種言語資源・ツールの整備を行っている。

音声認識に用いられる言語モデル構築のためには、語彙とその読みが重要となる。日本語においては、語彙の定義が形態素解析システムとそこで用いられる辞書に依存するため、我々は奈良先端科学技術大学院大学で開発されている茶筌 [4] を採用し、その形態素解析辞書の整備を行った。

また、音声認識に用いられる単語辞書では当該エントリの実際の読み方 (音素記号列で表される) が必要であるのに対し、これまで辞書を含めた言語資源という観点からは読みに相当する情報が現代かな遣いに基づくかな表記で表現されることが多かっ

た。そこで、我々は、読み表記のためのガイドラインを設定し、それに基づいて各種言語資源の整備を行った。

さらに、形態素解析結果をもとに正しい読みを付与するツールを作成した。

2 言語・音声資源の整備

我々は、一般全国紙の1つである毎日新聞の記事データを共通の言語・音声コーパスに採用している。また、JNAS新聞記事読み上げコーパス (以下、JNAS) を音響モデルの学習データ (の一部) と認識システムの評価データに用いている。音声認識という観点からは、言語資源としても読みに注目する必要があるため、読み表記のガイドラインを設定し、これに基づいて読み正解コーパスを作成した。さらに、この結果をもとに音声コーパスの整備を行った。

2.1 読み表記ガイドラインの設定

我々は、実際の発声に近い読みの表記法として、NHK日本語発音アクセント辞典 [2] に準拠したカタカナ表記を採用し、正解の基準も同辞典に準ずることとした。ただし、鼻音、無声化、およびアクセントに関する情報は取り扱っていない。これは、「チ」「ツ」の濁音をそれぞれ「ジ」「ズ」と表記したり、長音の表記に「ー」を採用しており、現代かなづかいに基づいた一般的な表記とは異なる。

また、一般に日本語の漢字かな混じり文では、「今日」のように同じ表記が複数の読みの可能性をもつことがある。このとき、ディクテーションを目的とすると可能な読みはすべて付与されることが望ましい[1]。このため、読みの併記方式を採り入れ、複数の読みが可能な場合には、それらを併記することとする。なお、今回は、併記された読みの間の尤度差は考慮せず、五十音順に並べることにした。更に、連接する形態素にともに読みが併記される場合に、それらの間の組合せに一定の制約が課せられることがある。このような場合には、それらの形態素を結合して一つの形態素とし、あり得ない読みの組合せが生じないようにする。

2.2 読み正解コーパスの整備

設定した読み表記ガイドラインに従って、JNASの読み表記を整備した読み正解コーパスの整備を行った。JNASには、読み上げ対象文に対して、現代かな遣いによってルビをふった文書 (LaTeX形式) が含まれている。これをもとにして、その表記を基本的に人手で読み表記ガイドラインに沿ったものに修正していった。作業手順は以下のとおりである。

1. LaTeX形式の文書から、表記とルビのペアを抽出する。
2. 全出現形態素のリストを作成する。
3. 上記リストの読みを修正する。
4. 修正リストをもとに元のルビを修正する。
5. 修正したルビを人手でチェックする。

このとき、文脈によって複数読みの可能なもの、一意に読みの定まるものはそれを読みに反映させた。また、全体を通しての形態素区切りの正規化も行った。

2.3 評価セット

JNASを認識システムの評価用データとして使用するために、約3万文から成る新聞記事読み上げ音声から標準的な評価用セットを定めた。不特定話者認識の評価用とするために、まず評価用話者のセットを定義し、評価用話者の発声した音声データから評価用文を選択した。

評価用話者としてJNASの話者306名から94名 (男女各47名) を選択した。評価用話者が読み上げた文セットは47 (約100文/セット) あり、各セットを評価用話者の男女各1名が読み上げている。これは、最終的な評価用文を男女各1名が必ず読んでいるようにしたかったために行った。さらに、94名の評価用話者は男女各23名計46名と、男女各24名計48名の2つのセットに分割した。最初の46名から以下に説明する評価用セットを作成した。残りの48名は今後のための予備話者である。

