(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3878998号
(P3878998)
(45)発行日　平成19年2月7日(2007.2.7)　(24)登録日　平成18年11月17日(2006.11.17)

(51)Int.Cl.　　F I
G06F 17/28 (2006.01)　　G06F 17/28 U
G06F 17/27 (2006.01)　　G06F 17/27 L

請求項の数 8　(全 14 頁)

(21)出願番号　特願2003-132528(P2003-132528)
(22)出願日　平成15年5月12日(2003.5.12)
(65)公開番号　特開2004-334730(P2004-334730A)
(43)公開日　平成16年11月25日(2004.11.25)
審査請求日　平成15年5月12日(2003.5.12)

特許法第３０条第１項適用　２００２年１１月１２日　社団法人情報処理学会発行の「情報処理学会研究報告情処研報２００２－ＮＬ－１５２－２０，Ｖｏｌ．２００２，Ｎｏ．１０４」に発表

特許法第３０条第１項適用　２００３年３月１８日　言語処理学会発行の「第９回年次大会発表論文集」に発表

(73)特許権者　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100130498
弁理士　佐野　禎哉
(72)発明者　木田　敦子
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　山本　英子
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　桝山　享子
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

最終頁に続く

(54)【発明の名称】呼応ペアデータベース生成支援装置、及び呼応ペアデータベース生成支援プログラム

(57)【特許請求の範囲】

【請求項１】

日本語の自然文からなる原文データの集合である原文データベースを検索することによって、同一文中において共起する２つの語のうち係り結びを形成する呼要素と応要素とを対にした呼応ペアの集合である呼応ペアデータベースを自動的に生成するためのコンピュータからなるものであって、

原文データベースに格納された原文データについてそれぞれ形態素解析を含む前処理を実行することによって、前記形態素解析の結果から所定の品詞に該当する語を削除し更にその削除後の残余の語を所定の語順で並べ替えて各語に品詞情報を付したデータである基礎データを生成するとともに、この基礎データの集合である基礎データベースを生成する基礎データベース生成手段と、

呼要素となり得る所定の対象語に基づき前記基礎データベースを検索して当該対象語と共起する語を含む基礎データについて、当該対象語及び該対象語と共起する語の２つの語に対する二値パターンをそれぞれ二値ｎ次元のベクトルとした場合に一方のベクトルが他方のベクトルにどれだけ類似しているかを表す補完類似度を演算する補完類似度演算手段と、

前記補完類似度演算手段による演算結果から所定以上のスコアを示した基礎データから対象語とその応要素とを対にした共起ペアを作成する共起ペア作成手段と、

前記共起ペア作成手段で作成した共起ペアのうち、当該共起ペアを含む基礎データに対応する原文データにおいて対象語と応要素とがこの順で記述されている原文データの数が

逆順で記述されている原文データの数よりも多い共起ペアを呼応ペア候補として選択する呼応ペア候補選択手段と、

前記呼応ペア候補に基づいて、呼応ペアの集合として呼応ペアデータベースを生成する呼応ペアデータベース生成手段とを具備していることを特徴とする呼応ペアデータベース生成支援装置。

【請求項２】

前記基礎データベース生成手段が前処理として、原文データを形態素解析したデータについて、用言の活用形を原形に変換するとともに固有名詞を削除し、さらに五十音順に並べ替える処理を行って基礎データを生成するものである請求項１記載の呼応ペアデータベース生成支援装置。

【請求項３】

前記呼応ペア候補選択手段で選択した呼応ペア候補を構成している対象語を含む基礎データ数に対する当該呼応ペア候補を含む基礎データ数の割合を、各呼応ペア候補について信頼度として演算し、その演算結果から得られた信頼度が所定の閾値以上のものを呼応ペアとして選定する信頼度判定手段を更に具備し、前記呼応ペアデータベース生成支援が、この信頼度判定手段で選択された呼応ペアの集合に基づいて呼応ペアデータベースを生成するものである請求項１又は２記載の呼応ペアデータベース生成支援装置。

【請求項４】

前記信頼度の閾値を、０．０４に設定している請求項３記載の呼応ペアデータベース生成支援装置。

【請求項５】

日本語の自然文からなる原文データの集合である原文データベースを検索することによって、同一文中において共起する２つの語のうち係り結びを形成する呼要素と応要素とを対にした呼応ペアの集合である呼応ペアデータベースを自動的に生成するためのコンピュータを、

原文データベースに格納された原文データについてそれぞれ形態素解析を含む前処理を実行することによって、前記形態素解析の結果から所定の品詞に該当する語を削除し更にその削除後の残余の語を所定の語順で並べ替えて各語に品詞情報を付したデータである基礎データを生成するとともに、この基礎データの集合である基礎データベースを生成する基礎データベース生成手段と、

