

(51) Int.Cl. ⁷	識別記号	F I	テ-マコード [*]	(参考)
G06F 17/21	550	G06F 17/21	550	L 5B009
3/16	320	3/16	320	H 5B091
17/28		17/28		X 5D015
G10L 15/00		G10L 3/00	551	B
15/18			537	D

審査請求 有 請求項の数 3 O L (全 8 頁) 最終頁に続く

(21)出願番号	特願2001 - 324096(P 2001 - 324096)	(71)出願人	301022471 独立行政法人通信総合研究所 東京都小金井市貫井北町 4 - 2 - 1
(22)出願日	平成13年10月22日(2001.10.22)	(72)発明者	村田 真樹 東京都小金井市貫井北町 4 - 2 - 1 独立 行政法人通信総合研究所内
		(72)発明者	井佐原 均 東京都小金井市貫井北町 4 - 2 - 1 独立 行政法人通信総合研究所内
		(74)代理人	100090893 弁理士 渡邊 敏

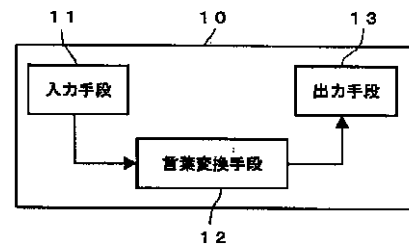
最終頁に続く

(54) 【発明の名称】話し言葉の書き言葉への変換装置

(57) 【要約】

【課題】 話し言葉で入力した文章等を書き言葉の文章等で出力することができる話し言葉の書き言葉への変換装置を提供すること。

【解決手段】 話し言葉の入力手段 1 1 と、前記入力手段により入力された話し言葉を書き言葉に変換する言葉変換手段 1 2 と、前記言葉変換手段により変換された書き言葉を出力する出力手段 1 3 とを有するものである。また、前記言葉変換手段は、話し言葉と書き言葉のデータを用いているものが好ましく、更に、前記言葉変換手段は、話し言葉と書き言葉の差分部分のデータを用いているものが好ましい。



【特許請求の範囲】

【請求項 1】 話し言葉の入力手段と、前記入力手段により入力された話し言葉を書き言葉に変換する言葉変換手段と、前記言葉変換手段により変換された書き言葉を出力する出力手段と、を有することを特徴とする話し言葉の書き言葉への変換装置。

【請求項 2】 前記言葉変換手段は、話し言葉と書き言葉のデータを用いていることを特徴とする請求項 1 記載の話し言葉の書き言葉への変換装置。

【請求項 3】 前記言葉変換手段は、話し言葉と書き言葉の差分部分のデータを用いていることを特徴とする請求項 1 又は請求項 2 記載の話し言葉の書き言葉への変換装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】この発明は、入力された話し言葉を書き言葉に変換し、変換後の書き言葉により表された文章等の保存、表示、プリント等の出力機能を備えた話し言葉の書き言葉への変換装置に関するものである。

【0002】

【従来の技術】話し言葉と書き言葉には違いがあり、例えば、書き言葉の「データ」、「え」と「=」が、話し言葉でそれぞれ「データー」、「えー」と「は」となったり、また話し言葉では「という」をいれて柔らかく言う場合等がある。話し言葉と書き言葉の差分を抽出し、前記のような話し言葉と書き言葉の違いを調べ、その差分結果により作成した書き言葉から話し言葉への変形規則を用い、書き言葉から話し言葉へ自動で言い換える方法に関しては、既に論文等で発表された周知の技術である。また、前記変形規則を用い、話し言葉から書き言葉へ自動で言い換えることも同様な方法により可能である。

【0003】前記話し言葉から書き言葉への変換は、様々な人の話し言葉を誰でも容易に理解できる書き言葉として残したり、話し言葉で入力したものを書き言葉の原稿として容易に得たり、テレビ、映画等において話す言葉を逐次画面の端等に表示する等に利用される。

