WWW を用いた書き言葉特有語彙から話し言葉語彙への用言の言い換え

鍛治　伸裕[†]　岡本　雅史[†]　黒橋　禎夫[†,††]

書き言葉で使われる語彙と，話し言葉で使われる語彙には大きな違いがある．そのため，書き言葉テキストから合成された音声は不自然なものとなってしまう．書き言葉テキストからでも自然な音声の合成を可能にするために，本論文では，書き言葉特有語彙から話し言葉語彙への言い換えを学習する手法を提案する．ある表現が書き言葉特有語彙であるか，話し言葉語彙であるかは，その表現の書き言葉コーパスでの出現確率と話し言葉コーパスでの出現確率をもとにして判断する．書き言葉コーパスと話し言葉コーパスは WWW から自動収集したものを用いる．実験の結果，書き言葉コーパスと話し言葉コーパスの収集精度は 94％，言い換え学習の精度は 79％であり，提案手法の有効性を示すことができた．

キーワード:　言い換え，書き言葉，話し言葉，暗示的意味，*WWW*

Paraphrasing Predicates from Written Language Specific Vocabulary into Spoken Language Vocabulary Using the World Wide Web

Nobuhiro Kaji[†] , Masashi Okamoto[†] and Sadao Kurohashi[†,††]

There are a lot of differences between expressions used in written language and spoken language. This paper represents a method of paraphrasing written language specific vocabulary into spoken language vocabulary. They can be distinguished based on the occurrence probability in written and spoken language corpora which are automatically collected from WWW. Experimental results indicated the effectiveness of our method. The precision of the collected corpora was 94％, and the accuracy of learning paraphrases was 79％.

KeyWords:　*paraphrase, written language, spoken language, connotation, WWW*

1　はじめに

音声情報は，我々にとって身近な情報形態であるうえに，効率的に情報を伝達できるという特徴を持っている (Hayashi, Ueda, Kurihara, Yasumura, Douke, and Ariyasu 1999; Nadamoto, Kondo, and Tanaka 2001)．しかしながら，現在，我々が利用できる電子情報の大半はテキスト情報であり，音声情報は比較的少ない．こうした背景から，音声合成技術を利用したアプリケーションが関心を集めている (Fukuhara, Nishida, and Uemura 2001)．音声合成を使えば，

† 東京大学大学院情報理工学系研究科, Graduate School of Information Science and Technology, the University of Tokyo
†† 科学技術振興機構 さきがけ, PRESTO, JST

既存のテキスト情報を音声情報に変換してユーザに提供することができる．

しかし，こうしたアプリケーションの開発には二つの問題点がある．一つは，合成された音声のアクセントやイントネーションが不自然であるという，音声合成の質の問題である．もう一つは，書き言葉と話し言葉で使われる表現に違いがあるため，音声合成の入力テキストが書き言葉だった場合，通常の話し言葉では殆んど使われなような表現が音声化されてしまうことがある，という問題である．これら二つの問題のうち，前者はよく知られた問題であるが，後者は今までほとんど指摘されなかった．そこで我々は，自然言語処理の言い換え技術を用いて，後者の問題を解決することを考えた．言い換え技術は，文の平易化 (Inui and Yamamoto 2001; Inui, Fujita, Takahashi, Iida, and Iwakura 2003) や質問応答 (Lin and Pantel 2001; Hermjakob, Echihabi, and Marcu 2002; Duclaye and Yvon 2003) などへの応用例が多いが，このような試みは初めてである．

以下では，まず書き言葉と話し言葉の相違について考察する．書き言葉と話し言葉のもっとも基本的な相違は，話し言葉の同時性である (畠 1987)．話し言葉を使ったコミュニケーションでは，話し手が情報の送信を行うと同時に，聞き手が受信することになる．こうした同時性は，聞き手に負担を強要することになる．例えば，話し言葉では，話し手が会話のペースを設定することになり，聞き手がそれを制御することはできない．もし書き言葉であれば，読み手はテキストをゆっくり読もうと早く読もうと自由であるし，途中で読むのを中断することさえできる．

そのため，話し言葉は，書き言葉に比べて聞き手の負担が低い表現を使う傾向にある．例えば「難解な語彙が用いられることが少ない」「一つ一つの文が比較的短い」などがあげられる．畠は，このような話し言葉の特徴を**冗長性**[1]と呼んでいる (畠 1987)．もちろん，これは大まかな傾向であり，例えば講演のように冗長性が低い話し言葉もある．しかし，畠も指摘しているように，冗長性とは，いわゆる話し言葉がもつ基本的な特徴であると言える．そのため，本論文では冗長性の高い話し言葉だけを想定して議論を行う．冗長性の高い話し言葉を定義することは難しいが，ここでは畠の分類にしたがって次のように考える．畠は，冗長性に着目して話し言葉を以下の四つに分類している (畠 1987)．

