音声言語情報処理 29－2
(1999. 12. 20)

〔特別講演〕 日本語ディクテーションシステムの現状と今後の課題

西村 雅史

日本アイ・ビー・エム (株) 東京基礎研究所
〒242-8502 神奈川県大和市下鶴間 1623-14, LAB-S77
e-mail: nisimura@trl.ibm.co.jp

あらまし

大語彙音声認識のアプリケーションの1つとして，ディクテーションソフトが日本の PC 市場に登場してからほぼ3年が過ぎた．「読み上げ」に分類される丁寧な発話に限ればソフトウェア製品として，一応完成の域に達した感もある．ここでは IBM のシステムを例として，その基本的な構成について述べるとともに，ディクテーションシステムの使いやすさを改善するために行ってきた細かな機能上の改良についても紹介する．一方，応用範囲をさらに広げて行くためには解決すべき課題もまだ多い．その中でも「自然で自由な発話」の書き起こしは最も重要な課題の一つである．これに対処できれば，講演の書き起し，会議の議事録作成，放送音声一般に対する字幕付与や情報検索など，さらに広い分野での応用が見えてくる．ここでは自由発話の書き起しに向けた取り組みの一部を紹介する．また，この自由発話の書き起しを前提とした，音声理解に関する取り組みについても簡単に触れる．

キーワード ディクテーション，大語彙連続音声認識，自由発話，音声理解

Japanese dictation system for now and the future

Masafumi NISHIMURA
IBM Research, Tokyo Research Laboratory, IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi, Kanagawa-ken 242-8502, Japan
e-mail: nisimura@trl.ibm.co.jp

Abstract

It is almost three years since the first shrink-wrapped dictation software for Japanese was announced. Various dictation products are now available in Japan, and have broad-based user acceptance. As a result of the rapid maturation of automatic speech recognition (ASR) technology, the type of "read speech" used for dictation no longer seems to be a target for ASR research. Instead, researchers are now working on large real-world problems such as "spontaneous speech". If ASR can be used to transcribe spontaneous speech precisely, its application areas will be expanded dramatically. In this paper, we describe the present IBM dictation system in detail, then give the preliminary results of our first attempt at spontaneous speech recognition. We also introduce our latest efforts in the area of speech understanding.

key words dictation, large-vocabulary continuous speech recognition, spontaneous speech, speech understanding

1. はじめに

一般の PC ユーザーが日本語ディクテーションソフトに触れるようになってほぼ 3 年がたつ．当初離散単語発声が必要であったが，音声認識技術ならびにハードウェアの急速な進歩にも助けられ，'97 年末には現行のディクテーションソフトと同様，連続発声が可能な市販ソフトが登場した．ユーザーのアクセプタンスも高く，少なく見積もっても既に 100 万本以上のディクテーションソフトが日本市場に出回ったと言われている．単体の製品だけでなく，今では多くのメーカー製パソコンに同梱あるいはプリインストールされるようになっているので，その数は現在も爆発的に増え続けている．このように，大語彙連続音声認識の研究はディクテーション(あるいは音声ワープロ)という一つのアプリケーション分野においてビジネス的に大きな成功を収めることが出来た．個人ユーザーだけでなく，医療所見の入力業務[1]などにも使われているし，ニュース音声の字幕付与といったシステム[2]も実用化されつつある．

これにともない研究の重点は，ディクテーションの対象とされる「読み上げ」のような丁寧な発話から将来の音声対話システムの実現に向けた自然な発話の認識や理解へと移行している．US では ARPA がスポンサーとなって行っている Hub-4 や Hub-5 などのプロジェクト[3]が有名であるが，日本でも ATR が相当量の対話コーパス[4]を完成させているし，また国立国語研究所等が講演などの発話を中心に，大規模な話し言葉コーパス[5]の収集作業を開始するなど今後の成果が期待されている．

本稿ではディクテーションの技術が音声認識/理解といった大きな枠組みの中でどのような役割をになっているのかについて概観した後，IBM のディクテーションシステムを例にして，その基本構成および特徴的な機能について紹介する．また，ディクテーションの次の課題として，自然な発話の書き起しに関する我々の取り組みについても紹介する．