【0004】また、話し言葉の入力手段としては、パソコン等に搭載された音声認識機能等があり、マイクを通して入力された話し言葉の音声から、単語、文章等の話し言葉データに変換する方法等、他にも多種多数の従来技術がある。前記変換後の単語、文章等の話し言葉データは、パソコン等のメモリに一時記憶あるいはハードディスク等の記録媒体に保存され、必要に応じた処理が行われる、尚、話し言葉は、マイクを通してのみではなく、磁気テープ等の記録媒体に記録された音声を再生する装置からの出力をパソコン等に入力し、前記同様に話し言葉データとして得られる場合もある。

【0005】また、パソコン等のキーボードにより、音声を聞きながらあるいは映画の台本、原稿等に話し言葉で記述された文章等をパソコン等に入力する方法もある。前記入力された文章等の話し言葉データは、パソコン等のハードディスク等の記録媒体に保存され、必要に応じて処理される。

【0006】尚、前記パソコン等では、前記記録媒体に記録する出力形態ばかりでなく、前記記録された文章等の話し言葉データを、専用のソフトウェアあるいは市販のワードプロセッサ等のソフトウェアを用いて CRT 等の表示装置に表示したり、プリンターによりプリントとして出力することができる。

【0007】

【発明が解決しようとする課題】しかしながら、前記従来のパソコン等では、入力された単語、文章等の話し言葉データが、そのまま話し言葉の単語、文章等として出力されるものであり、入力された話し言葉が、書き言葉に変換されて出力されるものがないという問題点があった。

【0008】本発明は、前記従来の問題点を解決するためになされたもので、話し言葉で入力した文章等を書き言葉の文章等で出力することができる話し言葉の書き言葉への変換装置を提供することである。

【0009】

【課題を解決するための手段】本発明は、話し言葉の入力手段と、前記入力手段により入力された話し言葉を書き言葉に変換する言葉変換手段と、前記言葉変換手段により変換された書き言葉を出力する出力手段とを有するものである。また、前記言葉変換手段は、話し言葉と書き言葉のデータを用いているものが好ましく、更に、前記言葉変換手段は、話し言葉と書き言葉の差分部分のデータを用いているものが好ましい。

【0010】

【発明の実施の形態】以下、図面を参照して、本発明の実施形態について説明する。図 1 は、本発明による話し言葉の書き言葉への変換装置の実施形態を示す図である。

【0011】図 1 に示すように、本発明の実施形態の話し言葉の書き言葉への変換装置 10 は、入力手段 11 と、言葉変換手段 12 と、出力手段 13 とを有するコンピュータとして構成されている。入力手段 11 は、話し言葉で記述された文章等、例えば、映画の台本、講演記録、会話の記録等を直接入力するキーボード、あるいはマイクから音声として入力若しくは磁気テープ等の記録媒体に記録された話し言葉の単語、文章等を再生する再生装置の出力を前記マイクと同様に音声として入力し、前記入力された音声又は再生装置の出力からの音声を単語、文章等の話し言葉データに変換するソフトウェアからなるものである。

【0012】前記入力された音声又は再生装置の出力が

らの音声を話し言葉データに変換する手段として、専用の変換ソフトウェアを用いる場合は、既知の技術である次に示す音声認識による音声処理技術を利用することができる。該音声処理技術は、一般的には、入力された音声波形から音素認識（記号化）、音声単位へのセグメンテーション、音響分析等の音響処理を行い、次に予測単語の認識、重要な自立語などのキーワードをボトムアップ的に抽出するワードスポッティング等の単語認識を行い、最後に意味解析、構文解析等の言語処理を行い話し言葉データを生成するものである。

【0013】尚、既存の音声認識システム等の音声から話し言葉データに変換するものを利用することも可能であり、この場合、言葉変換手段12で用いるファイル形式として、前記既存の音声認識システムにおける文章等のファイル形式と共通のものをを用いればよい。