呼要素となり得る所定の対象語に基づき前記基礎データベースを検索して当該対象語と共起する語を含む基礎データについて、当該対象語及び該対象語と共起する語の２つの語に対する二値パターンをそれぞれ二値ｎ次元のベクトルとした場合に一方のベクトルが他方のベクトルにどれだけ類似しているかを表す補完類似度を演算する補完類似度演算手段と、

前記補完類似度演算手段による演算結果から所定以上のスコアを示した基礎データから対象語とその応要素とを対にした共起ペアを作成する共起ペア作成手段と、

前記共起ペア作成手段で作成した共起ペアのうち、当該共起ペアを含む基礎データに対応する原文データにおいて対象語と応要素とがこの順で記述されている原文データの数が逆順で記述されている原文データの数よりも多い共起ペアを呼応ペア候補として選択する呼応ペア候補選択手段と、

前記呼応ペアの集合として呼応ペアデータベースを生成する呼応ペアデータベース生成手段とを具備する呼応ペアデータベース生成支援装置として機能させることを特徴とする呼応ペアデータベース生成支援プログラム。

【請求項６】

前記コンピュータを、前記基礎データベース生成手段における前処理として、原文データを形態素解析したデータについて、用言の活用形を原形に変換するとともに固有名詞を削除し、さらに五十音順に並べ替える処理を行って基礎データを生成するように機能させる請求項５記載の呼応ペアデータベース生成支援プログラム。

【請求項７】

前記コンピュータを、前記呼応ペア候補選択手段で選択した呼応ペア候補を構成している対象語を含む基礎データ数に対する当該呼応ペア候補を含む基礎データ数の割合を、各呼応ペア候補について信頼度として演算し、その演算結果から得られた信頼度が所定の閾値以上のものを呼応ペアとして選定する信頼度判定手段を更に具備する呼応ペアデータベース生成支援装置として機能させ、さらに前記呼応ペアデータベース生成支援において、この信頼度判定手段で選択された呼応ペアの集合に基づいて呼応ペアデータベースを生成するように機能させる請求項５又は６記載の呼応ペアデータベース生成支援プログラム。

【請求項８】

前記信頼度の閾値を、０．０４に設定している請求項７記載の呼応ペアデータベース生成支援プログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、日本語文において「係り」と「結び」を形成する「呼応ペア」となる一対の語に係るデータベースを生成するための装置及びプログラムに関するものである。

【０００２】

【従来の技術】

中世以前の日本語には、係助詞と文末の活用形とが形態的な呼応関係を持つ「係り結び」の用法が存在したが、「係り結び」が消滅した現代の日本語の文章の場合、述語が文末に置かれるため、文の終末まで進まないとその文章の内容が確定しない。そのため、長文で複雑な内容の文章では、その内容が肯定的なのか否定的なのか、或いは疑問を表しているのかが文末まで読まないことには明らかにならない。ここで、現代日本語の文構造の研究において、現代語ではある種の副詞などが古語の係助詞と似た役割を果たしており、後続要素を予告しているとの示唆がなされている（例えば、非特許文献１参照）。例えば「決して……ない」や「たぶん……だろう」や「おそらく……だろう」などといった組み合わせは、呼応関係を形成する先行要素（呼要素）及び後続要素（応要素）のペアとして内省や直感である程度予測がつくと考えられると指摘されている（例えば、非特許文献２参照）。

【０００３】

【非特許文献１】

大野　晋，「係り結びの研究」，第１版，岩波書店，１９９３年１月１２日，ｐ３５０－３５１

【非特許文献２】

益岡　隆志，「モダリティの文法」，第１版，くろしお出版，１９９１年５月２５日，ｐ２９－４６

【０００４】

【発明が解決しようとする課題】

ところが、このような呼応ペアについては未だ体系立てた研究がなされておらず、上述した文献や教科書等においても少数の呼応ペアが例示されるに留まっているのが現状である。すなわち、現代日本語における「係り結び」の研究では、内省や直感では予測し得ないような呼応ペアが不足しているために、ある語とそれと共に現れる（共起する）語とが本当に呼応関係にあるのか否かを明らかにするには基礎的データが不十分であるといわざるを得ない。また、斯かる研究成果を利用して、ある日本語文について中途まで進んだ状態で呼要素をキーにしてその文が肯定文なのか否定文なのかなどといった文の意味内容を機械的に推測したり、記述後の文章で用いられている呼応表現が正しいか否かを機械的に検証するというような応用を図ることも、現状では困難である。

【０００５】

そこで本発明は、以上のような問題に鑑みて、日本語文における「係り結び」に関する研究の基礎データとなり、また上述したような応用分野にも適用することができる「呼応ペア」のデータベースを適切に構築することができるようにすることを主たる目的としてい