評価セットは上記の46名の話者集合から選択された文から構成される。5種類の語彙規模に対して、男女各100文の計1000文である。各語彙規模の文は、ベースとなった新聞記事の文長・複雑さ (perplexity) の分布とおおよそ等しくなるように設計した。また、男女はまったく同じ文を読み上げている。

評価用データであるため、転記と発声が食い違っている文は適さない。選択した文の音声チェックを行い転記とずれのあるものは、文長・複雑さ・語彙規模が等価な他の文と入れ替えた。この際、男女共に同じ評価文とするために、男女共に入れ替えた。

より詳しくは [1]、評価セットのデータは以下の

[1] テキストの読み上げなどを目的とする場合は、このうち、特定の読みが選択される必要がある。また、ディクテーションに対しても、文脈によって読みが特定される場合は、可能な読みの種類を限定する何らかの仕組みが必要となる。

URLを参照のこと。
http://www.milab.is.tsukuba.ac.jp/jnas/test-set

2.4 転記の整備

JNASはもともと多くの機関がボランティアで作成したコーパスであるため、発声および転記の精度にはかなりのばらつきがある。JNASの音声データは新聞記事を読み上げたものであり、転記は読み上げられたテキストをそのまま用いている。このため、多くの問題は読み間違えたために生じた、転記と音声データのずれである。本プロジェクトでは、JNASの転記の整備を行っている。

今年度末(99年度末)の目標は、新聞記事読み上げ音声(306名約3万文)に対して、(1)ルビと音声を一致させる、(2)日本語文としておかしな音声を排除(軽微な誤りは排除しない)することである。JNASの新聞記事読み上げ音声は約3万文と多く、また実働可能な作業者数にも制限があるため、詳しいチェックはあきらめて、音声とルビとの一致を最大の目標としている。

実際の作業は、2.2で述べたJNASの読み上げテキストに対する読みの標準表記を元に作業を行っている。読みのテキストを見ながら音声を聞き、以下のような作業を行っている。

1. ルビに複数の選択肢がある場合はどれかに○をする。
2. ルビと音声が異なる場合、その音声が日本語として妥当かどうか判断する。妥当であればルビを修正。妥当でなければ読み誤りとして排除。
3. 長・短音化、無・有声化等の問題は目立った場合のみマークする。

日本語として妥当かどうかの判断は主観的に行っているが、誤りが明白なもの以外は妥当とする方針で行っている。

現在のところ、306名中192名のチェックが完了しているが、作業者によって異なるチェックのレベル調整、電子化や分析は今後の課題である。

3 形態素解析辞書の整備

日本語においては、語彙の定義が形態素解析システムとそこで用いられる辞書に依存するため、我々は奈良先端科学技術大学院大学で開発されている茶筌[4]を採用し、その形態素解析辞書の整備を行った。この際、活用体系・接続規則などの文法の一部に対しても手を加えた。作業にあたっては、1998年5月にリリースされた、茶筌1.51とipadic1.0b2を用いた。

3.1 辞書エントリの整備

ipadic1.0b2は、従来の茶筌の辞書をIPA品詞体系に機械的に移行したものであったため、各種の不具合があった。我々はipadic1.0b2に含まれる形態素解析辞書に対して、以下の作業を行った(以下では、ベースとなった辞書をipadic1.0b2、我々が整備した辞書をipadic1.0と呼ぶ)。

3.1.1 辞書のエントリの追加と削除、マージ

ipadic1.0b2から不要と考えられるエントリを削除した。このとき、形態素の粒度の統一を心がけた。また、従来の茶筌の辞書で採用されていた品詞体系との差異が原因となって、適切でない品詞分類が与えられているものに対しては、その変更を行った。