2. 統計的音声認識/理解システムの原理

2-1. 統計的音声認識システム

現在のディクテーションシステムの多くは，図 1 に示すような原理に基づいて設計されている(S を入力として W を復号化するまでの部分)．発話者は伝えようと意図した内容を頭の中で文書化し(W)それをマイクに向かってしゃべる．ディクテーションシステムはこの音声信号(S)を受け取り，周波数分析などの音響処理を行い特徴量に変換する．これにより，発話者が発声した文書はいくらかの情報を失う．次に受け取った情報(X)から事後確率最大となる単語列(最尤解) $\hat{W}$ を次式にしたがって推定する(言語復号)．

$$\hat{W} = \arg\max_{W} \Pr(W \mid X) \qquad \cdots(1.1)$$

$$= \arg\max_{W} \Pr(X \mid W)\Pr(W) \qquad \cdots(1.2)$$

ここで，$\Pr(X|W)$は音響モデル，$\Pr(W)$は言語モデルで，それぞれ隠れマルコフモデル(HMM)および N-gram モデルを使って表現される．ここで大切な点は，モデルは 2 つに分割されているが，確率という共通の尺度で 2 つのモデルが統合されていることである．つまり，まず音響特徴量の列を平仮名列に変換し，これを仮名漢字変換しているのではなく，音響特徴量の列から漢字仮名混じり文へ直接変換しているのである．より一般的にいうなら，モデルは複数に分割されているが，最尤解の探索に際しては，すべてのモデルの確率値が統一的に考慮されるのである．

2-2. 統計的音声理解システム

ディクテーションの枠組みを超えてしまうが，図 1 に示すように，この原理を発話者の意図(I)を理解するという枠組みに拡張することが可能である[6]．だが現状では意図の表現方法すら決まったものはなく，図 2 のような構文木(T)を推定することで音声理解の一助とすることが多い．その場合には次式で表されるような言語モデルを使

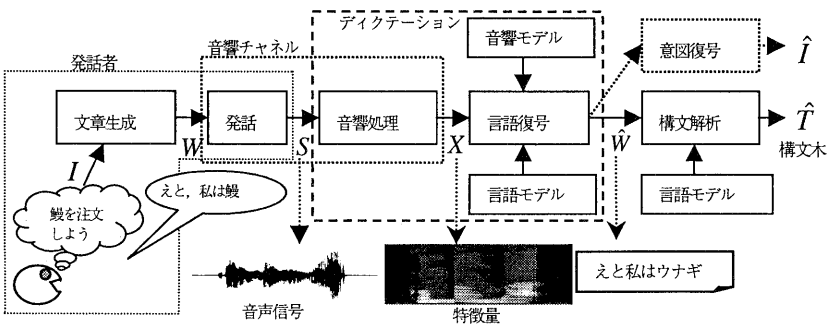

図 1 音声認識/理解の情報理論的解釈

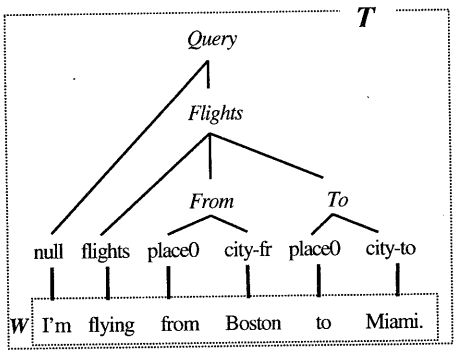

図2 構文木の例

って$\hat{T}$を推定することになろう[1].

$$\hat{T} = \arg\max_{T} \Pr(T \mid X) \quad \cdots(2.1)$$

$$= \arg\max_{T} \Pr(T \mid W, X) \Pr(W \mid X) \quad \cdots(2.2)$$

$$= \arg\max_{T} \Pr(T \mid W) \Pr(X \mid W) \Pr(W) \quad \cdots(2.3)$$

ただ，この式どおりに認識結果の出現確率まで考慮したシステムはまだ見当たらない．一般的には，Wは認識装置によって一意に決定されるものとして$\Pr(T|W)$を最大化するTを推定する．このような統計的パーザーがいくつか提案されている[7,8]．実際，IBMでは既に電話による音声対話システム開発用Toolkitの一部としてこのような統計的パーザーを提供している[9]．

なお，ポーズ位置や長さなど，Wには含まれない音響的な情報が構文木の推定に重要であると考えられる場合は式(2.2)に従い，$\Pr(T|W,X)$を最大化するTを求めることになる．