第 I 類　発話の形成から発話までほとんど時間的経過のないもので非常に冗長性が高い．(例) おしゃべり

第 II 類　伝達内容が多少準備されているが，言語化そのものは即興で行われるもの．第 I 類ほどではないが冗長性は高い．(例) 相談，打ち合わせ，連絡，座談会

第 III 類　かなり計画的で時間をかけた発話．言語化自体もある程度準備されていて，冗長性は低い．(例) 講演，講義

第 IV 類　言語化の即興がほとんどない発話．冗長性は非常に低い．(例) ニュース，青年の主張

本論文では，話し言葉として第 I 類と第 II 類を想定して議論を進める．

1 いわゆる「冗長性」という語とは違う意味で使われている．

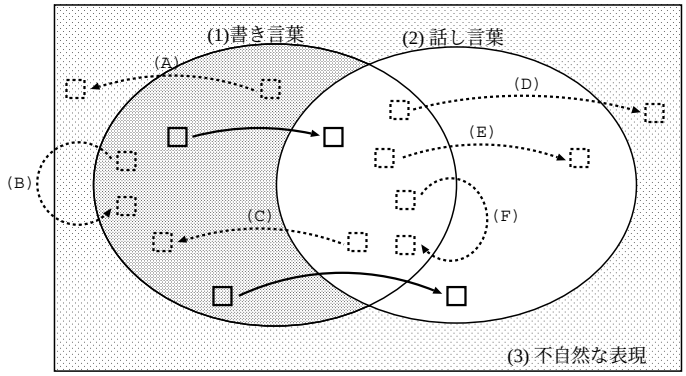

図 1　書き言葉特有語彙から話し言葉語彙への言い換え

書き言葉と話し言葉の差異の中でも，本論文は語彙の問題を扱う．以下では，書き言葉では使われるが話し言葉では殆んど使われない語彙を**書き言葉特有語彙**と呼び，話し言葉で通常使われる語彙を**話し言葉語彙**と呼ぶ．すなわち本論文では，書き言葉特有語彙を話し言葉語彙への言い換える，という問題を取り上げる．これを図で説明すると次のようになる (図 1)．まず，(1) 書き言葉で使う表現，(2) 話し言葉で使う表現，(3) どちらでも使われない不自然な表現，の三種類の表現を考える．不自然な表現を考慮しているのは，書き言葉特有語彙を自動処理で言い換えた先が，不自然な表現になる場合がたまにあるからである．図中の 2 つの円は，書き言葉で使う表現と話し言葉で使う表現を表している．そして，2 つの円の外の部分は，どちらでも使われない不自然な表現を表している．二つの円の重複部分は，書き言葉と話し言葉の両方で使う表現である．書き言葉特有語彙とは，書き言葉で使う表現から，書き言葉と話し言葉の両方で使う表現を除いたもので，左円の中の色がついた部分にあたる．話し言葉語彙とは，話し言葉で使う表現のことなので，図中の白い部分にあたる．

書き言葉特有語彙から話し言葉語彙への言い換えは，図 1 の矢印で表されているような言い換えであると言える．それ以外の言い換えは，破線矢印で表している．破線矢印の言い換えは，話し言葉で使われない表現 (書き言葉特有語彙または不自然な表現) が言い換え先になっているもの ((A)(B)(C)) と，話し言葉語彙が言い換え対象になっているもの ((C)(D)(E)(F)) がある．図 1 からも分かるように，入力となる書き言葉テキストには，言い換える必要のない表現 (二つの円の重複部分) が存在している．言い換えは，いわば「単言語内翻訳」と考えることができるため，機械翻訳との類似性がしばしば指摘されている (佐藤 1999)．しかし，このように言い換えは，変換対象とする必要のない表現が入力テキストに含まれているという点で，機械翻訳とは大きく異なっている．

本論文は，書き言葉特有語彙から話し言葉語彙への言い換えを学習する方法を提案する．あ

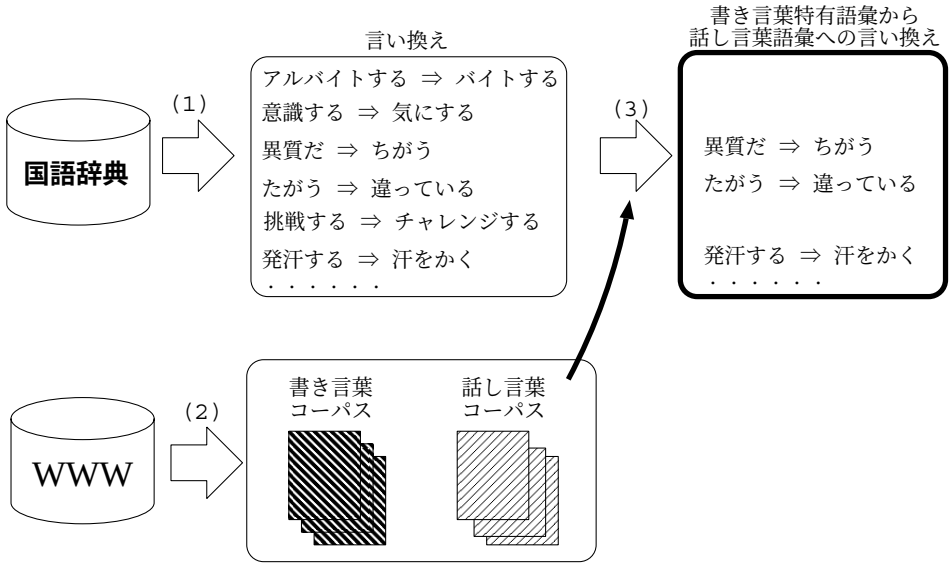

図 2　提案手法の流れ

る表現が書き言葉特有語彙であるか，話し言葉語彙であるかは，その表現の書き言葉コーパスでの出現確率と話し言葉コーパスでの出現確率をもとに判断する．コーパスは既存のものではなく，WWW から自動収集した大規模なものを利用する．提案手法の流れは以下のようになっている (図 2).

(1)　Kaji らの手法を利用して，国語辞典から用言の言い換えペアを学習する (Kaji, Kawahara, Kurohashi, and Sato 2002).

(2)　書き言葉コーパスと話し言葉コーパスを WWW から自動収集する．

(3)　それら二つのコーパスを用いて，(1) で学習した言い換えの中から，書き言葉特有語彙から話し言葉語彙への言い換えを選び出す.

入力テキストを，話し言葉に適したテキストに言い換えるためには，当然，用言以外の表現も言い換え対象とする必要がある．しかし，あらゆる表現を言い換え対象とすることは，現在の言い換え技術では困難なので，言い換え対象を用言に限定して議論を行う．

2　関連研究

2 つの表現が言い換えの関係にあるとは，それらがほぼ同一の意味を表しているということであるが，こうした場合に 2 つの表現が全く同じ意味を表していることは少なく，たいていは微妙な違いが存在している．この違いは 2 種類に分けて考えることができる．1 つ目は，指示的意味 (denotation) の違いで，2 つ目は，スタイルなどの暗示的意味 (connotation) の違いであ

