自 然 言 語 処 理　134－2
（1999．11．25）

口語体言語モデルのためのコーパス

伊東伸泰　西村雅史

(株)日本アイ・ビー・エム，東京基礎研究所

あらまし: 統計的手法による自然言語処理の進歩には、データとなるコーパスを欠かすことができない。日本語でもEDRをはじめとして、コーパスが整備されてきているが、これらの多くは「書き言葉」を対象としている。また「話し言葉」を対象として収集されたものも旅行予約など限定されたタスクの対話が中心である。筆者らは、講演・プレゼンテーションなど「まとまった内容を説明するさいの話し言葉」、言い換えればモノローグに重点を置き、かつ特定のドメインに限らないコーパス作成を目的として放送大学で行われている授業に着目した。そこでこれらの放送録音から書き起こし作業、単語単位への分割を行った。その結果、合計78科目(のべ148番組)の授業から単語数で、約110万語の音声データと対応付けられたコーパスを得ることができた。さらにこのコーパスを用いて、不要語も含めて予測する言語モデルを検討したところ、不要語をカウントしない場合は211.6であるのに対して、比較的簡単なモデルでパープレキシティは225.0となった。

キーワード　　コーパス，話し言葉，モノローグ，言語モデル，自由発話

A Corpus for Building a Language Model of Spoken Japanese

Nobuyasu ITOH and Masafumi NISHIMURA
Tokyo Research Laboratory, IBM Japan Ltd.

Abstract Corpus data is essential to the improvement of natural language processing by statistical methods. In Japan some text corpora have been created such as EDR. Most of them are written Japanese. The corpora of spoken Japanese reported are dialogue-based, the domains of which are restricted such as reservations on travel. We focused on a collection of spoken Japanese that is used in a presentation and a lecture (monologue) and selected classes broadcasted by the Air University. A total of 78 subjects (148 classes) are transcribed with disfluencies. A corpus of about 1.1 million words corresponding with speech data was obtained. We also conducted some experiments on models for predicting disfluencies (mainly *filled pause*). In a simple model, the perplexity is 225.0, 6% higher than the baseline (the case in which disfluencies are not counted).

Keywords Speech Corpus, Spoken Language, Monologue, Language Model, Spontaneous Speech

1. はじめに

近年、日本語においても言語を統計的に取り扱い処理するアプローチがさかんに試みられている。これは計算機環境の向上と日本語コーパスが整備されてきたことが大きく貢献している。実際、タグ付けされていないものなら、数年分以上の新聞記事テキストが利用可能である。しかしながら、これらのほとんどが「書き言葉」であり、そこでは統計的手法を用いた言語解析・処理の試みも、「書き言葉」に限られる。一方音声認識の分野でもディクテーションで用いる言語モデルの研究がさかんに行なわれているが、本来追究されるべき口語体は対象となっていない。

日本語で口語体データの収集が行なわれた例としては、ATR によるもの[ATR] [竹沢 95]と日本音響学会 (以下単に音響学会とよぶ)によるもの[小林 92]、JEIDA によって収集されたもの等がある。ATR では旅行や国際会議申し込みを中心とした対話が収集・テキスト化されていて、付けられている言語情報の豊富さや対訳の存在が評価されているが、対象とするタスクはかなり絞られている。また音響学会によるものも、観光他の案内の模擬対話で、やはり、特定のタスクを実現するための基礎データという色彩がつよい。コーパスの作成は非常にワークロードが大きいものであるため利用目的を明確にすることが効率的ではあるが、そこで得られた知見がどの程度汎用的であるかどうかについては検証する必要がある。

また対話 (Dialogue)の研究に至る基礎として、人があるまとまった内容を説明している話し言葉 (Monologue)の研究の重要性が指摘されている [古井 99]。筆者らも、講演や会議の書き起こしをアプリケーションとして、読み上げではない自発的 (Spontaneous) な発声の認識システム構築を目指しているが、今回そのための基礎データとして「放送大学」に着目し、その書き起こしコーパスを作成したので、その内容について報告する。

2. 放送大学

目的が明らかな場合を別にすれば口語コーパスを作成する上で望ましい条件に以下のようなことが考えられる。

・特定分野ではなく、様々なドメインのテキストが収集できる。
・一定数以上の発話者から収集できる。
・(音声データも同時に集めるとすれば)収録条件ができる限り一定である。

ことなどが考えられる。このような観点からみて放送大学の授業はかなりよく条件を満たしている。授業であるため発話のスタイルが限られるのではないかといった点や、授業で詳細な原稿を用意する講師がほとんどで、自由発話ではないのではないかといった批判が考えられるが、「人に内容のあることを丁寧に説明する」スタイルとしては一般性があると考えられるし、後者については、結果的にそうではないことを以下の節で示す。