3. ディクテーションシステム

3-1. ディクテーションシステムの構成

ディクテーションに話を戻し，IBMのシステム[10,11]を例として具体的な構成を紹介する．

図1中の言語復号部の処理，言い換えると式(1.2)の推定は図3に示すような構成で実現されている．一見複雑に見えるのは主に処理の高速化のためにFast Matchと呼ぶ単語予備選択，言語確率を使ったpruningなどを高速探索手法(Stack Decoder)と併用しているためである．なお，探索は時間非同期の前向き探索で行われている．音響モデルは前後の音素環境別に推定されたHMMと，そのHMMの各状態に割り当てられた混合正規分布によって表現されている．このモデルはFast MatchおよびDetailed Matchで参照されるがFast Matchでは音素環境を区別しないことでより高速な処理を実現している．

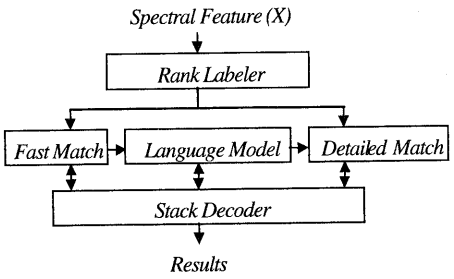

図3 ディクテーションシステムの構成

一方，Rank Labelerでは，これらHMMの各状態に対応付けらた混合正規分布に対し，入力音声の特徴量が出現する確率(実際は確率密度)を順位付けし，各状態に順位ラベルをつけている．そしてこの順位ラベルの出現頻度を正規化したものを各音響モデルの順位ラベルの出力確率としてFast MatchおよびDetailed Matchで参照する．このように$\Pr(X|W)$を混合正規分布の確率密度関数から直接算出するのではなく順位確率テーブルの参照で実現している点が本システムの一つの特徴となっている．

一方言語モデルとしては認識対象単語の表記のみをエントリーとする3-gramモデルを用いている．実際には式(1.2)に示したような言語モデルとDetailed Matchのスコアだけでなく，Fast Matchのスコアも併用する．これによってシステムとしてのロバストネスが高まることが分かっている．

3-2. ディクテーションシステムの改良点

ユーザーからの要望に答えるため，いくつかの機能をディクテーションシステムに付与してきた．同じくIBMのシステムを例にして，特に日本語に関係する機能上の改良点について述べる．

3-2-1. 読みの推定

英語では同じスペルの単語の発音が複数あった場合，実際にどの発音が認識されたのかが問題になることはほとんど無い．一方日本語では，発音が同じで表現が異なる同音異義語あるいは同音同義語が非常に多く，発音(「読み」)の推定値が返されると認識結果の修正や，新規単語登録時に役に立つ．

基本的な動作原理は前節に示したとおりで，$\Pr(x|w)$は，HMMによって推定される．日本語では同一表記に複数の読みがあることが多いので，wのエントリーとして表記だけでなく「読み」も含むように設計されたシステムもあるが，我々のシステムではwは表記のみをあらわすものとし，以下の近似を行っている．ここでpは単語wの「読み」をあらわす．

[1] TにはWの情報が含まれることに注意．

$$\Pr(x \mid w) = \sum_{p \in P_w} \Pr(x \mid p, w) \Pr(p \mid w) \quad \cdots (3.1)$$

$$\cong \max_{p \in P_w} \Pr(x \mid p, w) \quad \cdots (3.2)$$

この操作によって推定された p を認識結果 w と共に返す.

3-2-2. スペル入力

音声タイプライタとも呼ばれる機能で，日本語では音声仮名入力を実現する．未登録語などの入力用に実現したものである．仮名文字の 3-gram モデルを実装し，左側の音素環境も最大 5 音素まで考慮している．辞書を必要としないので任意音節列の入力が可能であるが，認識精度は必ずしも高くない.

3-2-3. トピック LM (特定分野への適応)

コンピュータ用語など特定分野のテキストを入力するために用意された LM で，単語出力確率を汎用の LM と次式で線形補間して用いる．なお，後で述べる句読点自動挿入機能もこのしくみを用いて実装されている.

$$\Pr(w_3 \mid w_1, w_2) = \lambda \Pr_G(w_3 \mid w_1, w_2) + (1-\lambda) \Pr_T(w_3 \mid w_1, w_2) \quad \cdots (4.1)$$

3-2-4. 不要語の処理[11]

原稿の読み上げの場合でも，不要語が挿入されることがある．このため，日本語文の読み上げにおいて特に出現頻度が高かった不要語として<エ>や<エー>等を発音辞書に登録した．これらの不要語は言語的には単語間に一定確率で出現するものと仮定し，さらに後続の単語の予測には影響を与えない"透過単語"として取り扱う(図 4). なお，不要語は認識結果としては表示されない.