【0014】言葉変換手段12では、話し言葉と書き言葉の対応関係を規定する言い換え（以下パラフレーズという）テーブルを予め作成しておく。パラフレーズの作成は、日本語の講演発表の音声を書き起こし、形態素情報を付与した日本語話し言葉集成（以下コーパスという）を話し言葉データベースとして、講演発表の元となる論文の電子化データを書き言葉データベースとして、それぞれ用いる。即ち、話し言葉と書き言葉のデータを用いる。次に、話し言葉データベースと書き言葉データベースの一致部分及び不一致部分を調べ、話し言葉データから書き言葉データへの変換規則を獲得しておく。即ち、パラフレーズテーブルの作成に、話し言葉と書き言葉の差分部分のデータを用いる。

【0015】言葉変換手段12では、図2に示すアルゴリズムを用いて話し言葉から書き言葉への言い換え（変換）を行う。このアルゴリズムは、大雑把に言うと、前記書き言葉データベースでの頻度が大きくなるように書き換える、つまり、書き言葉データベースで出現し易い表現に書き換えるものである。

【0016】先ず、入力された話し言葉データを、例えば、形態素解析プログラムであるJUMANを用いて形態素解析して、形態素列に分解する。そして、文頭の形態素から順に、形態素ごとに次の処理を行う。

【0017】先ず、現在の形態素で始まる形態素列S（形態素を一つも持たない場合、つまり空文字列も含む）と、前記変換規則を得たときの差分データの話し言葉の文字列A_iが一致した場合、その差分部分のデータR_iが規則として用いられる。そして、前記差分部分データの書き言葉の部分（文字列B_i）が、書き言葉の候補となる。また、Sの前後k-gramの形態素列S_{1i}、Sの後節k-gramの形態素列をS_{2i}とする。尚、前記kは、定数である。

【0018】次に、前記B_iに対して、前記書き言葉データベースでのS_{1i}、S_{2i}の文字列の頻度を求め、該頻度が最も大きかった時のiをmとする。

【0019】次に、書き言葉データベースでのS_{1i}、S_{2i}の文字列の頻度を求め、この値よりもS_{1i}、B_i、S_{2i}の文字列の頻度の方が大きい時、A_iをB_iに置き換え、処理を次の形態素に移す。

【0020】以上の処理を文末の形態素まで行うことにより、話し言葉データから書き言葉データへの書き換え（変換）を行うことができる。

【0021】出力手段13では、言葉変換手段12で変換して得られた書き言葉データを、所定のファイル形式でハードディスク等の記録媒体に保存したり、CRT等の表示手段に書き言葉データの表す文章等を表示したり、プリンターで前記文章等をプリントする等の出力を行う。

【0022】話し言葉の文章等の入力から書き言葉の文章等の出力まで、専用の処理プログラムを用いて行うように構成することもできるし、専用の処理プログラムと市販のワードプロセッサ等の処理プログラムとを用いて行うように構成することも可能である。

【0023】尚、前記表示手段として、テレビ画面、映画画面等であってもよい。また、前記所定のファイル形式を、市販のワードプロセッサで取り扱えるファイル形式とすることにより、容易に変換後の書き言葉での文章等を表示したり、プリントしたりすることができる。

【0024】以上、本発明の実施形態の話し言葉の書き言葉への変換装置を用いると、映画あるいはテレビにおける台詞、解説等をほぼ同時に書き言葉として読みやすく、理解しやすい文章等で映画あるいはテレビの画面の端に表示できる等、気軽に使っている話し言葉から、形式的である故誰でも理解し易い書き言葉の文章等に容易に変換できる。

【0025】

【実施例】次に、言葉変換手段12において用いる話し言葉と書き言葉の差分を調べた実施例について説明する。

【0026】差分をとる話し言葉と書き言葉のデータの形態素解析を行い、図3に示すように、各形態素が各行に分かれた形にデータを变形する。前記形態素解析は、形態素解析プログラムであるJUMANを用いて行う。