る。
【０００６】
【課題を解決するための手段】
すなわち、本発明は、日本語の自然文からなる原文データの集合である原文データベースを検索することによって、同一文中において共起する２つの語のうち係り結びを形成する呼要素と応要素とを対にした呼応ペアの集合である呼応ペアデータベースを自動的に生成するためのコンピュータからなる呼応ペアデータベース生成支援装置、並びに当該コンピュータを呼応ペアデータベース生成支援装置として機能させるためのプログラムである。
【０００７】
図１に基本的な機能構成図を実線で示すように、この呼応ペアデータベース生成支援装置Ａは、原文データベースＤＢ１に格納された原文データについてそれぞれ形態素解析を含む前処理を実行することによって、前記形態素解析の結果から所定の品詞に該当する語を削除し更にその削除後の残余の語を所定の語順で並べ替えて各語に品詞情報を付したデータである基礎データを生成するとともに、この基礎データの集合である基礎データベースＤＢ２を生成する基礎データベース生成手段１と、呼要素となり得る所定の対象語に基づき前記基礎データベースＤＢ２を検索して当該対象語と共起する語を含む基礎データについて、当該対象語及び該対象語と共起する語の２つの語に対する二値パターンをそれぞれ二値ｎ次元のベクトルとした場合に一方のベクトルが他方のベクトルにどれだけ類似しているかを表す補完類似度を演算する補完類似度演算手段２と、前記補完類似度演算手段による演算結果から所定以上のスコアを示した基礎データから対象語とその応要素とを対にした共起ペアを作成する共起ペア作成手段３と、前記共起ペア作成手段で作成した共起ペアのうち、当該共起ペアを含む基礎データに対応する原文データにおいて対象語と応要素とがこの順で記述されている原文データの数が逆順で記述されている原文データの数よりも多い共起ペアを呼応ペア候補として選択する呼応ペア候補選択手段４と、前記呼応ペアの集合として呼応ペアデータベースを生成する呼応ペアデータベース生成手段６とを具備していることを特徴とするものである。
【０００８】
ここで、原文データベースＤＢ１には、現代日本語の自然文データを原文データとして多数格納してある。これら原文データを品詞ごとの単語に分解し、それら単語ごとに品詞名等を付与する前処理を行うのが、基礎データベース生成手段１の機能であり、この機能は、一般的な日本語の形態素解析プログラム及び形態素解析用辞書を利用することによって実現することができる。
【０００９】
補完類似度とは、本来文字認識システムにおいて、劣化印刷文字を高い精度で人敷くできるようにするための尺度として開発されたものである。すなわち、補完類似度を用いた文字認識方法である補完類似度法では、文字を二値画像特徴として扱い、補完類似度を用いて、そのパターンとテンプレートとする文字のパターンとの類似度を計算して文字が認識される。この手法は、汚れた文字においては人間による文字認識と同等の精度を持ち、かすれた文字においては人間による文字認識よりも高い精度を持つとされている。ここで、日本語自然文において２つの語句が出現するパターンとして捉えると、その出現パターンは二値パターンであるため、仮にこれを上述のような文字パターンと置き換えたとすれば、二つの語句の出現パターンが異なる部分はかすれや汚れと解釈することができるため（参考文献；山本英子、梅村恭司，「コーパス中の一対多関係を推定する問題における類似尺度」，「自然言語処理」，ｖｏｌ．９　Ｎｏ．２，２００２年，ｐ４５－７５）、本発明において補完類似度を日本語自然文における共起ペア乃至呼応ペアの出現パターンに適用したものである。
【００１０】
また、対象語とは、現代日本語文において「係り結び」を構成する２単語のうち「係り」に該当する呼要素を意味する。呼要素たる対象語となり得る単語には、例えば『基礎日本語文法　改訂版』（益岡隆志、田窪行則、くろしお出版、１９９２年）の分類によると、

「提題助詞」、「取り立て助詞」、「陳述の副詞」が該当する。具体的な対象語としては、「こそ」、「しか」、「さえ」、「は」、「も」、「ばかり」、「のみ」、「すら」、「なら」、「くらい（ぐらい）」、「だけ」、「なんて」、「けっして（決して）」、「おそらく（恐らく）」、「たぶん（多分）」、「ぜひ（是非）」、「まるで」、「もし」、「きっと」等の語を挙げることができるが、必ずしもこれらに限定されるわけではなく、これら以外の適宜の語を対象語に加えたり、これらの一部のみを対象語とすることも可能である。そして、「結び」に該当する「応要素」は、現代日本語自然文において「呼要素」が出現した場合にそれと同時に出現する単語である。本発明では、「呼要素」と「応要素」とが同一文中に同時に出現することを「共起する」と定義するとともに、この「共起」する「呼要素」と「応要素」との組み合わせを「共起ペア」と呼び、その「共起ペア」のうち「呼要素」と「応要素」とがこの順で出現することを「呼応する」と定義するとともに、この「呼応」する「呼要素」と「応要素」との組み合わせを特に「呼応ペア」と呼ぶものとする。