る．本研究は，ある表現が話し言葉として使われるかどうか，という暗示的意味の違いを扱った研究と位置づけることができる．

言い換えの自動学習に関する研究は多いが (Lin and Pantel 2001; Barzilay and Lee 2003)，学習された言い換えの間にどのような違いがあるかについて，議論を行っている研究は少ない．本研究のように，言い換え間の暗示的意味の違いを扱った研究には，以下のようなものがある．Edmonds らは，暗示的意味の違いを計算機で表現するモデルを提案している (Edmonds and Hirst 2002)．そして Inkpen らは，そのモデルのパラメータを同義語辞書から自動学習する方法を提案している (Inkpen and Hirst 2001)．だが，同義語辞書のような既存の言語資源に暗示的意味の違いが十分に記述されているとは考えにくい．これに対して本論文で提案する手法は，コーパスでの出現確率にもとづくもので，既存の言語資源に依存しない．

一方，Inui らは，暗示的意味の違いの中でも可読性の違いに焦点をあて，トレーニングコーパスを使って可読性の違いを判定する手法を提案している (Inui and Yamamoto 2001)．トレーニングコーパスは，まず大量の言い換えのペアを用意して，そして，どちらが読み易いかを専門家にタグ付けしてもらって作成されている．Inui らの試みと本研究は，コーパスを使って暗示的意味の違いを判定するという点で類似しているが，Inui らは統語的な言い換えを扱っているのに対して，本研究は語彙的な言い換えを扱っているという違いがある．

村田らは，我々と同様に，書き言葉から話し言葉への言い換えを扱っている (村田，井佐 2001; Murata and Isahara 2002)．村田らの手法は，話し言葉コーパスにより多く出現する表現に言い換えるというものである．これに対して本論文は，書き言葉コーパスと話言葉コーパスの両方を利用した機械学習に基づく手法を提案する．また，村田らの手法は，人手で用意された書き言葉コーパスと話し言葉コーパスを前提としているが，本論文では，書き言葉コーパスと話し言葉コーパスを自動収集する．

文体が異なる 2 つのコーパス (書き言葉と話し言葉，イギリス英語とアメリカ英語など) を統計的に比較して，どちらか一方の文体に特有の語彙を発見する試みは多い (Kilgarriff 2001)．しかし，こうした研究では言い換えとの関連は十分に議論されていない．

特定の文体のコーパスを自動収集するという試みに関しては，Tambouratzis らが Demotiki コーパスと Katharevoua コーパス[2]を収集する方法を提案している (Tambouratzis, Markantonatou, Hairetakis, Vassiliou, Tambouratzis, and Carayannis 2000)．また，Bulyko らは，我々と同様に WWW から話し言葉コーパスを自動収集する手法を提案している (Bulyko, Ostendorf, and Stolcke 2003)．自動収集の手がかりとして，Bulyko らは既存の話し言葉コーパスの N-gram の情報を使っている．これに対して我々の手法は，後述するように，待遇表現に着目しているという点が異なる．

2 Demotiki と Katharevoua は，ギリシャ語の変異 (variation) の一つである

3　用言の言い換え学習

用言の定義文には，その用言の言い換えが含まれているので，言い換えの学習に利用することができる．そこで，先行研究の手法を用いて，国語辞典の定義文から用言の言い換えを学習する (Kaji et al. 2002)．以下に，用言とその定義文の具体例を示す．用言は太字で示している．

1. (a) **激怒する**
[激しく 怒る] こと
(b) **相乗りする**
乗物などに [いっしょに のる] こと
(c) **発汗する**
[汗を かく] こと

一般に，用言の定義文の主辞は用言であり，主辞には副詞や名詞がかかる場合がある．例えば (1a)，(1b)，(1c) の定義文の主辞は,「怒る」,「のる」,「かく」である．主辞にかかる副詞は一重線で，主辞にかかる名詞は二重線で表している．そして，定義文に含まれる言い換えは括弧でくくっている．

主辞とそれにかかる副詞は，必ず言い換えに含まれると考えることができる．主辞に名詞がかかっている場合，それが言い換えに含まれるかどうかは見出し語によって異なるが，(Kaji et al. 2002) の手法で判断できる．ほとんどの場合，言い換えに含まれる名詞は 0 個か 1 個である．したがって，定義文から取り出される言い換えは,「副詞 $*$ 名詞 ? 用言」という形をしていると仮定できる[3]．以下では，見出し語である用言を source と呼び，定義文から取り出された言い換えを target と呼ぶ．そして，この二つのペアを言い換えペアと呼ぶ．

この手法によって，例解小学国語辞典 (田近 1997) から 5,836 の言い換えのペアを学習した．例えば，上記の定義文からは次のような言い換えペアが学習された．

2. a　激怒する → 激しく怒る
b　相乗りする → いっしょにのる
c　発汗する → 汗をかく

こうして学習された言い換えペアの中から，次節以降に述べる方法を用いて，書き言葉特有語彙から話し言葉語彙への言い換えを選び出す．

3 $*$ は 0 以上，? は 1 以上を表す．

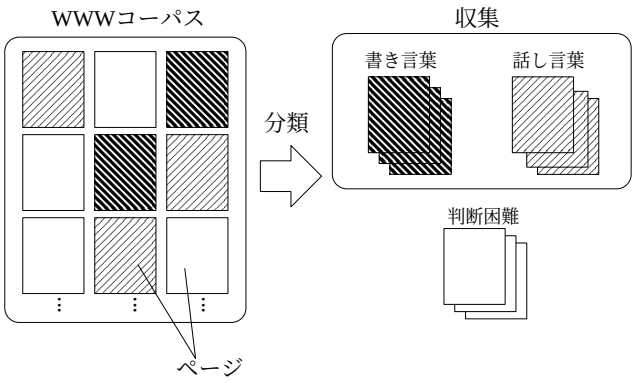

図 3　コーパスの自動収集

4　WWW からの書き言葉コーパスと話し言葉コーパスの自動収集

提案手法は，書き言葉特有語彙から話し言葉語彙への言い換え (図 1) を，source と target の書き言葉コーパスでの出現確率と，話し言葉コーパスでの出現確率に基づいて選び出す．そのためには，大規模な書き言葉コーパスと話し言葉コーパスが必要となるが，問題は，どのようにして大規模な話し言葉コーパスを準備するのかということである．日本語の話し言葉コーパスは，(Maekawa, Koiso, Furui, and Isahara 2000; Takezawa, Sumita, Sugaya, Yamamoto, and Yamamoto 2002) など最近少しずつ整備されてきているが，いずれも規模が小さい．