一方、新聞記事に瀕出する固有名詞に対処するため、人名や地名、組織名などを中心に新たにエントリを追加し、解析能力の向上を図った。エントリの追加にあたっては、郵政省の郵便番号簿など、ネットワークを通じて入手可能なフリーデータを活用した。

さらに、同一表記同一品詞であるにもかかわらず、読みが異なることによって複数エントリにわかれていたものについては、それらをマージし、読みを併記した上で単一のエントリとした。例えば、「貴い」に対して「タットイ」「トートイ」という2種類の読みが存在する場合がこれにあたる。この結果、原型と品詞の組に対して、常に唯一のエントリが対応することになった。

整備前後の品詞分類毎のエントリ数の比較を表1に示す。

3.1.2 辞書の読み表記の修正

辞書中のすべての形態素の読み表記を、読み表記ガイドラインに沿ったものに修正した。この結果、辞書中の各エントリには、可能な読みがすべて記載されることとなった。その中から、前接語や後接語、さらには文脈に応じて、適切な読みを選択する仕組みとして、形態素解析結果に対する読み付与ツールを想定している。これについては4章で述べる。

3.1.3 形態素コストの調整

解析済み正解コーパス(RWCP毎日新聞テキストコーパス[3]の人手修正済3000記事)を用いて、辞書中に持つ各エントリのコスト学習を行った。ただし、新たに追加したものを中心に、解析済み正解コーパスに現れない形態素についてのコストは未学習(一律200)である。これらについては、今後、

表 1: 辞書エントリ数の比較

	ipadic1.0b2	ipadic1.0
名詞 一般	69,296	58,438
名詞 固有名詞 一般	36,262	26,404
名詞 固有名詞 人名	6,895	32,147
名詞 固有名詞 組織	3,055	18,084
名詞 固有名詞 地域	2,528	67,420
名詞 代名詞	97	96
名詞 副詞可能	958	778
名詞 サ変接続	11,499	9,420
名詞 形容動詞語幹	3,251	2,970
名詞 数	73	41
名詞 その他	1,484	1,387
接頭詞	260	221
動詞	11,947	14,450
形容詞	1,309	1,651
副詞	2,472	2,894
連体詞	159	137
接続詞	174	150
助詞	170	181
助動詞	35	35
感動詞	195	207
記号	194	128
その他	4	3

正解コーパスの整備、コストのスムージング処理などによって対応を図る予定である。

3.2 活用体系・接続規則の整備

ipadic1.0b2で用いられていた活用体系・接続規則に対して、読み付与の観点から以下の整備を行った。

3.2.1 「う」「よう」の活用語尾への移行および長音化

活用語に助動詞の「う」が後続する場合、その読みは長音化される。ところが、ここに形態素の切れ目があると、長音のみの形態素が生じるため、現在の枠組みでは認識時に不都合を引き起こす。このため、助動詞-不変化型の「う」「よう」を活用語の活用語尾に追加して、意志形または推量形という活用形を追加し、「う」「よう」の部分の読みを長音化することとした。「よう」を追加するのは、「う」と同様の扱いにするためである。

3.2.2 接続コストの調整

ipadic1.0b2で用いられていた接続規則に対して、接続が存在しないか、または誤った接続が学習されていた以下の項目について、形態素コスト学習で用いた正解コーパスを修正もしくは新たに正解例を収集し、接続コストを学習した[2]。接続コスト学習の際は、bi-gramマルコフモデルのもとで連接確率を最尤推定し、これを接続コストに変換した。なお、接続コスト学習は、次の品詞分類の単位で行った。

- 品詞分類は、助詞以外は、最も細かい品詞レベル
- 助詞は語彙レベル
- 活用語は、活用型・活用形まで区別

「う」「よう」の活用語尾への移行 「う」「よう」の活用語尾への移行に関連して、接続コスト・形態素コスト学習用コーパスで対応する部分を変換し、接続コストの学習を行った。