【００１１】

そして、上述のように求められた補完類似度の演算結果に基づいて、共起ペア作成手段３によって、所定値以上の補完類似度を得た基礎データから対象語である呼要素とそれに対応する応要素とを共起ペアとして得る。さらに、本発明で目的とするところは、「呼応」する語の組み合わせであるので、呼応ペア候補選択手段４では、先に得られた共起ペアのうち、呼応関係にあるもののみを抽出したり、共起はするが呼応はしないものを削除するなどして、対象語と応要素とが呼応しているもののみを選択する。すなわち、この選択された「共起ペア候補」から「呼応ペア」が得られることになる。このようにして選択された呼応ペア候補の集合から呼応ペアのデータベースを作成するのが呼応ペアデータベース生成手段６の機能であり、それによって呼応ペアデータベースＤＢ３が得られることになる。

【００１２】

以上のようにして本発明により得られる呼応ペアのデータベースには、大量の原文データから非常に多くの呼応ペアを機械的に得ることができるので、その呼応ペアの数は従来のように人間の直感や内省から得られていたものとは比較にならないといえる。すなわち、直感等では「呼応する」とは決して把握できなかった語のペアを新たに見出すことができる。したがって、現代日本語の構造解析の研究分野に多大な貢献をなすことができるのはもちろんのこと、そのような研究に基づいてなされる応用分野、すなわち日本語入力プログラムや日本語解析プログラム等を作成しているコンピュータ産業分野にも極めて有益なものとなる。

【００１３】

また、本発明による呼応ペアのデータ及びそれを格納した呼応ペアデータベースＤＢ３の生成をより効率的なものとするためには、基礎データベース生成手段１によって前処理を行う際に、原文データを形態素解析したデータについて、用言の活用形を原形に変換するとともに固有名詞を削除し、さらに五十音順に並べ替える処理を行って基礎データを生成するとよい。すなわち、対象語である呼要素にはなり得ない語を省くことで、基礎データベース生成手段１による前処理以後の処理の効率化を図ることができる。

【００１４】

さらに、本発明の呼応ペアデータベース生成支援装置Ａは、図１に破線で示すように、信頼度判定手段５をも有するように構成することもできる。この信頼度判定手段５は、呼応ペア候補選択手段４で選択した呼応ペア候補を構成している対象語を含む基礎データ数に対する当該呼応ペア候補を含む基礎データ数の割合を、各呼応ペア候補について信頼度として演算し、その演算結果から得られた信頼度が所定の閾値（例えば、０．０４）以上のものを呼応ペアとして選定するものである。そして、呼応ペアデータベース生成支援６では、信頼度判定手段５で選択された呼応ペアの集合に基づいて呼応ペアデータベースを生成するようにする。このようにすれば、呼応ペア選択手段４で得られた膨大な数の呼応ペアのなかから、呼応ペアとして日本語自然文中に出現する確率が低いものを省略し、真に

呼応関係を形成するものと信頼できる語のペアを絞り込むことで、呼応ペアデータベースＤＢ３の信頼性を向上することができる。

【００１５】

【発明の実施の形態】

以下、本発明の一実施形態を、図面を参照して説明する。

【００１６】

この実施形態に係る呼応ペアデータベース生成支援装置Ａは、図１に機能構成を示したように、日本語自然文から「係り結び」を形成する呼要素と応要素とのペアである呼応ペアデータを収集した呼応ペアデータベースＤＢ３を生成するためのものである。この呼応ペアデータベース生成支援装置Ａは、日本語自然文のデータである原文データを多数格納した原文データベースＤＢ１を内蔵し又は外部に接続して検索することができる状態にあるコンピュータにより構成されるものである。このコンピュータは、図２に概略的な機器構成図を示すように、バス線等で電気的に接続されたＣＰＵ１０１、メモリ１０２、ハードディスク等の記憶装置１０３、モニタ等の表示装置１０４、キーボードやマウス等の入力装置１０５、各種通信インターフェース１０６等を備えた通常のパーソナルコンピュータ等からなり、例えば外部に原文データベースＤＢ１、基礎データベースＤＢ２、呼応ペアデータベースＤＢ３等を通信線を介して接続してある。なお、これらデータベースＤＢ１～ＤＢ３に格納されるデータは、ハードディスク等の記憶装置に格納させることもできる。