そこで次のような解決方法を考案した．まず，話し言葉語彙は，チャットや掲示板，日記，メールといったくだけたテキストに使われている語彙で近似できるという仮説をたてた．そして，WWW 上のそうしたテキストを話し言葉コーパスとして自動収集する，新聞記事などの典型的な書き言葉テキストを書き言葉コーパスとして自動収集する，ということを考えた．自動収集が可能になれば，大規模なコーパスを用意することが可能となる．

ここで問題は，上記のような仮説の妥当性であるが，次の二つの理由からこれは妥当であると考えている．まず第一に，冒頭で述べた通り話し言葉の基本的な特徴は冗長性であると考えることができるので，書かれたテキストであっても冗長性が高ければ代用できると予想できる．畠は，冗長性が高いテキストの例として，日記や家族への手紙をあげているが (畠 1987)，チャットやメールなどにも同様のことが言えると考えられる．第二に，WWW からチャットのようなくだけたテキストを自動収集して，そこから言語モデルを学習することによって，音声認識の精度が向上するという報告があ (Bulyko et al. 2003)．これは，チャットのようなテキストと，実際の発話が非常に類似していることを意味する．このことも，我々の仮説の裏付けと

捉えることができる．

図 3 に収集手法の概要を示す．まず，WWW から Web ページを集めてきて，そこから html タグなどの不要な部分を取り除く．こうして得られたコーパスを WWW コーパスと呼ぶ．そして，各ページを (1) 書き言葉 (2) 話し言葉 (3) 判断困難の三つに分けて，(1) または (2) に分類されたページだけを，書き言葉コーパス，話し言葉コーパスとして利用する．各ページを，書き言葉と話し言葉の二種類ではなく，三種類に分類するのは，WWW コーパスには人間でも書き言葉か話し言葉かの判断の難しいページがあり，そうしたページを無理に分類して利用しようとすれば，質の悪いコーパスが集まる原因になると考えられるからである．

4.1　待遇表現にもとづく分類

各ページを，書き言葉，話し言葉，判断困難に分類するために，待遇表現の一種である親愛表現と丁寧表現に着目した．

「聞き手に対する尊敬や親愛や軽侮などの態度を表す言語表現」を待遇表現という．待遇表現は，書き言葉よりも話し言葉で多く使われるという傾向がある．話し言葉は，特定の聞き手を想定して使われることが多いので，その聞き手に対する待遇表現が使われることが多い．これに対して，新聞記事などの書きことばは，不特定の読み手を想定して使われるのが普通である．そのため，基本的には待遇表現を使わないという傾向がある．

日本語は，用言の活用形や用言に付属する機能語によって，待遇表現の一種である親愛表現と丁寧表現を表すことができる．親愛表現とは，聞き手に対する親愛の感情を表す表現で，機能語「ね」や「よ」などを使うことによって表すことができる．例えば (3) は，機能語「ね」をつけることによって聞き手への親愛を表している．

3.　今度お稽古できる時には，もっとやりますから ね

丁寧表現とは，聞き手に対する丁寧な態度を表す表現のことで,「〜です」「〜ます」といった表現によって表すことができる．

4.　a　すぐ近くに見える虹がとても きれいです
　　b　またあんな感動を味わえたら，と思い ます

(4a) は形容詞「きれいだ」を「ですます形活用」で使うことによって丁寧さを表している．(4b) は機能語「ます」を使うことによって丁寧さを表している．

親愛表現と丁寧表現は，話し言葉で非常に多く出現するうえに，形態素解析と簡単なルールによって認識できる．そこで，各ページが書き言葉，話し言葉，判断困難のいずれであるかを判定するために，次の二つの数値を用いた．

- 親愛表現を含む文数 / ページに含まれる全文数

表 1　分類規則

親愛表現率 = 0 かつ 丁寧表現率 = 0	→ 書き言葉
親愛表現率 > 0.2 または 親愛表現率 > 0.1 かつ 丁寧表現率 > 0.2	→ 話し言葉
上記以外	→ 判断困難

表 2　コーパスの規模

	ページ数	語数
WWW コーパス	336,341	391,582,073
書き言葉コーパス	38,472	38,941,503
話し言葉コーパス	33,186	51,801,168

- 丁寧表現を含む文数 / ページに含まれる全文数

前者を親愛表現率，後者を丁寧表現率と呼ぶ．

4.2　自動収集

各ページの親愛表現率と丁寧表現率を求める．そのためには，親愛表現と丁寧表現を含む文を判定する必要がある．判定は以下のように行う．まず，WWW コーパスを Juman[4]を用いて形態素解析する．そして，次の機能語のうちいずれか一つでも含む文は親愛表現を含むとする．

ね，よ，わ，さ，ぜ，な

次に，それ以外の文で，次の機能語のうちいずれか一つでも含む文，もしくは「ですます活用」の用言を含む文は丁寧表現を含むとする．

です，ます，ください，ございます

このようにして，親愛表現と丁寧表現を含む文を判定し，各ページの親愛表現率と丁寧表現率を求める．

収集された WWW コーパスの一部を人手で調査し，表 1 のような分類規則を作成した．この規則にしたがって，ページを書き言葉，話し言葉，判定困難の三つにわける．そして，書き言葉または話し言葉に分類されたページだけを収集する．

4.3　評価

収集に使用した WWW コーパスと，収集された書き言葉コーパスと話し言葉コーパスの規模を表 2 にしめす．WWW コーパスのサイズは 336,341 ページ，391,582,073 語である．そして，収集された書き言葉コーパスのサイズは 38,472 ページ，38,941,503 語で，話し言葉コーパ

4 http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html

表 3　適合率

	被験者 1	被験者 2
書き言葉コーパス	95 %(119/125)	89 %(110/125)
話し言葉コーパス	94 %(109/115)	97 %(111/115)
合計	**95 %**(228/240)	**92 %**(221/240)