【0027】次に、前記データを照合し、話し言葉と書き言葉のデータの差分部分と一致部分の検出を行う。ここでは、UNIX（登録商標）コマンドのdiffを用いて行う。得られた結果を図4に示す。

【0028】図3において、セミコロンで始まる行は一致部分、差分部分を示すためのもので、「 ; 」から「 ; 」までの部分は、話し言葉データでのみ出現したものの、「 ; 」から「 ; 」

までの部分は、書き言葉のみ出現したものの、「 ; 」から「 ; 」までの部分は、話し言葉と書き言葉でともに出現したものを意味する。ここで取り出したいものは、話し言葉と書き言葉の違いで

あるので、「 ; 」から「 ; 」までの部分となり、差分部分は図5のようになる。

【0029】前記図5の「え今日は」「本論文では単語の羅列を」は、話し言葉で「え今日は」とあったのが、書き言葉では「本論文では単語の羅列を」となったということを意味する。この結果では、差分部分のデータとして精度が悪すぎるので、diffの結果から、ある程度よさそうな話し言葉と書き言葉の差分部分を抽出する。ここでは、1.珍しい(出現頻度の低い)文字列に囲まれた不一致部分ほど、パラフレーズとしては確からしい、2.複数箇所に出現した不一致部分ほど、パラフレーズとしては確からしい、という二つの特徴を利用する。

【0030】先ず、前記特徴1「珍しい文字列に囲まれ

$$P(S2) \simeq (d+1) * \frac{\text{文字列}S2\text{の出現数}}{\text{文字総数}}$$

【0033】この時、差分部分が確からしい確率をP(差分、S1、S2)とすると、P(差分、S1、S2)はS1、S2がともに図6に示すような形で現れ難い確率であると仮定すると、S1とS2が独立であることを仮定して、数3のようになる。

【0034】

【数3】

$$P(\text{差分}, S1, S2) \simeq (1 - P(S1))(1 - P(S2))$$

【0035】次に、前記特徴2「複数箇所に出現した不一致部分ほど、パラフレーズとしては確からしい」の方を考える。これは、複数箇所での確率をうまく組み合わせればよい。複数箇所のうち一箇所でも正しければ、その差分部分は正しいものとして抽出できると考える。即ち、差分部分が正しい事象は、任意のS1、S2に対してS1、S2に囲まれる差分部分が全て確からしくない場合の余事象なので、差分部分が確からしい確率をP(差分)とすると、P(差分)は、各差分部分が独立であることを仮定して、数4で表される。

【0036】

【数4】

$$P(\text{差分}) \simeq 1 - \prod_{S1, S2} (1 - P(\text{差分}, S1, S2))$$

【0037】差分部分の取り出しは、diffの結果を前記数4のP(差分)の値でソートし、その値の大きいものから取り出すことによって行われる。

【0038】尚、最初の差分部分の候補の取り出しについては、次に示す改良を行うことができる。図7に示すように、一致部分と差分部分が出現している時に、“「差分部分1」「一致部分1」「差分部分2」”、“「差分部分1」「一致部分1」「差分部分2」「一致部分2」「差分部分3」”といったものも差分部分の候補とする。

【0039】前記改良は、単に「差分部分1」だけで

た不一致部分ほど、パラフレーズとしては確からしい」の方を考える。ここでは、差分部分(不一致部分)が、図6に示すように、一致部分である文字列S1、S2にはさまれている、S1とS2の間がd文字以内に図7中の方向にS2及びS1が現れる確率を、P(S1)、P(S2)とすると、P(S1)とP(S2)は、それぞれ近似的に数1と数2で表される。

【0031】

【数1】

$$P(S1) \simeq (d+1) * \frac{\text{文字列}S1\text{の出現数}}{\text{文字総数}}$$

【0032】

【数2】

は、「一致部分0」「一致部分1」から求まるP(差分)の値が小さくて取り出せないような時も、“「差分部分1」「一致部分1」「差分部分2」を差分部分と考えることで、「一致部分0」「一致部分2」から求まるP(差分)の値が大きくなって取り出しうるという効果を持つ。ここでは、この連結によって生成する差分部分は、元の差分部分を3個以下しか含まないものに限る。