3. 書き起こしの手続き

3-1 音声データ収録

特別講義等を除いた 153 科目について、1998 11 月 22 日より同 12 月 19 日までの 4 週間、CS デジタル放送から授業を DAT (48KHz)に収録した。各科目は 1 週間に 1 回、45 分の授業が行われるので、各科目について連続した 4 回分の授業 (のべ 612 コマ)を収録したことになる。

3-2 対象科目の選択

つぎに、なるべく各専攻・分野に偏らないように 78 科目(のべ 312 コマ)を選択した。ただ

し、本研究の目的は現代日本語の口語コーパスを作成することであるから、その目的に合致しないであろう科目、たとえば外国語や古典文学、言語学等は意図的に避けた。また数式等の列挙が多くを占める数学もはずしている。結果として選択された科目を分野、専攻別に表1に示す。

表 1　分野別科目数

分野・専攻	数
人文科学	12
社会科学	6
自然科学	6
生活と福祉	4
発達と教育	5
社会と経済	11
産業と技術	12
人間の探求	8
自然の理解	14

今回は収録した音声データの中で、表1に示した選択科目の前半2回分(のべ156コマ)を書き起こしの対象とした。さらに本コーパスは単にテキストの解析にとどまらず、対応付けられた音声データを作成し、音響モデル・認識システムの構築にも使用することを考えた。そのためには書き起こし部分については音データとしても使用可能であることが望ましい。そこで書き起こし対象について、以下の条件を設定し、当該部分は書き起こしの対象としないことにした。

・主として講義を行っている講師以外の発声
・講師が女性[1]
・屋外で収録されていたり、ビデオを再生しながらの講義で発声以外の音が連続的に混入している。

結果としてのべ合計148回分の授業が書き起こされた。

3-3 書き起こし規則

書き起こしやその後のテキスト処理は、詳細な規則がないと、揺れを生じ易い。Text Encoding Initiative [TEI]でも Spoken Language に関して様々な指摘が行われているし、日本では、JEIDAが対話コーパスを作成するにあたってそのガイドラインを提供している[JEIDA 97]。一方詳細な規則・タグを定義したとしても清書法が確立していない部分では主観による不一致が生じ[小林 95]、また詳細規則のためあまりに書き起こし時間がかかると、最小限のデータサイズすら確保できない。そこで今回の書き起こしでは、情報漏れのない文字化に重点を置き、比較的緩やかな規則を採用した。その主なものは、

・発声を句点位置または Long pause (1sec 以上)を単位として、分割し、その単位ごとに文字化する[2]。
・間投詞的な発話、言いよどみで単語にならないようなもの (以下不要語と呼ぶ)は'<', '>'で囲み、聞こえたとおりの音を片仮名(音節を構成しない場合はアルファベット)で表記する。語尾の母音が伸ばされ「を」が「オー」となっている場合は当該単語を'<','>'で囲み、その中に伸びた音とともに記述する (i.e. <を<オー>>)。
・咳、喉ならしなど片仮名列として表記することが困難な音 (非単語)は'<<', '>>'で囲み、中に音の種類を<<咳払い>>、<<唾を飲み込む音>>などのように記述する。
・言いよどみに近いかあいまいな発声ではあるが言い直しが行なわれず、かつ該当単語が明らかな場合は当該単語を記述した上で'{', '}'で囲む。
・Pause (0.3sec を閾値とする)とコンテキストを元に読点を挿入する。
・句点位置で分割を行った場合は、最後に句点

[1] 女性話者が少なく、すべてを書き起こしても音響モデルを構築するだけのデータ量を確保できない。

[2] 本報告の範囲を離れるが、該当する単位で音声データも分割し、テキストと対応づけている。

を付ける。

である。表記の統一は、書き起こし時点では行わず、次節で述べる後処理として行った。

また作業者による判断の異なりを少なくするため、作業を3段階に分け、詳細書き起こしにおいては2人/1組で作業すると同時に、チェック時には作業者を入れ替え、1つのデータを複数人が判断するシステムをとった。

4. 後処理

3節で得られるテキストデータは不要語等を除いて単なる文字列であり、このままでは分析に使用することができない。そこで以下の二段階の後処理を行った。

4-1 単語単位への分割

筆者らは、音声認識の言語モデル作成の単位として、解析的に決められた形態素ではなく、「人の感覚」に基づく単語単位を用いることを提案し、その作成法、言語モデルの評価・有効性について報告した[伊東 99]。そこでは活用語の多くが後続の付属語(連鎖)とともに1単語を構成している。本研究のコーパスでもこの単位を採用し、語彙(約 60K 単語)と言語モデル(*N*-gram モデル)を用いて文字列 $S=c_1c_2,\ldots,c_n$ を以下の式、すなわち $P(W)$を最大にする単語列 $W=w_1w_2,\ldots,w_m$ に分割した[伊東 97]。

$$\underset{w}{argmax}\, P(W|S) = \underset{w}{argmax}\, P(W)$$

ここで確率 $P(W)$は単語 *N*-gram モデルによって求める。ただし、'<','>'で囲まれた不要語や非単語については透過単語として扱い、確率算出には含めていない。得られた分割結果はすべて人手により、分割エラーを修正した。