接続助詞「ちゃ」の用例の作成・接続規則学習 接続助詞「ちゃ」に関する接続が学習されていなかったため、用法が同じと考えられる「ては」(接続助詞の「て」、係助詞の「は」)の解析例を前後の形態素の品詞パターンが網羅される程度にまで収集した。次に、これらの用例の「ては」を「ちゃ」に置き換え、その中で文として成立するものを正解例として、接続コストの学習を行った。

仮定縮約1形、仮定縮約2形の用例の作成・接続規則学習 仮定縮約1形、仮定縮約2形とは、「—りゃ」「—きゃ」などの口語的な表現である。これらの接続も学習されていなかったため、以下の作業を行った。

- **前接形態素**
  前接形態素については、活用形の違いを無視してよいので、各活用型について可能な全ての前接品詞パターンを網羅するように用例を収集し、活用形を「仮定縮約1形」「仮定縮約2形」に置き換えてみて、文として成立するものを正解例として蓄積する。
- **後接形態素**
  用法としては、未然形+接続助詞の「ば」と同じなので、後接形態素のパターンとしても、接続助詞の「ば」に後接する品詞パターンを網羅する程度に接続助詞の「ば」の用例を収集し、活用形を「仮定縮約1形」「仮定縮約2形」に置き換えてみて、文として成立するものを正解例として蓄積する。

前接形態素部分およびそれより前と、後接形態素部分およびそれより後を適当に組み合わせて文を作成

[2] 以下では、多くの場合、正解例を機械的に作成しているが、これは暫定的措置であり、将来的には、生テキストから対応する用例を収集し、これを人手で解析することにより正解例を収集することが望ましい。

した。ただし、特定の形態素が頻出しないように、同一品詞の語で適当に置き換えた。その後、接続コストの学習を行った。

助動詞「です」の未然形「でしょ」、「ます」の未然ウ接続形「ましょ」の接続規則の暫定的整備　これらについても、学習用コーパスでは、用例が不足しており、接続規則が学習できていなかったため、以下のように接続規則を機械的に追加することで対処した。

- 「でしょ」「ましょ」の前接規則
  「です」「ます」の全活用形の前接規則を流用する。
- 「でしょ」「ましょ」の後接規則
  「記号」との接続規則のみ記述する。

4　読み付与ツールの整備

形態素解析辞書の読みを整備したことにより、形態素解析結果に付与される読みは、人手で整備した読み正解コーパスのものに近付いたが、単独の形態素レベルでは解決しない問題として以下のものが残った。

1. 数詞・助数詞に対する読み
2. 同型異音語の読み分け
3. 連濁
4. その他

このうち、高頻度で出現し、バリエーションも多いが、規則適用により対処可能な数詞・助数詞を中心にして、形態素解析結果の読みのつけ換えを行うツールを作成した。同形異音語のように、文脈による読み分けが必要な現象に対しては、同じく形態素解析の後処理ツールによって今後対応する予定である。

4.1　数詞・助数詞の読み変化

茶筌では、数字は1文字ずつのエントリになっているため、たとえば、「１６本」に対しては「イチ|ロク|ホン」という結果が返される。これら数詞・助数詞に対する読み付与パターンを整理し、ツール化した。

日本語の数詞の読み変化は、位取りを行うものと、行わないものの大きくは2通りがある。本ツールでは、以下に示す基準により、これを自動的に判別するようにした。また、数詞と助数詞が連接することで、一部の読みが音便化したり助数詞の先頭の音節が半濁音や濁音に変化する現象にも対応した。

4.1.1　数詞の表記方法と読みの関係

日本語の文章においては、数字の表記形態が多数存在することは周知の事実である。これは、従来より縦書きで漢数字を用いて記述されていたものが、明治以来、横書きでアラビア数字を用いて記載されるに至った歴史的な経緯も遠因である。日本語の文章中にみられる数字の表記方法を、文字の種別で類別すると、大きく次の3つに分類することができる。