【0040】次に、話し言葉、書き言葉のデータとして、開放融合プロジェクトにおける82編の学会講演の論文の電子版を利用した。話し言葉データは、開放融合コーパス(集成)の内、前記論文データに対応するもの(330679文字)である。書き言葉データは、前記82編の論文データ(打ち込み、352660文字)である。

【0041】話し言葉データには、図8に示すタグが埋め込まれていたもので、次の処理を施した。基本的に各タグのリストの第二要素をタグの代りに本分に埋め込む。例えば、“(F あの)”の場合、“あの”を本分の該当箇所に挿入する。但し、セミコロンで区切られているものについては、最後のものを、カンマで区切られているものについては、最初のものを用いる。

【0042】フィルターや言い直しなどは省いた方がよいとも考えられるが、あえてそのような表現も差分部分として抽出することを目的として残した。

【0043】書き言葉データとして用いる論文データには、表題や著者名、所属なども含まれているが、これらはそのまま残して利用した。

【0044】以上の条件で話し言葉データと書き言葉データの差分部分を抽出した。抽出総数を図9に示す。図9における確率値は、数4で算出された値である。差分部分を前記数4で算出される値で分類した結果の上位500個を図10に示す。図10における頻度は差分部分の

出現回数である。

【0045】前記図10における「データ」「データ」の食い違いは、コーパス(集成)の定義によるもので、書き言葉で「データ」と書くが、話し言葉で「データ」と伸ばして発音し易いということを意味しているものではない。また、話し言葉で“<C>”が得られているが、これはコーパスにおいて単語の途中を意味するタグで、これが得られてもあまり意味はない。その他目立つものとしては、「え」「えー」などのフィラーが検出できていたり、「=」は「は」と読むということが分かったり、話し言葉では「という」をいれて柔らかくいう場合があることが分かる。

【0046】抽出された差分結果を分析したところ、主に次のものがあった。

1. 表記ゆれ

表記ゆれの例を図11に示す。これはコーパスの定義にも関係するところであるが、ここでは大規模コーパスを使用してコーパスで多く出現するものを自然な表現として定義した。

【0047】2. 表記・読みを与えるもの

この例を図12に示す。図12により「=」は「は」と読めばよいとか、「s」は「秒」を意味するときと「S」を意味する時があるなどが分かる。

【0048】3. 同義関係のもの

この例を図13に示す。論文に書いていたことを少し違えて言ったりするために、図13のような同義な意味を示す言い換え表現を獲得することができる。ここでは、研究がらみの同義表現が得られている。

【0049】4. 口語調のもの

この例を図14に示す。話し言葉で丁寧語にするものから、「。」と書いているところを「訳ですが」と文をつなげるものなども得られている。また、最後の行に「これ」が得られているが、これは「明瞭に発声したもの(これ)を」という形で使われていた。

【0050】5. 省略をしているもの

この例を図15に示す。話し言葉の方では「処理」を省いて言ってみたり、データの値を「11.25」を「11.3」に丸めていってみたり、はしょっているところがある。

【0051】6. 補完をしているもの

この例を図16に示す。これは、前記「省略しているもの」の逆の例である。書き言葉では「損失の平均」となっているが、「損失の値の平均」と「値」をいれて分かりやすいように言い換えている。また、値も正確に「七十五デシベル」と言っている場合もある。一般に、話し言葉の方が書き言葉よりも省略が多いとされており、

この場合は逆の現象である。

【0052】7. コーパスの誤り検出に関わるもの
この例を図17に示す。もともと、話し言葉データ、書き言葉データ自体に誤りがあった場合、その部分が差分として得られることがある。1行目のデータは、「速報」を「速記」と誤ったものと思われる。この誤りは論文データをオンライン化する時に生じたものと思われる。また、話し言葉データの方にも誤りが見受けられる。最後の行のデータは、「死活」と「生活」どちらでも正しいような感じがする。