スのサイズは 33,186 ページ，51,801,168 語であった．

既存のコーパスとの比較　収集された話し言葉コーパスと，既存の話し言葉コーパスの規模を比較した．我々の知る限り最も大規模なものは「日本語話し言葉コーパス」であり (Maekawa et al. 2000)，その規模はおよそ 7,000,000 語である．一方，収集された話し言葉コーパスは 51,801,168 語を含んでおり，こうした既存のコーパスと比較しても十分に大きなものを構築できたといえる．

適合率による評価　提案手法にとって重要なのは適合率であり，再現率は問題ではない．なぜなら，我々の手法は大規模な WWW コーパスからの自動収集というアプローチなので，たとえ再現率が低くても適合率が高ければ，高い質で大規模なコーパス収集が可能になるからである．逆に適合率が低ければ，たとえ再現率が高くとも質の良いコーパス収集は不可能である．

書き言葉または話し言葉として収集されたページをランダムに 240 ページ取り出し，そのうち何ページが正しく集められいるかという適合率による評価を行った．240 ページの中には，手法が書き言葉と判断したページが 125 ページ，話し言葉と判断したページが 115 ページ含まれていた．

評価は，2 人の被験者 (以下では被験者 1，被験者 2 と呼ぶ) が個別に次のような手順で行った．各被験者は，各ページを書き言葉，話し言葉，判断困難の 3 つに分類する．このとき，そのページが手法によって書き言葉として収集されたか，話し言葉として収集されたのかは，被験者に知らせていない．そして，提案手法が書き言葉/話し言葉として選んだページを被験者も書き言葉/話し言葉として選んだ場合にのみ，そのページは正しく収集されたと考えた．

提案手法の適合率を表 3 に示す．被験者 1 は 240 ページ中 228 ページが正しく収集されていたと判断し，被験者 2 は 240 ページ中 221 ページが正しく収集されていたと判断した．各被験者による適合率は 95(=228/240)%と 92(=221/240)% で，その平均は 94 %あった．この数字を見るかぎり十分な質のコーパスが得られたと考えられる．

具体例　以下に，自動収集された書き言葉コーパスの一部を示す．

- 世界的に患者が広まっているＳＡＲＳに関して，２００３年４月３日から上海の各種新聞，テレビ，ラジオも報道し始めた．３月下旬ごろに１度「上海では確認されていない.」

という報道がされて以来だ．報道によれば現在上海市では一人の患者がＳＡＲＳの疑いがあるということで病院に隔離されている．

- 静岡市を流れる安倍川流域で，江戸時代に「友釣り禁止令」が出された．友釣りという漁法が流行して，その面白さに若者が熱中した.「肝心な農業を放棄して困る」というのが，禁止令の理由である．百四十年以上も昔から，友釣りに夢中だったお国柄のことだ．本県が今「友釣り人口日本一」といわれる理由は，このように歴史が証明している．アユ特集の県内各河川漁協で見るように，今年は天然アユが極めて大量に川を上った．
- 診察室で医療者がみる姿かたちだけでなく，患者の目や心に映るこうした波風を理解し，それに基づいて日々の医療を創りあげて行くことが，いわゆるＱＯＬを重視した医療であろう．一人ひとりの患者のＱＯＬを知るには，基本的には彼女に問いかけ，話をよく聴く以外に術はない．

次に，話し言葉コーパスの一部を示す．

- 美味しいキムチを食べましょ～！最近よくお客さんから質問される事があります.「スーパーでよくキムチを買うんやけどスーパーで売っているキムチって何で酸っぱいの？」結構こんな印象を持ってる人がたくさんいますよね．いつも聞かれたらこう答えるようにしています.「スーパーで売ってるキムチって大半がキムチと違う物やからやで」
- 今日、久しぶりにガンダムやってたねー．新聞見りゃ判ることだけど，一応教えとこと思って．電話したら，繋がりませんでした，ＰＨＳ．山奥へ走りに行ってたのかな？
- これに２５６ＭＢのメモリ乗せてマシンに負担のかかることをするのが楽しい．もうほとんどホビーだね，この感覚は．だってすぐベンチとりたくなるしね．やっぱり新しいシステムってわくわくするんだよ．クロックアップはしていない．ほら、いちおう業務マシンだから．
- 谷口：私ぐらいの歳になるとね … なかなか見つからないよ～そう言う黒田君　最近仕事はどう？忙しい？
  黒田：仕事の量は同じですけどネ～単価を値切られて困ってます．同じだけの枚数書いても半分の値段ですから …
  谷口：でもあるだけでもイイじゃないの？世の中不景気だからネ～

議論　話し言葉として収集されたコーパス (115 ページ) を分析したところ，多くの部分は，掲示板 (チャットも含める) や個人の日記のページから収集されていることが分かった．29 ページが掲示板から，18 ページが個人の日記から収集されていた．この結果は，我々の置いた仮説と一致する．また，具体例の最後にあるような，発話や対談の書き起こし (または書き起こし風) のページが 10 ページ含まれていた．

提案手法は，親愛表現と丁寧表現を全く含んでいないページを，書き言葉コーパスとして収

集する．しかし実際には (5a) と (5b) のように，親愛表現と丁寧表現を全く含んでいないにもかかわらず，話し言葉的である文は存在する．提案手法は，こうした文をうまく扱うことができず，それが収集の適合率を下げる大きな原因となっていた．

5. a　昔から買ってあるのに読んでいない本を ちんたら 読み進める．
   b　前はこんな せこい こと言う店ではなかったのに．

今後は，今回の実験で収集された書き言葉コーパスと話し言葉コーパスを利用して，こうした文も扱えるようにしていきたい．具体的な方法としては，収集されたコーパスと (Kilgarriff 2001) などで議論されている手法を使い，書き言葉や話し言葉に特有の語彙を自動学習し，それを利用することを考えている．例えば,「ちんたら」や「せこい」といった語が話し言葉に特有の語彙であることを学習できれば，(5a) と (5b) は，いずれも話し言葉的な文であると判断することができる．