- アラビア数字のみ　　例: 198000
- 漢数字のみ　　例: 十九万八千
- アラビア数字と漢数字の混在　　例: 198千

この他にローマ数字による表記が与えられる場合があるが、ローマ数字による表記に必要な文字コードが一部の計算機システムでは、外字コードとして扱われているため、ローマ数字を用いた数字表記に対する読み付与の処理は対象外とした。

これらに対し、数詞のみで構成されるときは、先頭の文字が「〇 (ゼロ)」以外の文字から始まる場合は、原則として位取りして読むことにした。ただし16桁 (千兆) をこえた場合は、位取りはしない。

4.1.2　数字表記における特殊記号とその読み方

数字表記におけるもう一つの大きな課題は、特殊記号の取扱いである。数字を表現するにあたって、数字だけではなく、特殊記号と合わせて使用することにより、意味を付加することがある。数字表記における特殊記号として、以下の場合に対処した。

(1) 数字表記中に特殊記号「, (こんま)」が入る場合　「, (こんま)」が入る場合は、読み上げには影響しない。これは、数字の表記における欧米式の書式であり、通常のように位取りをして読む。なお、欧米式の数字表記では「,」は3桁毎に入るが、それ以外の箇所に入っている場合も、読み付与の観点からは無視する。

(2) 数字表記中に特殊記号「. (てん)」が入る場合　「. (てん)」が1個だけの場合は、小数点とみなし、位取りして読み、小数点以下は、位取りせずに読む。ただし、「２．５千」のように位を表す漢数字と混在している場合には、位取りを全体に施してから読む。一方、「. (てん)」が2個以上の場合は、位取りせずに読む。

(3) 数字表記中に特殊記号「・(なかてん)」が入る場合　「・(なかてん)」は、「. (てん)」と同様に扱う。

(4) 数字表記中に特殊記号「ー(ばー)」が入る場合　「ー(ばー)」または「—(ばー)」が入る場合に、数字は原則として位取りし、当該記号は「マイナス」と読む。表記が、数前置詞の郵便記号(〒)から始まる場合か、地名からの後処理として、数字を読む場合に限って、「ノ」と読む。

(5) 数字表記中に特殊記号「() (かっこ)」が入る場合　「() (かっこ)」が入る場合には、二通り考えられる。一つは、数字を単に括弧で括ったもの、もう一つは、電話番号等の表記である。「() (かっこ)」に数字が後続すれば、電話番号と考え位取りせず、後続しなければ位取りして読む。

4.1.3　数詞と助数詞が連接した場合の読み変化

数詞の読み方は、後続する助数詞によって変化する。日本語発音アクセント辞典[2]をもとに、これを分類したものを表2に示す。

また、数詞と助数詞の組み合わせによっては、助数詞の先頭の読み方が変化する場合がある。この現象は、通常の単語においては、連濁と呼ばれるような変化であるが、数詞と助数詞の組み合わせでは、単純な連濁の他に、複数の読み変化をとるパターンが存在する。例えば、「本(ホン)」では、「一本(イッポン)」、「二本(ニホン)」、「三本(サンボン)」というように読みが変化する。これに対応するためには、助数詞の方のデータとして、連濁するかどうかの情報を蓄えておく必要があり、これらの情報を一元的に管理するようにした。

4.1.4　数詞および助数詞における読み併記

読み表記ガイドラインに従って読みを併記して出力させる場合に、数詞と助数詞に関しての読み付与においては、助数詞との間で読みが複数考えられるものは、併記して出力させることとした。また、場合によっては、読み併記に伴い、数詞と助数詞を分離せずに出力させる必要が生じる。