【0053】以上、話し言葉と書き言葉の差分を取り出した。図1に示す言葉変換手段12において、以上で得られたデータをデータベースとして利用する。

【0054】

【発明の効果】本発明によれば、話し言葉で入力した文章等を書き言葉の文章等で出力することができる話し言葉の書き言葉への変換装置が得られる。

【図面の簡単な説明】

【図1】本発明による話し言葉の書き言葉への変換装置の実施形態を示す図である。

【図2】図1の言葉変換手段で用いる話し言葉から書き言葉への言い換え(変換)のアルゴリズムである。

【図3】書き言葉データと話し言葉データの形態素への分割を示す説明図である。

【図4】書き言葉データと話し言葉データのdiffの結果を示す説明図である。

【図5】差分部分の抽出結果を示す説明図である。

【図6】差分の出現モード図である。

【図7】差分部分の拡張を示す図である。

【図8】話し言葉データに使用されているタグの説明図である。

【図9】差分部分の抽出数の結果を示す図である。

【図10】話し言葉データと書き言葉データの照合結果の上位50個を示す図である。

【図11】表記のゆれの例である。

【図12】表記・読みを与えるものの例である。

【図13】同義関係のもの例である。

【図14】口語調のもの例である。

【図15】省略しているもの例である。

【図16】補完しているもの例である。

【図17】誤り検出の例である。

【符号の説明】

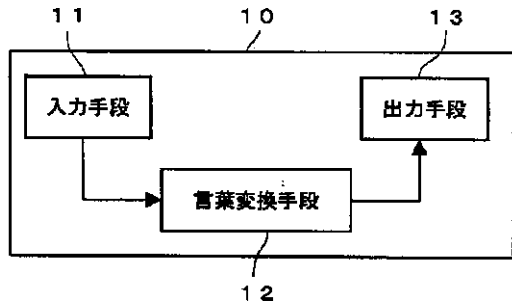
10 話し言葉の書き言葉への変換装置

11 入力手段

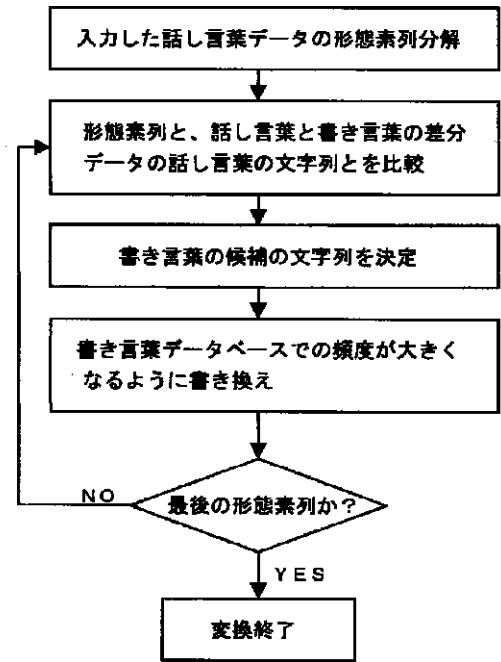
12 言葉変換手段

13 出力手段

【図 1】



【図 2】



【図 3】

話し言葉データ	書き言葉データ
え	本
今日は	論文
意味	で
ソート	は
に	単語
について	の
述べ	羅列
ます	を
え	意味
普通の	で
ソート	ソート
って	する
いう	と
の	いろいろな
は	とき
だいたい	に
え	便利である
50	と
音	いう
順	こと
と	に
か	ついて
EUC	記述
コード	する
順	。

【図 9】

	抽出数
確率 99.99%以上	1,011
確率 99.9%以上	3,245
確率 99%以上	7,846
確率 90%以上	14,777
総数	72,835
頻度 2 以上	421