こうした改良を行えば，収集の精度が上がるだけでなく，収集されるコーパスの規模を増やす効果も期待できる．今回の実験では，WWW コーパス 336,341 ページのうち，書き言葉又は話し言葉コーパスとして収集されたのは，71,658(38,472+33,186) ページであったが，さらに大規模なコーパス収集が可能になると考えられる．

5　言い換えペアの選択

収集された書き言葉コーパスと話し言葉コーパスを用いて，3 節で学習された言い換えペアの中から，source が書き言葉特有語彙で target が話し言葉語彙であるような言い換えペアを選択する．このような言い換えペアを正例，それ以外の言い換えペアを負例と考えれば，解くべき問題は二値分類であると考えることができる．

本論文は，Support Vector Machine(Vapnik 1995) を用いた手法を提案する．ある表現が書き言葉特有語彙であるか話し言葉語彙であるかは，その表現の書き言葉コーパスでの出現確率と話し言葉コーパスでの出現確率から判断できると考えられる．そこで，SVM に与える素性は，以下の 4 つの出現確率を用いた．各素性は 0 から 1 までの連続値をとる．

(1)　source の書き言葉コーパスでの出現確率
(2)　source の話し言葉コーパスでの出現確率
(3)　target の書き言葉コーパスでの出現確率
(4)　target の話し言葉コーパスでの出現確率

5.1 出現確率

以下では，ある表現 e のコーパスでの出現確率 ($P(e)$) の計算方法を説明する．上記の 4 つの素性は，この計算方法にしたがって求める．

出現頻度　$P(e)$ を求めるためには，まず，出現頻度 ($F(e)$) を求める必要がある．基本的には，コーパスを形態素解析，構文解析すれば求まるが，以下の三点に留意した．なお，形態素解析には JUMAN，構文解析には KNP を用いた．

まず第一に，同じ用言でも態が異なれば，異なる用言として出現頻度を数えた．これは，用言の中には使役や受身などの態の違いによって出現頻度が大きく異なるものがあるからである．例えば「漂う」と使役形「漂わせる」では，出現頻度が大きく異なる．

次に，データスパースネスの問題に対処するため，副詞を無視するという近似を行った．3 節ですでに述べたように，言い換えペアの target は「副詞* 体言+ 用言」という形をしている．e が target で，たくさんの副詞や体言を含んでいるときには，すべて頻度が 0 になってしまうことが考えられる．そこで，e の頻度を求めるさいには，副詞を無視するという近似を行った．例えば「早く走る」の出現頻度は，コーパス中の「走る」の出現頻度で近似する．副詞ではなく体言を無視するという近似も考えられるが，(6a) や (6b) の下線部のように体言と用言は慣用句を形成することがあり，そうした場合に体言を無視してしまうと意味が大きく変わってしまうので，副詞を無視するという近似を採用した．

6. a　**入手する**
　　手に入れる こと
　b　**憤る**
　　腹を立てる

2 節で学習された 5,836 の言い換えのペアのうち，target が副詞を含んでいるものは 1436 個，体言を含んでいるものは 839 個であった．

最後に，構文解析結果はすべて利用するのではなく，信頼度の高い部分だけを利用した．e が体言を含んでいる場合に $F(e)$ を求めるには，コーパス中の体言と用言の係り受け情報が必要なので，構文解析が必要となる．しかし，構文解析精度 (90 %) は形態素解析の精度 (99 %) と比べると低く，誤りが多い．そこで (Kawahara and Kurohashi 2001) と同様のヒューリスティクスを使って，構文解析の信頼度の高い部分だけを利用した．(Kawahara and Kurohashi 2001) の報告によると，このヒューリスティクスによって信頼度が高いと判断された部分では，構文解析の精度は 97 %である．

出現確率　$F(e)$ から $P(e)$ を求める．一般に，$P(e)$ は次のような式で定義される．

$P(e) = F(e)/$ コーパス中の全表現数

表 4　各コーパスの用言数と「体言-用言」数

	用言数	「体言-用言」数
書き言葉	3,169,253	769,876
話し言葉	5,186,414	849,905

e が名詞を含んでいない場合，$F(e)$ を計算するのにコーパス全体を使用するのに対して，e が名詞を含んでいる場合，コーパスは解析の確信度の高い部分しか使わない．そのため，e が名詞を含んでいる場合と含んでいない場合で，上式の分母「コーパス中の全表現数」の値を変えるべきである．このことを踏まえて，$P(e)$ を次のように計算する．

e が名詞を含まない

$P(e) = F(e)/$ コーパス中の用言数

e が名詞を含む

$P(e) = F(e)/$ コーパス中の「体言-用言」数

「体言-用言」は，係り受けの関係にある体言と用言のペアをあらわす.「体言- 用言」数は，$F(e)$ と同様に，構文解析の信頼度が高い部分だけで数える．

表 4 に，学習された書き言葉コーパスと話し言葉コーパスにおける用言数と「体言-用言」数を示す．いずれのコーパスでも，用言数は「体言-用言」数よりも非常に大きな値となっており，上記のように 2 通りの計算方法を用意する必要があることが分かる．

5.2　評価

2 人の被験者が SVM で学習するためのデータセットを作成し，提案手法の評価を行った．

データセット　データセットは以下のように作成した．まず Kaji らの手法を用いて，例解小学国語辞典 (田近 1997) から言い換えペアを自動学習した．そして，各言い換えペアを 2 人の被験者が正例と負例に分類して，2 人の被験者の判断が一致した 200 の言い換えペアをデータセット (正例が 70 個，負例が 130 個) とした．

データセットを作成するために要した言い換えペアは 247 個であった．つまり，被験者の判断が一致しなかった言い換えペアが 47 個あった．被験者の判断の一致度を示す Kappa 統計量を求めたところ 0.627 であった．以下に，被験者の判断が一致しなかった例を示す．

7.　　殺到する → 一度にどっと、おしよせる

(7) の判断が食い違ったのは，target がまわりくどい表現であることに一因があると考えられる．target は定義文から自動抽出されているが，定義文は一般のテキストよりもまわりくどい傾向があり，target にも同様の傾向がみられる．そうした表現を自然な話し言葉と考えるかど