［例］一日　{ツイタチ/イチニチ}
　　一　{イチ/イッ}
　　食　ショク

読み併記を行う場合には、表2に示した数詞の読み変化パターンに対して以下のルールを適用する。

1. A型変化において助数詞の先頭がp,t,k,sの場合、{イチ/イッ}、{ハチ/ハッ}になる傾向が高い。
2. A型、B型変化において全て{シチ/ナナ}になる。
3. A型、B型変化においてA1型の標準的な読み方から変化するものは、すべてA1型を許容する。
4. 「ジュッ」の読みでは「ュ」が脱落して、「ジッ」となりえる。
5. C型、E1型変化では、2桁以上の読みは、A型で変化する。

この他に読みを併記する場合として、小数点を含む場合に対応した。

［例］８０・６　ハチ{ジュッ/ジッ}テンロク
　　９０・６　キュー{ジュッ/ジッ}テンロク

4.2　その他の読み修正

数詞以外に前接形態素によって読みが変化するものとして「々」「々々」がある。また、現在の品詞体系、活用型・活用形体系で処理が不可能なものとして活用語の語幹の読みが変化する場合がある。このうち、カ行変格活用動詞「来る」のように、特殊な活用型が予め割り当てられているものは、その活用形体系の中に語幹の読み変化まで含まれているため、この限りではない。以下で、これらの取り扱いについて述べる。

4.2.1　「々」「々々」に対する読み付与

「々」および「々々」は前接する形態素によって読みがかわるため、これが独立の形態素として出現した場合は、前接する形態素の読みに応じて読みを付与する必要がある。このとき、読みによっては連濁を施す必要のあるものがある。

［例］神　カミ
　　々　ガミ

4.2.2　動詞「言う」の読み

「言う」は五段・ワ行促音便の動詞として分類されている。このとき、その読みは「イワ、イー、ユウ、イエ、イオ、イッ」のように変化するため、語幹「言」に対して一意な読みを割り当てることができない。そこで、辞書ではこの語幹に対しては「イ」という読みを割り当て、「言う」の場合のみ、後処理でこれを「ユウ」に修正することにした。

表 2: 数詞の読み変化

分類	一	二	三	四	五	六	七	八	九	十
A1	イチ	ニ	サン	ヨン	ゴ	ロク	ナナ	ハチ	キュー	ジュー
A2	イチ	ニ	サン	ヨン	ゴ	ロク	ナナ	ハチ	キュー	ジュッ
A3	イッ	ニ	サン	ヨン	ゴ	ロク	ナナ	ハチ	キュー	ジュッ
A4	イチ	ニ	サン	ヨン	ゴ	ロッ	ナナ	ハチ	キュー	ジュッ
A5	イッ	ニ	サン	ヨン	ゴ	ロク	ナナ	ハッ	キュー	ジュッ
A6	イッ	ニ	サン	ヨン	ゴ	ロッ	ナナ	ハチ	キュー	ジュッ
A7	イッ	ニ	サン	ヨン	ゴ	ロッ	ナナ	ハッ	キュー	ジュッ
B1	イチ	ニ	サン	シ	ゴ	ロク	シチ	ハチ	ク	ジュー
B2	イチ	ニ	サン	ヨン	ゴ	ロク	シチ	ハチ	ク	ジュー
B3	イチ	ニ	サン	ヨン	ゴ	ロク	シチ	ハチ	キュー	ジュー
B4	イチ	ニ	サン	ヨ	ゴ	ロク	シチ	ハチ	ク	ジュー
B5	イチ	ニ	サン	ヨン	ゴ	ロク	ナナ	ハチ	ク	ジュー
B6	イチ	ニ	サン	ヨン	ゴ	ロク	ナナ	ハチ	キュー	ジュー
C1	ヒト	フタ	ミ	ヨ	イツ	ム	ナナ	ヤ	ココノ	ト
C2	ヒト	フタ	ミ	ヨ	イツ	ム	ナナ	ヤ	キュー	ト
C3	ヒト	フタ	ミ	ヨ	イツ	ム	ナナ	ハチ	キュー	ト
C4	ヒト	フタ	ミ	ヨ	イツ	ム	ナナ	ハチ	キュー	ジュー
C5	ヒト	フタ	ミ	ヨ	ゴ	ム	ナナ	ハチ	キュー	ジュッ
C6	ヒト	フタ	サン	ヨン	ゴ	ム	ナナ	ハチ	キュー	ジュー