【００１７】

そして、記憶装置１０３に格納した呼応ペアデータベース生成支援プログラムをＣＰＵ１０１が読み出してメモリ１０２に記憶させ、当該ＣＰＵ１０１が前記プログラムに従った処理を行い、メモリ１０２、ハードディスク等の記憶装置１０３、モニタ等の表示装置１０４、キーボードやマウス等の入力装置１０５、各種通信インターフェース１０６等の機器を駆動させることによって、このコンピュータは、呼応ペアデータベース生成支援装置Ａとして機能する。ここで、呼応ペアデータベース生成支援装置Ａの機能とは、図１に示したように、基礎データベース生成手段１、補完類似度演算手段２、共起ペア作成手段３、呼応ペア候補選択手段４、信頼度判定手段５、呼応ペアデータベース生成手段６を指す。

【００１８】

また、原文データベースＤＢ１に格納される原文データには、例えば新聞記事等から収集した日本語自然文のデータを利用することができ、本実施形態では「毎日新聞記事データ」と「日経新聞記事データ」の各１０年分を利用している。図３に、原文データの例を示す。各原文データは、記事ＩＤと記事本文とから構成されており、記事本文は、上述した新聞記事のテキストデータである。

【００１９】

次に、このような原文データベースＤＢ１を利用した呼応ペアデータベース生成支援装置Ａの動作手順を、図４以下を参照して説明する。なお、呼要素となり得る対象語は、本実施形態では「提題助詞」、「取り立て助詞」、「陳述の副詞」に分類される語がこの呼応ペアデータベース生成プログラムに予め選定されて記述されているものとする。具体的な対象語は、「こそ」、「しか」、「さえ」、「は」、「も」、「ばかり」、「のみ」、「すら」、「なら」、「くらい（ぐらい）」、「だけ」、「なんて」、「けっして（決して）」、「おそらく（恐らく）」、「たぶん（多分）」、「ぜひ（是非）」、「まるで」、「もし」、「きっと」である。ただし、このうちの一部を対象語としたり、他の語を対象語群に追加することも可能である。

【００２０】

呼応ペアデータベース生成支援装置Ａの動作は、基礎データベース生成手段１、補完類似度演算手段２、共起ペア作成手段３、呼応ペア候補選択手段４、信頼度判定手段５、呼応ペアデータベース生成手段６の各機能に対応して、概略的には図４に示すように、基礎データベース生成段階Ｓ１、補完類似度演算段階Ｓ２、共起ペア作成段階Ｓ３、呼応ペア候

補選択段階Ｓ４、信頼度判定段階Ｓ５、呼応ペアデータベース生成段階Ｓ６の６段階からなる。

【００２１】

まず、基礎データベース生成段階Ｓ１では、基礎データベース生成手段１の機能により、移行の処理のための前処理として、図５に示すように、原文データベースＤＢ１から原文データを読み込んでメモリに格納し（Ｓ１１）、各原文データについて形態素解析を実行する（Ｓ１２）。ここで、形態素解析プログラムには、例えば日本語形態素解析プログラムとして、「ＪＵＭＡＮ」（ｈｔｔｐ：／／ｗｗｗ．ｌａｂ２５．ｋｕｅｅ．ｋｙｏｔｏ－ｕ．ａｃ．ｊｐ／ｎｌ－ｒｅｓｏｕｒｃｅ／ｊｕｍａｎ．ｈｔｍｌ／）や「茶筅」（ｈｔｔｐ：／／ｃｈａｓｅｎ．ａｉｓｔ－ｎａｒａ．ａｃ．ｊｐ／）を用いることができる。また、この前処理における形態素解析ステップＳ１２に引き続き、当該形態素解析プログラムの品詞体系において「未定義」、「特殊」、「名詞」のうち「普通名詞」「固有名詞」「人名」「地名」「サ変動詞」に該当する語を削除して（Ｓ１３）、原文データごとの各語を五十音順に並べ替え（Ｓ１４）、これを基礎データとして収集することによって基礎データベースＤＢ２を生成する（Ｓ１５）。このようにして得られた基礎データベースＤＢ２に格納された基礎データの例を図６に示す。このように、各基礎データは、記事ＩＤごとに区分されており、各基礎データに含まれる語には品詞情報が付されている。