【図 4】

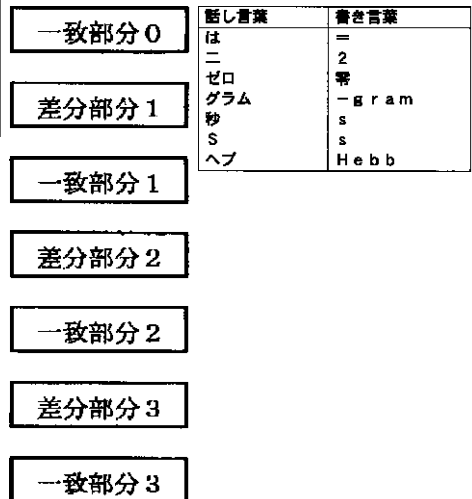
【図 5】

: ▲▲▲▲▲	: ▼▼▼▼▼
本	ソート
論文	: ▲▲▲▲▲
で	する
は	と
単語	いろいろな
の	とき
羅列	: ●●●
を	: ▼▼▼▼▼
: ●●●	に
え	: ▲▲▲▲▲
今日は	便利である
: ▼▼▼▼▼	と
意味	いう
: ▲▲▲▲▲	こと
で	に
: ●●●	: ●●●
(右欄につづく)	: ▼▼▼▼▼

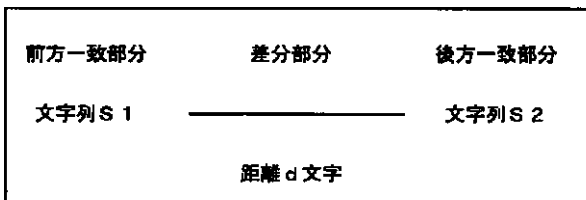
話し言葉データ	書き言葉データ
え今日は	本論文では単語の羅列を
	で
	するといろいろなとき
	便利であるということに

【図 7】

【図 12】



【図 6】



【図 15】

【図 16】

前方一致部分	話し言葉	書き言葉	後方一致部分
スムージング		処理	を
各		C (V) - {k}	兼片
スポーツニュース	の	における	会話部分を
平均時間が	11.3	11.25	分まで

前方一致部分	話し言葉	書き言葉	後方一致部分
に対する損失の	値の		平均として
会話に	関しましては全然		不便はない
音圧レベル	七十五デシベル	70 dB	で表示

【 図 8 】

タグ	説明	使用例
(F)	フィラー・感情表出系感動詞	(F あの)
(D)	言い直し	(D こ)これ
(W)	いい間違い	(W ミダリ;ヒダリ)
(?)	聞き取りなどに自身がない	(? タオングー)
(M)	音や言葉の引用	(? あのー、あんのー)
(O)	外国語や古語、方言など	(M わ)は(M は)と表記
(R)	個人名など	(O ザッツファイン)
(A)	基本型で漢字仮名以外の文字を使用する場合	(R 小林)さんが
(K)	何らかの原因で漢字表記できなくなった場合	(A イーユー;EU)
(S)	未登録の口語表現が出現した場合	(K たち(F んー)ばな;橋)
(笑)、(泣)、(嘆)、(あくび)	非言語音との共起	(S こりゃ)
(L)	ささやき声や独り言などの小さな声	(笑 なにそれ)
		(L アレコレナンダク)

【 図 11 】

話し言葉	書き言葉
データー	データ
クラスター	クラスタ
パラメーター	パラメータ
モダリティ	モダリティ
データーベース	データベース
掛かる	係る
越える	超える
全て	すべて
あ	ため
行なう	行う
言い替え	言い換え