表 5　実験結果

	linear	poly2	poly3
精度	72 %	**79 %**	76 %
適合率	60 %	**69 %**	63 %
再現率	59 %	**72 %**	62 %
F 値	59 %	**70 %**	62 %

うかが，判断を別れさせる要因になったと考えることができる．

実験結果　20 分割の交差検定によって評価を行った．実装には学習パッケージ TinySVM[5]を用いた．カーネル関数を用いない場合 (linear) と，カーネル関数に 2 次，3 次の多項式関数を使った場合 (poly2，poly3) で実験を行った．表 5 に，それぞれの場合の正例と負例の分類精度，正例に分類された言い換えペアの適合率，正しく正例に分類された言い換えペアの再現率の値を示す．また，F 値もあわせて示す．各値は交差検定を行ったときの平均値となっている．カーネル関数に 2 次の多項式を使った場合の精度が最も高く，そのときの精度は 79 %であった．交差検定を行ったときの F 値のばらつきを調べたところ，標準偏差は 23 であった．ばらつきは大きいが，これは，20 分割した一つ一つのデータセットの規模が小さいからであると考えることができる．

議論　SVM に与える素性として使った出現確率は，自動収集された書き言葉コーパスと話し言葉コーパスから求めている．つまり，素性の値は自動学習されている．さらに，今回用意したデータセットも比較的小規模なものである．このことを踏まえると，実験結果の 79 %は十分に高い精度であり，使用した 4 つの素性は有効に働いていると考えることができる．

求められた出現確率の中には，人間の直感と反する不適切な値が計算されたものがあり，精度を下げている原因となっていた．そのような例として「観劇する」がある.「観劇する」は，書き言葉特有語彙であると考えることができるが，書き言葉コーパスで 4 回，話し言葉コーパスで 43 回出現していた．これを出現確率になおすと，話し言葉コーパスでは書き言葉コーパスの約 8 倍出現していたことになる．これは，コーパスを収集した WWW に一因があると考えられる．つまり，観劇に関する Web ページの大部分は劇好きの掲示板などであり，ニュース記事などは数が少ないと考えることができる．そのため，収集された書き言葉コーパスは観劇の話題を殆んど含まず,「観劇する」が話し言葉コーパスの方により多く出現したと考えることができる．

書き言葉コーパスと話し言葉コーパスの 2 コーパスを利用するという提案手法の有効性を確かめるために，次のような 2 つの実験を行い，その結果を提案手法と比較した．

5 http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/

表 6　具体例

分類結果	言い換えペア	素性			
		source		target	
		書き言葉	話し言葉	書き言葉	話し言葉
正例	商う → 商売をする	4/3,169,253	6/5,186,414	15/769,876	21/849,905
	激化する → はげしくなる	151/3,169,253	17/5,186,414	152/3,169,253	88/5,186,414
	受諾する → 引き受ける	50/3,169,253	5/5,186,414	280/3,169,253	259/5,186,414
	発汗する → 汗をかく	1/3,169,253	5/5,186,414	50/769,876	119/849,905
	×きざだ → 気取っている	1/3,169,253	0/5,186,414	40/3,169,253	129/5,186,414
	×上京する → 東京へ行く	80/3,169,253	132/5,186,414	3/769,876	23/849,905
負例	へばる → へとへとに疲れる	6/3,169,253	36/5,186,414	32/3,169,253	366/5,186,414
	食事する → 食べる	27/3,169,253	170/5,186,414	3,114/3,169,253	17,239/5,186,414
	優先する → 先に扱う	312/3,169,253	249/5,186,414	2,008/3,169,253	2,125/5,186,414
	引越しする → 転居する	3/3,169,253	89/5,186,414	35/3,169,253	14/5,186,414
	×伝聞する → 伝え聞く	0/3,169,253	1/5,186,414	16/3,169,253	14/5,186,414
	×軟化する → 軟らかくなる	25/3,169,253	5/5,186,414	11/3,169,253	2/5,186,414

- 日本語には漢語と和語があり，大雑把にいうと前者は書き言葉特有語彙であり，後者は話し言葉語彙であると言える．そこで，source が漢語で target が和語である言い換えペアを正例，それ以外を負例に分類した場合の，精度，適合率，再現率を求めた．ここでは，用言がサ変名詞であれば漢語，それ以外ならば和語とした．実験の結果，精度，適合率，再現率はそれぞれ，59％，40％，39％であった．
- 書き言葉コーパスと話し言葉コーパスの 2 コーパスを利用しなくとも，source と target の単純な使われやすさが分かれば，正例と負例をうまく分類できると考えることもできる．そこで，収集した書き言葉コーパスと話し言葉コーパスを 1 つにまとめた混合コーパスを使って，提案手法と同様に SVM を使って分類を行った．ここで素性に使ったのは，source の混合コーパスでの出現確率と，target の混合コーパスでの出現確率の 2 つである．この二つの素性によって，source と target の単純な使われやすさを学習できると考えられる．実験の結果，カーネル関数に 3 次の多項式関数を使った場合の結果が最も良かったが，精度，適合率，再現率はそれぞれ，71％，57％，46％であった．

提案手法は，上記の手法をいずれも上回る結果となり，書き言葉コーパスと話し言葉コーパスを利用することの優位性を示すことができた．

具体例　表 5.2 に，正例に分類された言い換えペアと，負例に分類された言い換えペアの具体例と，それらの言い換えペアのもつ素性の値を示す．素性の列は左から，source の書き言葉コーパスでの出現確率，source の話し言葉コーパスでの出現確率，target の書き言葉コーパスでの出現確率，target の話し言葉コーパスでの出現確率を表す．出現確率は「出現頻度/コーパスの全表現数」という表記になっている．target の「コーパスの全表現数」は，target が体言を含

む場合とそうでない場合で違った値になっている．×印は，言い換えペアの分類結果が間違っていることを表す．

「商う → 商売をする」「食事する → 食べる」「優先する → 先に扱う」などは，source と target が漢語でるか和語であるか，だけに注目していたのでは，正しく分類するのが難しい例であるが，提案手法ではうまく分類できている．