【００２２】

次に、補完類似度演算段階Ｓ２では、補完類似度演算手段２の機能により、図７に示すように、基礎データベースＤＢ２から各基礎データを読み込んでメモリに格納した後（Ｓ２１）、所定のパラメータの数え上げを行う（Ｓ２２）。ここで、パラメータには、「ａ：二つの単語が同時に現れる基礎データ数」、「ｂ：一方の単語が現れ、他方の単語が現れない基礎データ数」、「ｃ：ｂとは逆に、一方の単語が現れず、他方の単語が現れる基礎データ数」、「ｄ：二つの単語がどちらとも現れない基礎データ数」が用いられる。次いで、これらパラメータａ，ｂ，ｃ，ｄを利用した補完類似度の演算式により、補完類似度（Ｓｃ）を算出する（Ｓ２３）。すなわち、各補完類似度（Ｓｃ）は、二つの語句に対する二値パターンをそれぞれ二値ｎ次元のベクトルとした場合、一方のベクトルが他方のベクトルにどれだけ類似しているかによって表され、具体的には、パラメータａとパラメータｃの和とパラメータｂとパラメータｄの和との平方根を分母として、パラメータａとパラメータｄとの積とパラメータｂとパラメータｃとの積の差を分子とする演算式（図中、式Ｘ）によって求められる。そして、この演算結果から得られた補完類似度の高いものから降順に、二つの単語のペアの並べ替えを行う（Ｓ２４）。ここで図８に、共に出現した（共起した）二つの単語の補完類似度の演算結果の一部を一覧にして示す。同図における中欄と右欄に記載された語が、同一文中で共起した可能性のあるペア（以下、「共起候補ペア」）であり、左欄は各共起候補ペアについての補完類似度を示す。

【００２３】

次に、共起ペア作成段階Ｓ３では、共起ペア作成手段３の機能により、図９に示すように、まず補完類似度が所定の閾値以上の共起候補ペアを抽出し（Ｓ３１）、それらを共起ペアとして対象語ごとにまとめる処理を行う（Ｓ３２）。なお、補完類似度の閾値は、本実施形態では例えば「０．０００１」としている。ここで図１０に、対象語「きっと」についての共起ペアを一覧にして示す。すなわち、「きっと」を呼要素とする応要素となり得る語が「応要素候補」として挙げられることになる。

【００２４】

次に、呼応ペア候補選択段階Ｓ４では、呼応ペア候補選択手段４の機能により、共起ペアのうち一対の語が呼応関係にないものを除き、呼応関係にあるもののみを呼応ペア候補として選択する。すなわち、図１１に示すように、共起ペアを含む原文データを原文データベースＤＢ１から取得し（Ｓ４１）、当該共起ペアを含む原文データ数を出現頻度として実数で求める（Ｓ４２）。そして、各共起ペアについて対象語（呼要素）と応要素とがこの順（「呼」「応」の順）で出現した原文データ数（以下、「正順データ数」）を計数するとともに（Ｓ４３）、その逆順すなわち「応」「呼」の順で出現した原文データ数（以

下、「逆順データ数」）を計数し（Ｓ４４）、ステップＳ４３で得た正順データ数からステップＳ４４で得た逆順データ数を差し引いて（Ｓ４５）、その値が正数であったもの（Ｓ４５；Ｙｅｓ）を呼応ペア候補として抽出する（Ｓ４６）。なお、正順データ数と逆順データ数の差が０以下であったもの（Ｓ４５；Ｎｏ）は削除される（Ｓ４７）。ここで図１２に、対象語「きっと」についての呼応ペア候補を一覧にして示す。同図最右欄の「判定」において、「○」が付されているものは、呼応ペア候補として抽出されるものであり、「×」が付されているものは、共起はしたものの呼応したとは認められず削除対象となるものである。

【００２５】

次に、信頼度判定段階Ｓ５では、信頼度判定手段５の機能により、呼応候補ペアから真に呼応関係にあると認められるものの絞り込みを行う。呼応していると一応は認められた「呼応ペア候補」の数は極めて膨大であり、その中には真に呼応関係にあるとは認めがたいものが多数含まれているからである。すなわち、図１３に示すように、各呼応ペア候補について、呼応ペア候補の構成要素である対象語を含む基礎データを基礎データベースＤＢ２から抽出し（Ｓ５１）、当該基礎データ数を計数するとともに（Ｓ５２）、これら基礎データのうち当該呼応ペア候補を含む基礎データ数を計数し（Ｓ５３）、後者の基礎データ数の前者の基礎データ数に対する割合を信頼度として演算する（Ｓ５４）。そして、信頼度が所定の閾値以上のもの（Ｓ５４；Ｙｅｓ）を呼応ペアとして選定・抽出する（Ｓ５５）。信頼度が０．０４を下回ったもの（Ｓ５４；Ｎｏ）は削除される（Ｓ５６）。なお、信頼度の閾値は「０．０４」としている。信頼度の基準値をこのように設定したのは次の理由による。すなわち、基準値を０．０５以上とすると、例えば「まるで……みたい（名詞-非自立-形容動詞語幹）」、「きっと……ね（助詞－終助詞）」、「おそらく……ようだ（助動詞／ナ形容詞）」等の呼応関係として着目する可能性を残すべきペアが撥ねられてしまい、その一方、基準値を０．０４より下げると、例えば「おそらく……初めて（副詞）」、「は……者（名詞-接尾-一般）」、「は……的（名詞-接尾-形容動詞語幹）」等の呼応関係にあるとはいえず除外すべきものを拾ってしまうからである。ここで、図１４に、対象語「きっと」についての信頼度判定結果を一覧にして示す。同図最右欄の「判定」において、「○」が付されているものは、信頼度が０．０４以上であり呼応関係にあると認められる呼応候補ペアであり、「×」が付されているものは、信頼度が０．０４よりも小さく真に呼応しているとは認められず削除対象となるものである。