4-2 付加情報タグ

以上の作業により、単語単位に分割された書き起こしが得られるわけであるが、言語モデル作成上はまだ足りない情報がある。第一は不要語中、いわゆる間投詞や英語で *filled pause* と呼ばれているもの[Stolcke 96]と言い誤りで単語と認められないものの区別がなされていないこと。第二は「あの」「その」などの指示代名詞と間投詞の「アノ」「ソノ」が区別されていないことである[3]。そこでこれらについては別途タグを定義して付加することとした。

ただし、これらの区別は困難な場合があり、あいまいな場合は、間投詞に分類してタグ付けしている。

5. コーパスの統計量

5-1 不要語

得られたコーパスの諸元を表 2 に示す。単語数は約 1,100K (110万)であるが、形態素数を調べるため、われわれの形態素解析プログラムにより解析したところ、のべ 1,263K の形態素が得られた[4]。この中で不要語の数を現在までの研究と比較してみる。

表 2 本コーパスの諸元

話者数	97
総単語数	1,098,888
単語数(異なり)[5]	23,929
文数[6]	27,135
非単語	8,570
不要語	99,441
不要語(異なり)	2,263

過去に対話コーパス等を用いて、自然発話(Spontaneous Speech)を分析した研究には中野[中野 95]や、中川 [中川 95]、村上 [村上 91]

[3] 語尾が長音化して<あの<オー>>となっている場合は、指示代名詞である確率はきわめて低く区別する識別子にはなりえる。

[4] 人手によるエラー修正は行っていない。

[5] 非単語、不要語を含まない。

[6] 話し言葉の場合、句点を挿入する個所にはあいまい性があるため、参考数値である。

の研究がある。いずれも間投詞的表現の数について調べているが、出現確率について陽には述べていない。ただし提示された頻度等から推定すると中野の場合で 2.5-3%、中川らが用いた音響学会のデータで 7-10%程度と考えられ、これらの値はわれわれのコーパス(9.1%)と同程度か、より低い。したがって、一部の講師で原稿を読み上げている可能性は残るものの、全体としては十分自発的な発声であると思われる。図 1 は不要語の割合からみた話者の分布を示す。

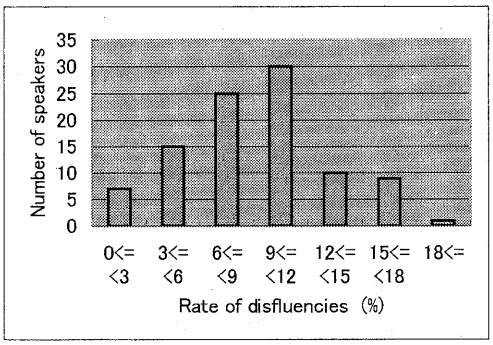

図 1 不要語の出現頻度からみた話者数の分布

5-2 単語カバレージ・パープレキシティ

今回収集したコーパスは既存の語彙や言語モデルでどの程度対応が可能であろうか？ このことを検証するため、既存の語彙・言語モデルを用いて、本コーパスの検証を行った。用いた語彙・言語モデルのベースとしたのは[伊東 99]で報告を行ったものである。ただし単語数は 44K から 60K、言語モデルの学習用コーパスは 98 年度産経新聞を中心に約 30%増やし、合計 148M 単語となっている。表 3 に毎日、日経の新聞、そして放送大学コーパスの内、無作為に選んだ約 1/10 のデータをテストデータとして既存の語彙･言語モデルを評価した結果を示す。放送大学コーパスのテストにおいては不要語・非単語を削除しており、またパープレキシティの計算においては、未知語はカウントしていない。科目単位でみたとき放送大学のカバレージの標準偏差は±1.3%であり、書き言葉の一般用語彙であるにもかかわらず、不要語を除けばカバレージは比較的よいことがわかる。

表 3 既存の語彙・言語モデルによる評価

テストデータ	カバレージ(%)	パープレキシティ
日経新聞	99.5	91.1
毎日新聞	98.4	146.5
放送大学	*98.4*	*211.6*

5-3 不要語を考慮した言語モデル

さて、実用的な音声認識システムにおいては、非単語や不要語に対する考慮を避けてとおれない。この内、リップノイズなど非単語については音響的な対応策がとられるべきであろう。しかし不要語については、言語的にも何らかのモデル化を行うことが認識率向上に役立つと期待される。ここでは比較的簡単な以下の 2 つのモデルをテストした。