【 図 14 】

【 図 10 】

頻度	前方一致部分の例	話し言葉	書き言葉	後方一致部分の例
182	IPALの形容詞			形容詞同の
72	を母国語	の	、	キーワードで効率良く検索
43	ばたばた			ばたばた
49	の拡張である	え		識別的特徴抽出法
56	L R表への複数	えー		接続制約の組み込みによる一般化LR法の
54	引き上げでNTT株は百万円になる		」	と予言したとか
39	に属するベクトル		の	和の距離に
28	本文間のハイパーリンク	を		自動生成
19	修飾要素		を	取り込む
21	ロストファイル群から導出した絞り込み題は		「	米映画
22	名詞の繰り返しの場合	えー		先頭の名詞のみを余剰語の挿入
20	Yは)	Xと事情が違う
21	経時的損失は全学習	データー	データ	に対する損失の
11	および用言の	が	ガ	格情報を付与したコーパスの作成
19	計算機科学の学会発表論文	<C>		25534表題対からなる
13	大量のコーパスを用いて	え	、	機械学習により
15	認識数	に		より切り出された
12	原言語と目標言語		と	の間の同時進行性が
10	叙述を示す言葉を補って	ー	ー	つの文を
10	雑音スペクトル	を	の	減算
14	その結果を人手	で		修正していくプロジェクトについて
20	また頻度も少なくない	という		ことから
10	番組	の	を	最初から最後まで
11	合計強度がN		- {	E
10	記事の	その		構造と特徴
16	ニュース番組自動字幕化の	為	ため	の音声認識システムを
10	近い	クラスター	クラスタ	が意味
6	ベクトルを異なる	K	k	個の集合にクラスタリングした
8	VQコードブックの	二	2	種類の話しモデル
7	適合率	は	=	認識結果が
8	その文中で	ん		話題となっている要素
9	十分良い精度で	は		推定できないという問題が
6	連結学習と	お		混合数を倍増する操作を繰り返す
11	ピッチの上昇	というもの		が有声音
7	基礎的	な	の	検出を
5	生起制限	が	/	厳しい形態素も
5	狭帯域			広帯域CELP
7	深さDの	おー		上位概念に抽象化
7	をコンテキストに持つ	ような		言語モデルに
14	手法	というの		を提案
4	かぎ括弧部分が	掛かる	係る	語の叙述を示す言葉ではないが
7	状態遷移をセットにして遺伝的探索を	行なう	行う	ことを提案し
4	E		}	以下のリンクを検出
4	時間的に逆向きの遷移	が	は	起こらないという
5	科学論文における要旨		ー	本文間のハイパーリンク
4	最も結果	の	が	良かった混合数
4	良い意味合い		」	悪い意味合い
6	概念が	あー		支配する概念数
4	が持つて	る	いる	常識を
6	性能	が		改善

話し言葉	書き言葉
という	
いたしました	した。
ですね	、
です	
られます	られる。
ってどうか	
とか	や
こう	
非常にこう	
います	いる。
分かりました	分かった。
ません	ない。
訳ですが	、
ってうちの	
れるんですが	れた。
であって	であり、
訳ですけれども	ことである。
ありますけれども	ある。
なくとも、	なくとも、
ないって	ないと
これ	

【 図 1 3 】

話し言葉	書き言葉
と とか 研究 異なりで それぞれ ターム 1 動詞 認識 違えば	および や 論文 ・異なり 各 1 番目のターム 述語 識別 違えば、

【 図 1 7 】

前方一致部分	話し言葉	書き言葉	後方一致部分
話し言葉データでの誤り			
かぎ括弧の種類	表層的な	表層的な	情報のみで
書き言葉データでの誤り			
ニュース 日本語が 日本語複数	速報 延べ 名詞構造	速記 述べ 名詞構造	記事 1 7 8 0 9 1 冊 解析法
どちらかのデータが誤り、もしくは不明			
マイクホン 社会の	および声帯 生活	及び生体 死活	アンブ にかかわる問題

フロントページの続き

(51)Int.Cl.⁷

識別記号

F I

テラコード⁸ (参考)

G 1 0 L 15/22

G 1 0 L 3/00

5 6 1 E

F ターム(参考) 5B009 QA17 QB14

5B091 AA11 AA15 BA19 CA02 CA21

CB12

5D015 HH23 KK03