【００２６】

最後に、呼応ペアデータベース生成段階Ｓ６では、呼応ペアデータベース生成手段６の機能によって、図１５に示すように、前段階で選定した呼応ペアを対象語ごとに収集し（Ｓ６１）、これらを呼応ペアデータベースＤＢ３に出力して格納する（Ｓ６２）。図１６に、対象語「きっと」についての呼応ペアデータベースＤＢ３の内容の一部を示す。同図に示した一例からも明らかなように、呼要素「きっと」についてだけでも、経験的に呼応していると考えられていたよりも極めて多数の応要素が得られる。すなわち、本実施形態により得られる呼応ペアデータベースＤＢ３を利用することで、従来からは指摘されていなかった呼応関係や直感では気付きにくい呼応関係の発見も可能になるため、現代日本語の構文研究の促進が図られるだけでなく、構文解析ソフトウェア等の開発にも資するものであるといえる。

【００２７】

なお、本発明は上述した実施形態に限られるものではなく、各手段の具体的機能等も、本発明の趣旨を逸脱しない範囲で種々変形が可能である。

【００２８】

【発明の効果】

以上に詳述したように、本発明に係る呼応ペアデータベース生成支援装置又はそのためのプログラムによれば、従来は人間の感覚や経験で「呼応関係」にあると考えられていた現代日本語における「係り結び」を構成する対語である呼応ペアよりも、遙かに多くの呼応ペアを得ることができる。そのため、現代日本語の構文に関する研究を一挙に促進するこ