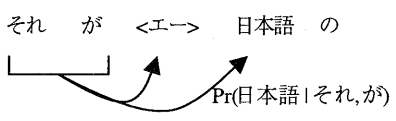

図4 透過単語の扱い

3-2-5. 句読点の自動挿入[11]

英語では句読点(Punctuation mark)を N-gram 言語モデルによって精度良く予測することは難しいとされ[12]，ディクテーションでは句読点を明示的に発話するスタイルが踏襲されてきた．一方，日本語の読点(、)は英語のカンマに比べると文意に与える影響が小さく，息継ぎの位置に対応させるだけでも妥当な自動挿入が出来る．また，句点(。)についても，基本的に息継ぎ位置に対応するうえ，日本語は SOV 型の言語であるため，文末に出現する品詞や活用形から句点を予測しやすいという特徴がある．具体的な実装方法としては，句点および読点に「まる」，「てん」といった発音だけではなく，無音をあらわす発音記号を割り当てて辞書に登録しておけばよい．このようにするだけで，認識時に無音部分が検出された際，一般的にはその部分には何も表示されないが，N-gram モデルの確率を参照して，句点あるいは読点が出現する可能性が言語的により高いと判断された場合にはその部分に句読点が自動的に挿入されることになる.

3-2-6. テキスト解析ツール (未知語処理)

個人のテキストデータベースを活用して，個人用途の言語モデルおよび未登録語のリストを自動生成するためのツールである．この時，英語ではアラビア数字や短縮形等の処理が問題になるだけだが，日本語では入力テキストを単語単位に分割する精度の高い手段が別途必要になる．一般には形態素解析プログラムを使うことが多いが，我々のシステムでは統計的なモデルで抽出した単語を認識単位としたこともあり，N-gram モデルを用いて入力単語列のゆう度が最大になるような単語分割位置を推定している．またこの際，未知語部分を他の部分と矛盾なく統計的に表現するため，文字種クラスの N-gram と文字種テンプレートを組み合わせたモデルを未知語モデルとして使用した[11].

4. 自由発話の音声認識

4-1. 次の課題としての自由発話

現状のディクテーションシステムでは，「読み上げ」に分類されるような，丁寧でまた文法的にもある程度正しい音声しか認識できない．先に述べたような不要語処理も入ってはいるが，あくまでも補助的な役割しか果たしていない．将来の音声理解/対話システムの実現に向け，また，それ以前に大語彙音声認識技術の応用分野を広げるためにも自然な発話(自由発話)に対する認識精度の向上が必要である.

ただ，読み上げ以外の発話をすべて自由発話と分類してしまうことは少し危険である．もう少し目標を段階的に場合分けしておく必要がある．たとえば，ARPA のプロジェクトで使用されている ATIS と Switchboard の 2 つの自由発話コーパスでも，human-computer dialogue である ATIS では不要語(Disfluency)の出現頻度が 1%程度であるのに対し，human-human dialogue である Switchboard では 6%程度の頻度で観測される[13]など，自由発話の程度には差があることが知られている．特に，発話が不明瞭で，認識にはコンテキストの情報も必要とされるような一般対話音声と，ある程度原稿が用意され，その上比較的丁寧に発声された講演や議会の答弁などのような音声は明確に区別して取り扱うべきである．前者は単なるディクテーション技術の拡張ではなく，2-2.節で述べたような高次情報を利用する枠組みの中で検討する必要があるだろう.