今後の課題　畠の分類からも明らかなように，一口に話し言葉といっても様々な種類が存在する．本論文では，冗長性の高い話し言葉のみを扱っており，この問題に深く立ち入らなかったが，今後は話し言葉のより詳細な分類を行い，その分類も考慮した言い換えに取り組む予定である．

本論文で扱った用言の言い換え以外にも，複合名詞の言い換えや統語構造の変換なども「書き言葉から話し言葉への言い換え」に必要であり，今後，取り組んでいきたい．

6　まとめ

本論文では，話し言葉特有語彙から話し言葉語彙への言い換えを学習する方法について述べた．提案手法は，書き言葉コーパスと話し言葉コーパスを WWW から自動収集して，それらのコーパスでの出現確率を利用するものである．実験の結果，書き言葉コーパスと話し言葉コーパスの収集精度は 94 %，言い換え学習の精度は 79 %であり，提案手法の有効性を確認することができた．

参考文献

Barzilay, R. and Lee, L. (2003). “Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment.” In *Proceedings of HLT/NAACL2003*.

Bulyko, I., Ostendorf, M., and Stolcke, A. (2003). “Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures.” In *Proceedings of HLl/NAACL2003*, pp. 7–9.

Duclaye, F. and Yvon, F. (2003). “Learning Paraphrases to Improve a Question-Answering System.” In *Proceedings of the 10th Conference of EACL Workshop Natural Language Processing for Question-Answering*.

Edmonds, P. and Hirst, G. (2002). “Near-Synonymy and Lexical Choice.” *Computational Linguistics*, **28** (2), pp. 105–144.

Fukuhara, T., Nishida, T., and Uemura, S. (2001). “Public Opinion Channel: A System for Augmenting Social Intelligence of a Community.” In *Workshop notes of the JSAI-Synsophy International Conference on Social Intelligence Design*, pp. 22–25.

畠弘巳 (1987). “話しことばの特徴ー冗長性をめぐってー.” 「国文学 解釈と鑑賞」, **52** (7), pp. 22–34.

Hayashi, M., Ueda, H., Kurihara, T., Yasumura, M., Douke, M., and Ariyasu, K. (1999). “TVML (TV program Making Language) - Automatic TV program Generation from Text-based Script -.” In *ABU Technical Review*.

Hermjakob, U., Echihabi, A., and Marcu, D. (2002). “Natural Language Based Reformulation Resource and Web Exploitation for Question Answering.” In *Proceedings of TREC2002 Conference*.

Inkpen, D. Z. and Hirst, G. (2001). “Building a Lexical Knowledge-Base of Near-Synonym Differences.” In *Proceedings of Workshop on WordNet and Other Lexical Sources*, pp. 47–52.

Inui, K., Fujita, A., Takahashi, T., Iida, R., and Iwakura, T. (2003). “Text Simplification for Reading Assistance: A Project Note.” In *Proceedings of the Second International Workshop on Paraphrasing*, pp. 9–16.

Inui, K. and Yamamoto, S. (2001). “Corpus-Based Acquisition of Sentence Readability Ranking Models for Deaf People.” In *Proceedings of NLPRS2001*.

Kaji, N., Kawahara, D., Kurohashi, S., and Sato, S. (2002). “Verb Paraphrase based on Case Frame Alignment.” In *Proceedings of ACL2002*, pp. 215–222.

Kawahara, D. and Kurohashi, S. (2001). “Japanese Case Frame Construction by Coupling the Verb and its Closest Case Component.” In *Proceedings of HLT2001*, pp. 204–210.

Kilgarriff, A. (2001). “Comparing Corpora.” *International Journal of Corpus Linguistics*.

Lin, D. and Pantel, P. (2001). “Discovery of inference Rules for Question Answering.” *Journal of Natural Language Engneering*, **7** (4), pp. 343–360.

Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). “Spontaneous Speech Corpus of Japanese.” In *Proceedings of LREC2000*, pp. 947–952.

村田真樹，井佐原均 (2001). “言い換えの統一的モデルー尺度に基づく変形の利用ー.” 言語処理学会第 7 回ワークショップ論文集, pp. 21–26.

Murata, M. and Isahara, H. (2002). “Automatic Extraction of Differences between Spoken and Written languages, and Automatic Translation from the Written to the Spoken Language.” In *Proceedings of the LREC2002*.

Nadamoto, A., Kondo, H., and Tanaka, K. (2001). “WebCarousel: Restructuring Web Search Results for Passive Viewing in Mobile Environments.” 7th International Conference on Database Systems for Advanced Applications, pp. 164–165.

佐藤理史 (1999). “論文表題を言い換える.” 情報処理学会論文誌, **40** (7), pp. 2937–2945.

田近洵一（編） (1997). 例解小学国語辞典. 三省堂.

Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., and Yamamoto, S. (2002). “Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world.” In *Proceedings of LREC2002*, pp. 147–152.

Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Tambouratzis, D., and Carayannis, G. (2000). “Discriminating the registers and styles in the Modern Greek language.” In *Proceedings of Workshop on Comparing Corpora 2000*.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

略歴

鍛治 伸裕: 2000 年京都大学工学部電気電子工学科卒業．2002 年京都大学大学院情報学研究科修了．現在，東京大学大学院情報理工学系研究科博士後期課程在学中．自然言語処理の研究に従事．

岡本 雅史: 1994 年早稲田大学政治経済学部政治学科卒業．1996 年京都大学大学院人間・環境学研究科博士前期課程修了．1999 年同大学院博士後期課程単位所得認定退学．現在，東京大学大学院情報理工学系研究科学術研究支援員．博士 (人間・環境学)．認知言語学，語用論，コミュニケーション論の研究に従事．

黒橋 禎夫: 1989 年京都大学工学部電気工学科第二学科卒業．1994 年同大学院博士課程修了．京都大学工学部助手，京都大学大学院情報学研究科講師を経て，2001 年東京大学大学院情報理工学系研究科助教授，現在に至る．自然言語処理，知識情報処理の研究に従事．

(2004 年 1 月 9 日 受付)
(2004 年 5 月 21 日 再受付)
(2004 年 7 月 5 日 採録)