4.2.3 形容詞・アウオ段・連用ゴザイ接続の読み

形容詞・アウオ段「長い」に「ございます」を後接すると、表記は「長うございます」となるが、その読みは「ナゴーゴザイマス」である。すなわち、「長い」の語幹部分「長」の読みが「ナガ」から「ナゴ」に変化する。これも現在の品詞体系、活用型・活用形体系では処理できないので、辞書では「ナガ」という読みを割り当てておき、連用ゴザイ接続の場合にこれを「ナゴ」に修正することにした[3]。

4.2.4 動詞・五段・カ行イ音便・連用タ接続の読み

例えば「聞く」という五段カ行イ音便の動詞の連用タ接続は、「聞いた」となるが、その読みは「キータ」というように長音化する。このとき、「キイ」から「キー」への修正を行う。

4.3 数字表現の位取り正規化

形態素解析では、数字表現は全体で一つの形態素となるか、一文字づつ分解されることが多い。本読み付与ツールの出力では、全体で一つの形態素となっている。これをそのまま一語として計量してしまうと、数の種類だけ語の種類が増え、無数に形態素の種類が増えることになる。また、4.1.1 で述べたように、数字の表記方法には様々な方法があるが、表記の違いで語を区別すると、その分だけ語の種類が増えることになる。

そこで、言語モデルを作る際には、数字表現について表記を漢数字に統一し、読みが一意に決まる程度の単位 (位取り) に分割する処理をおこなった。たとえば,「３６．３３」という数字は、「サンジュー|ロク|テン|サン|サン」と読むので、以下のような正規化を行なう。

三十 サンジュー
六　ロク
・　テン
三　サン
三　サン

5 評価

本稿で報告した処理の成果を評価するため、オリジナルの ipadic1.0b2 と改良後の ipadic1.0 を用いて、毎日新聞 91 年 1 月から 94 年 9 月までの計 45 か月分の形態素解析結果を用いて、テストセットパープレキシティと音声認識率で評価した。

5.1 言語モデルの作成

学習データ 45 か月分から、文章でない記事や段落、読み上げに適さない表現などを削除すると、総文数は、2,352,471 となった。これらを形態素解析し、「形態素、読み、原形、品詞」の組で形態素を区別し、言語モデル構築の際の単位とした。以後、この単位を単語と呼ぶ。

2 種類の形態素解析結果の比較結果を表3 に示す。

単語被覆率などについては、大幅な改善は見られなかったが、頻度上位 20000 語語彙に含まれる未定義語の種類数は、122 から 34 に減少しており、固有名詞を追加した効果があらわれている。

この形態素解析結果を学習データとして、言語モデルを作成した。語彙サイズは 20000 語とし、前

[3] 将来的には、品詞体系、活用型・活用形体系を詳細化し、後処理をなくすことが望ましい。

表 3: 整備前後の形態素解析結果の比較

	ipadic1.0b2	ipadic1.0
総単語数	66,548,749	66,236,904
異なり単語数	164,280	190,975
5k 被覆率 (%)	88.1	88.2
20k 被覆率 (%)	96.6	96.5
未定義語数 (全体)	43,603	56,692
未定義語数 (20k)	122	34

向き単語バイグラムと逆向き単語トライグラムを作成した。カットオフは、バイグラムで 1、トライグラムで 2 とした。この 2 種類の言語モデルについて、JNAS 男性話者 20000 語彙評価セット 100 文 [1] と、毎日新聞 94 年 10 月から 12 月の 3 か月分の二つのテストセットに対して、テストセットパープレキシティを求めた結果を表 4 に示す。