4-2. 放送大学講義コーパス

我々は自由発話の中で最もディクテーションに近い対象として；講演や議会の答弁等が自由発話研究の最初の足がかりとなると考えている．ただ，実際の講演を収集するには多大な費用と時間を要するし，収集環境による音響的なばらつきなど，安定した品質のデータを入手することは容易でない．

そこで，安定した品質でかつ大量のデータが継続的に入手できる対象として「放送大学」の講義音声に着目した．CS デジタル放送を収録して文単位に時間情報を付与するとともに，不要語を含めて正確に人手で書き起す作業を進めている[14]．既に，1998 年 11 月 22 日から 12 月 19 日までの全講義中，語学や数学，それに女性講師担当分を除く授業計 148 回分の講義コーパス(男性話者 97 名，約 100 万単語，発話部分約 83 時間)の整備を終えた．

4-3. 講義音声の認識実験

整備の済んだ放送大学講義コーパスを各統計モデルの学習データとして用い，性別のみ既知の不特定話者連続音声認識による講義音声の書き起しを試みた．

4-3-1. 音響モデル

音響モデルを「読み上げ」コーパスから学習した場合(BL-AM)と，上記放送大学講義コーパスから学習した(SP-AM)場合とを比較する．なお，この実験で使用した BL-AM 用学習音声コーパスのサイズは SP-AM の約 5 倍である．

4-3-2. 言語モデル

言語モデルは主に新聞記事(約 200M 単語)によって学習した(BL-LM)場合と，放送大学コーパスによって学習(SP-LM)し，(4.1)式に従って BL-LM と補間した場合について比較した．

言語モデル学習時の語彙はいずれの場合も主に新聞記事に高頻度に含まれた約 75K 単語である．なお，SP-LM の学習時には，出現頻度の高かった不要語 45 単語(主に間投詞)を，図 4 に示したような透過単語として学習[15]している[2]．

4-3-3. 認識タスク

認識対象語彙は言語モデル学習時と同じ約 75K 語とした．講義音声の認識用に特に追加した単語はない．ただし，不要語処理を行う場合のみ 45 単語分の不要語を透過単語として発音辞書に追加している．

4-3-4. 評価用データ

評価用データは放送大学講義コーパスから無作為に抽出した 5 つの講義(世界の教育，金融論，計測と制御，カオスの数理と技術，現代生物学)の冒頭各 60 文，計 300 文(6310 単語)とした．当然これらの講義は各モデルの学習用データには一切含めていない．このデータに限ると，間投詞や言いよどみなどの不要語が 8.6%(意味のある単語として認識可能な言い直しや，語尾の長音化[3]はこれに含めない)，それ以外の未知語が 1.2%で，計 9.8%を占めた．

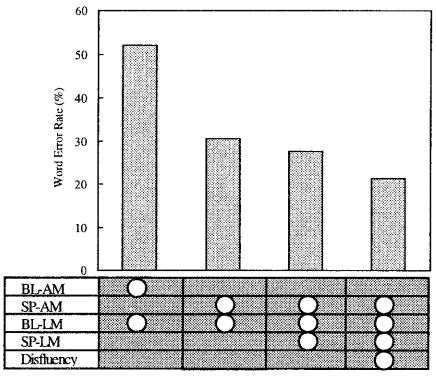

図5 講義音声の不特定話者音声認識実験結果 (75K 語)

4-3-5. 評価方法

言い直し以外の，間投詞，語尾の長音化(e.g.それで<デー>)，言いよどみなど，書き起し作業時に不要語のラベルが付与された部分をすべて正解テキストから除去し，認識システムの出力と比較する．つまり，不要語部分についての認識性能，言い換えると「不要語部分を正しく不要語として処理したか」という点も含めて評価を行う．なお，不要語処理をする場合，不要語は認識結果に明示的には出力されない．

4-4. 実験結果と考察

既存のディクテーションシステムによる結果(BL-AM&BL-LM)の場合との比較の形で，自由発話用 AM, LM および不要語処理を適用した場合の単語誤り率を図 5 に示す．なお，いずれの場合も Pentium-II 300MHz クラスの CPU でリアルタイム処理されている．

この結果を見る限り，最良の場合でも単語誤り率が約 20%となるなど，数値的にはまだ満足できるレベルにはないが，情報検索などの用途ならばこの程度の性能でも実用化の可能性はある．

音響的には自由発話音声データの整備がいかに重要であるかということを強く示唆する結果が得られている．なお，SP-AM および BS-LM を用いてある「読み上げ文」を認識したところ，単語誤り率は 5.9%であった．一方，BL-AM と BL-LM では 4.6%であり，総学習データ量が BL-AM の 1/5 と少なかったにもかかわらず，SP-AM は

[2] 一定確率を用いるのではなく，左側コンテキストに依存した透過単語の出現確率を学習データから推定する．

[3] 語尾の長音化は全体の 1.5%の単語で観測された．

読み上げ文に対してもおおむね良好に動作するモデルとなっていた.