[モデル 1] 不要語予測時には放送大学コーパスから得られた当該 1-gram 確率を用い、後続の単語予測時には不要語を未知語として取り扱う。すなわち以下の式で表される。ただし、w_{fil} は不要語、w_{UNK} は未知語、P_{au}、P_{base} はそれぞれ放送大学コーパスから学習した確率とベースの言語モデルから得られる確率を表す。

$$P(w_n|w_{n-1}w_{n-2}) = \begin{cases} \lambda P_{au}(w_{fil}|\{w_{fil}\}) & (\text{if } w_n \in \{w_{fil}\}) \\ (1-\lambda)P_{base}(w_n|w'_{n-1}w'_{n-2}) & (\text{otherwise}) \end{cases}$$

$$\lambda = P_{au}(\{w_{fil}\})$$

$$w'_{n-i} = \begin{cases} w_{n-i} & (\text{if } w_{n-i} \notin \{w_{fil}\})\ i=1,2 \\ w_{UNK} & (\text{otherwise}) \end{cases}$$

[モデル 2] 不要語予測時はモデル 1 と同じで、後続の単語予測時には不要語の存在を無視し、透過単語とする。つまり上式の最初は同じで、

2 番目が以下のようになる。これは甲斐らのモデル [甲斐 99] に近い。

$$P(w_n|w_{n-1}w_{n-2})$$
$$= (1-\lambda)P_{base}(w_n|w_{n-i}w_{n-j})\ (w_n \notin \{w_{fil}\})$$
$$w_{n-i}w_{n-j} \notin \{w_{fil}\}$$
$$w_{n-k} \in \{w_{fil}\}, {}^{\forall}k: 0 < k < i \text{ and } k \neq j$$

実験では放送大学コーパスを科目単位で 10 個に分け、9 個を不要語の学習に、1 個をテストとして用いた。 結果はパープレキシティが 251.1 と 225.0 で、不要語をカウントしない場合 (211.6)に比べると(後者の場合で)約 6%大きいものの、モデルの単純さの割にはよい結果が得られた。

6. まとめ

口語体コーパス作成の基礎資料として、放送大学の書き起こしテキストを作成し、それに基づくいくつかの統計量について報告した。間投詞をはじめとする不要語については、比較的簡単なモデルで 225.0 を得たが、全体としてのパープレキシティは新聞に比べまだ大きく、口語体のスタイルを学習する必要性を示唆している。今後はこのスタイルの学習に取り組みたい。

また本コーパスの明らかな問題点としては女性話者が含まれないということであるが、これについても別途コーパスの構築を考えたい。

謝辞

産経新聞社、日本経済新聞社、毎日新聞社 (CD 毎日新聞 91-95)、および放送大学で番組制作にあたられた関係者各位に深謝したい。

参考文献

ATR. http://www.itl.atr.co.jp/Japanese/overview/index.html

古井 (1999). 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.

伊東, 西村 (1997). *N*-gram を用いた日本語テキストの単語単位への分割, 情処 自然言語処理研究会, NL122-9, pp. 57-62.

伊東他 (1999). 単語単位による日本語言語モデルの検討, 自然言語処理, Vol. 6, No. 2, pp. 9-27.

JEIDA (1997). 自然言語処理システムの動向に関する調査報告書, pp. 162-178.

甲斐, 廣瀬, 中川 (1999). 単語 *N*-gram 言語モデルを用いた音声認識システムにおける未知語・冗長語の処理, 情報処理学会論文誌, Vol. 40, No. 4, pp. 1383-1394.

小林他 (1992). 日本音響学会研究用連続音声データベース, 日本音響学会誌, Vol. 48, No. 12, pp. 888-893.

小林, 北澤 (1995). 習熟による対話音声情報の書き起こし精度の定量的評価, 情処 音声言語情報処理研究会, SLP 7-10, pp. 61-66.

村上, 嵯峨山 (1991). 自由発話音声における音響的および言語的な問題点の検討, 電子情報通信学会技術報告 SP91-100, pp.71-78.

中川, 小林 (1995). 自然な音声対話における間投詞・ポーズ・言い直しの出現パターンと音響的性質, 日本音響学会誌, Vol. 51, No. 3, pp. 202-210.

中野他 (1995). 対話文の文法構築に向けた分析, 情処 自然言語処理研究会, NL107-5, pp. 35-42.

Stolcke, A., Shriberg, E. (1996). Statistical Language Modeling for Speech Disfluencies, *Proc. ICASSP 96*, pp. 405-408.

竹沢, 末松 (1995). 音声・テキストコーパスとその標準化動向, 人工知能学会誌, Vol. 10, No. 2, pp. 168-180.

TEI. http://www.uic.edu/orgs/tei.