表 4: 整備前後のテストセットパープレキシティの比較

	ipadic1.0b2		ipadic1.0	
	JNAS	毎日	JNAS	毎日
バイグラム	55.3	72.5	55.5	71.6
トライグラム	37.7	49.1	37.8	48.5
未知語	9	210.409	7	213,146

テストセットパープレキシティにおける両者の差はほとんどない。

これらの言語モデルを用いて、JNAS 男性話者 20000 語語彙評価セット 100 文 [1] の認識実験を行ない、音声認識率を比較した。認識実験には、IPA のディクテーションツールキットの Julius と男性話者 triphone HMM (総状態数 2000、16 混合分布) を用いた。認識率の比較結果を表 5 に示す。正解率と正解精度の値は、形態素を単位として自動算出した [6]。

表 5: 整備前後の音声認識率の比較

	ipadic1.0b2	ipadic1.0
Corr/Acc	87.8/86.2	91.2/89.6

単語誤り率は約 3 ポイント減少した。これは、例えば「は」などの品詞によって読みが変わるエントリの読み分けを的確に行なったこと、併記形式を用いることにより、同一品詞の語にも複数の読みを出力できるようになったことが大きく寄与している。また、認識対象に含まれる未知語が 2 語減ったことも影響していると思われる。認識対象に未知語があるとその周辺の単語も含めて認識結果に悪影響を及ぼすからである。

6 おわりに

本稿では、大語彙連続音声認識に用いられる言語モデル構築のために行った各種言語資源・ツールの整備について述べた。ここで述べた言語資源・ツール類は日本語ディクテーション基本ソフトウェア—1998 年度版—に含まれて、一般に公開されている。ここで述べた以外に、言語モデル圧縮ツール [5] や単語正解率判定ツール [6] なども含まれている。また、本プロジェクトで整備した形態素解析用辞書は ipadic1.0 として、茶筌 2.0b6[4] とともに公開されている[4]。今後は，本稿で述べた言語資源や各種ツール類の整備を更に進めていく予定である。

謝辞: 本プロジェクトは情報処理振興事業協会 (IPA) の「独創的先進的情報技術に係わる研究開発」の支援を受けている。また、日本語形態素解析システムとして、奈良先端科学技術大学院大学で開発されている茶筌を利用している。さらに、言語・音声資源として毎日新聞社による CD-毎日新聞データ集、また技術研究組合新情報処理開発機構による RWC テキストデータベース、音響学会データベース委員会による JNAS 新聞記事読み上げコーパスを利用している。これらを提供し、また利用を許可して下さり、ご協力頂いた関係各位に感謝する。

参考文献

[1] 萬崎弘、山本幹雄、板橋秀一:「日本音響学会新聞記事読み上げ音声コーパスからの評価用文セットの作成」、日本音響学会講演論文集 (I)、1-R-13, pp.143-144 (1998).

[2] NHK 放送文化研究所: NHK 日本語発音アクセント辞典 新版 (1998).

[3] データベースワークショップ テキストグループ: テキストデータベース報告書, 技術研究組合 新情報処理開発機構 (1995).

[4] 松本裕治, 北内啓, 山下達雄, 平野善隆: 日本語形態素解析システム『茶筌』version 2.0 使用説明書, Information Science Technical Report NAIST-IS-TR99008, 奈良先端科学技術大学院大学 (1999).

[5] Norimichi Yodo, Kiyohiro Shikano, Satoshi Nakamura: Compression Algorithm Of Trigram Language Models Based On Maximum Likelihood Estimation, Proc. of ICSLP 98, pp.716–719 (1998).

[6] 山本 俊一郎、伊藤 克亘、鹿野清宏、中村 哲: ディクテーションにおける日本語の特質を考慮した単語正解率判定ツール, 日本音響学会講演論文集, pp. 155–156 (1999).

[4] URL:http://cl.aist-nara.ac.jp/lab/nlt/chasen.html