一方，講義コーパスから構築した言語モデルに関しては，認識精度上まだ十分な効果をあげているとは言えない．ただ，未知語および不要語以外の部分について算出したテストセットパープレキシティで見ると，BL-LM のみの場合は 512.9，SP-LM を併用すると 226.2 であり，SP-LM が発話スタイルの学習にとって有効であったことはあきらかである．なお，SP-LM 併用時に不要語も予測した場合のパープレキシティは 356.0 となっていた.

最後に，不要語処理についてはおおむね良好に動作したことが分かる(図 5，Disfluency)．しかし，N-gram のような局所的な情報だけによってすべての不要語を処理できるのかというと疑問が残る．たとえば，「あのー」と言った場合に，それが間投詞なのか，指示代名詞なのかと言った問題については 2-2.節で述べたような統計的パーザーによる処理が必要となろう.

5. おわりに

大語彙音声認識技術はコンピュータ技術の絶え間ぬ改良にも支えられ，急速に発展し成熟してきた．'90 年代はじめにはまだ夢と思われていたディクテーションソフトが今ではビジネス上の成功を収め，音声認識は PC ユーザーにとってごくあたりまえの技術となった．さらに，研究室レベルとはいえ，自由発話の大語彙認識に関しても，講義やテレビニュース中での対談といった少し丁寧な部類の自由発話ならば，発話内容を十分把握できるレベルにまで認識精度が高まりつつある．基本的な認識性能はさらに高い目標に向けて今後も改良を続ける必要があるが，一方で，情報検索や議事録作成時の一次近似等の応用に限れば既に実用可能なレベルに達した部分もある.

また，発話の書き起しにとどまらず，より高次の情報も利用した枠組みの中で，音声認識から理解そして対話へと想像以上のスピードで実用化が進む可能性がある．基本的な道具は揃いつつある．今後の展開が非常に興味深い.

謝辞 放送大学の番組制作にあたられた方々に深謝します.

参考文献

[1] 新島 他, "音声認識による剖検記録システムの開発," 日本法医学会誌, Vol.52, p62, 1998.

[2] 安藤, "放送ニュース番組への字幕放送を目指した音声認識システム," Proc. of TAO Workshop, pp.121-129, 1999.

[3] http://www.nist.gov/speech/

[4] 松井 他, "地域や年齢的な広がりを考慮した大規模な日本語音声データベース," 日本音響学会 1999 年秋季研究発表会講演論文集, 3-Q-26, pp169-170, 1999.

[5] 古井, "国内外の音声言語資源," 言語資源共有機構設立シンポジウム資料, 1999.

[6] 西村 他, "音声認識・理解のための統計的言語処理," 電子情報通信学会誌, Vol.82, No.8, pp.828-831, 1999.

[7] D. M. Magerman, "Natural language parsing as statistical pattern recognition," Doctoral dissertation, Stanford university, 1994.

[8] S. Miller et al., "A fully statistical approach to natural language interfaces," Proc. of the 34th annual meeting of ACL, pp.55-61, 1996.

[9] K. Davies et al., "The IBM conversational telephony system for financial applications," Eurospeech-99, Vol.1, pp. 275-278, 1999.

[10] L.R.Bahl et al., "Performance of the IBM large vocabulary continuous speech recognition system on the ARPA wall street journal task," Proc. of ICASSP'95, pp.41-44, 1995.

[11] 西村 他, "単語を認識単位とした日本語大語彙連続音声認識," 情報処理学会論文誌, Vol.40, Nol.4, pp.1395-1403, 1999.

[12] D. B. Paul et al., "The design for the Wall Street Journal-based CSR corpus," Proc. of 5th DARPA speech and natural language workshop, 1992.

[13] E. E. Shriberg, "Preliminaries to a theory of speech disfluencies," Doctoral dissertation, University of California at Berkeley, 1994.

[14] 伊東 他, "口語体言語モデルのためのコーパス," 情報処理学会自然言語処理研究会, NL-134, pp.9-14, 1999.

[15] A. Stolche et al., "Statistical language modeling for speech disfluencies," Proc. of ICASSP'96, pp.405-408, 1996.