とができるだけでなく、効率のよい日本語構文解析プログラムの作成の基礎データとしても大いに役立つものである。

【図面の簡単な説明】

【図１】本発明及びその一実施形態に係る呼応ペアデータベース生成支援装置の機能構成を概略的に示す図。

【図２】同実施形態に係る呼応ペアデータベース生成支援装置を構成するコンピュータの概略的機器構成図。

【図３】同実施形態において利用される原文データの一部を示す図。

【図４】同呼応ペアデータベース生成支援装置の動作の概観を示すフローチャート。

【図５】基礎データベース生成段階を詳細に示すフローチャート。

【図６】同基礎データの一部を示す図。

【図７】補完類似度演算段階段階を詳細に示すフローチャート。

【図８】補完類似度の演算結果の一部を示す図。

【図９】共起ペア作成段階を詳細に示すフローチャート。

【図１０】対象語「きっと」についての共起ペアの一部を示す図。

【図１１】呼応ペア候補選択段階を詳細に示すフローチャート。

【図１２】対象語「きっと」についての呼応ペア候補の一部を示す図。

【図１３】信頼度判定段階を詳細に示すフローチャート。

【図１４】対象語「きっと」についての信頼度判定結果の一部を示す図。

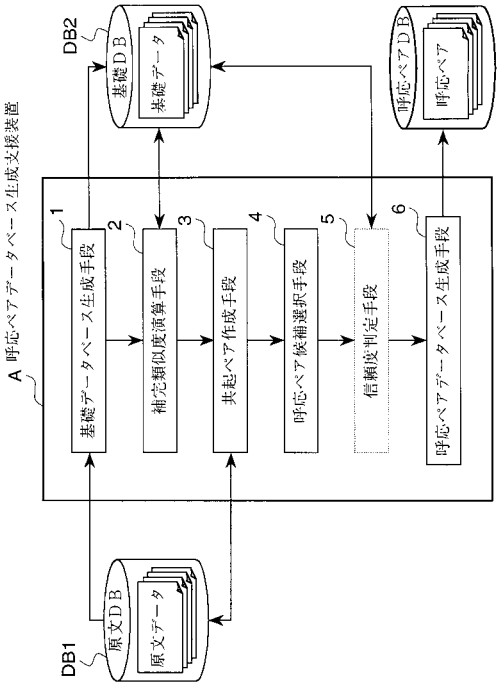
【図１５】呼応ペアデータベース生成段階を詳細に示すフローチャート。

【図１６】対象語「きっと」についての呼応ペアの一部を示す図。

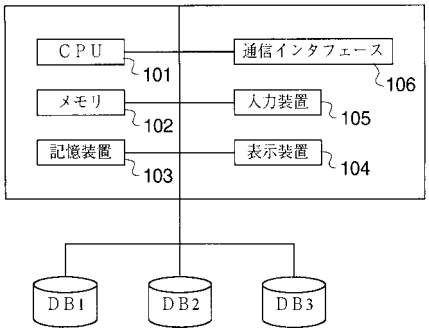
【符号の説明】

Ａ…呼応ペアデータベース生成支援装置
ＤＢ１…原文データベース
ＤＢ２…基礎データベース
ＤＢ３…呼応ペアデータベース
１…基礎データベース生成手段
２…補完類似度演算手段
３…共起ペア作成手段
４…呼応ペア候補選択手段
５…信頼度判定手段
６…呼応ペアデータベース生成手段

【図１】



【図２】

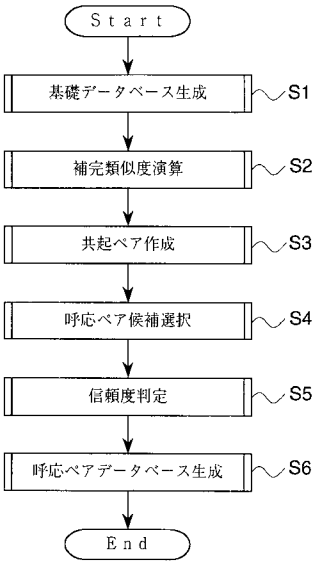


【図３】

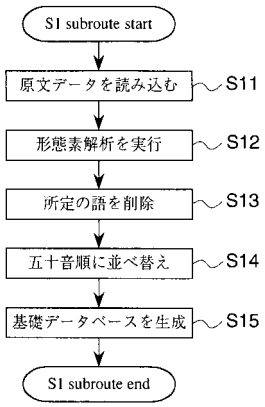
＜原文データ＞

記事ＩＤ	記事本文
000218244-8	国に相談なんてしてたら、きっとつぶされていたはず。
19921219 JITYMLL 1400010117866-14	見学しているうちにきっと宝くじにかける夢が広がるはずだ。
950212154-33	砂ではなく乾いた土なので、水さえあればたちまち緑の野と化すであろう。
920304001-23	これは日、米、カナダ三国にとって決して好ましいことではない。

【図４】



【図５】

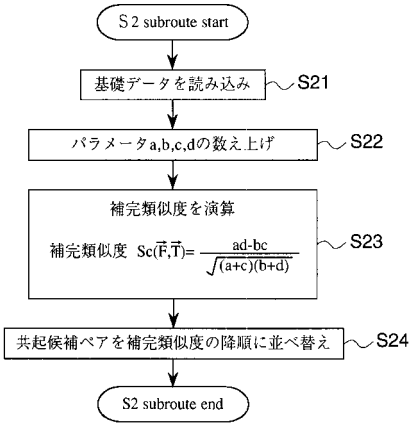


【図６】

＜基礎データ＞

記事ＩＤ	記事本文
000218244-8	いる〔動詞－非自立／一段〕きっと〔副詞－一般〕する〔動詞－自立／サ変・スル〕た〔助動詞／特殊・タ〕た〔助動詞／特殊・タ〕つぶす〔動詞－自立／五段・サ行〕て〔助詞－接続助詞〕てる〔動詞－非自立／一段〕なんて〔副詞－一般〕に〔助詞－格助詞－一般〕はず〔名詞－非自立－一般〕れる〔動詞－接尾／一段〕相談〔名詞－サ変接続〕
19921219 J1TYMLL 1400010117866-14	いる〔動詞－非自立／一段〕うち〔名詞－非自立－副詞可能〕かける〔動詞－自立／一段〕が〔助詞－格助詞－一般〕きっと〔副詞－一般〕する〔動詞－自立／サ変・スル〕だ〔助動詞／特殊・ダ〕て〔助詞－接続助詞〕に〔助詞－格助詞－一般〕に〔助詞－格助詞－一般〕はず〔名詞－非自立－一般〕見学〔名詞－サ変接続〕広がる〔動詞－自立／五段・ラ行〕
950212154-33	ある〔助動詞／五段・ラ行アル〕ある〔動詞－自立／五段・ラ行〕う〔助動詞／不変化型〕さえ〔助詞－係助詞〕た〔助動詞／特殊・タ〕たちまち〔副詞－助詞類接続〕だ〔助動詞／特殊・ダ〕だ〔助動詞／特殊・ダ〕だ〔助動詞／特殊・ダ〕と〔助詞－格助詞－一般〕ない〔助動詞／特殊・ナイ〕の〔助詞－連体化〕ので〔助詞－接続助詞〕は〔助詞－係助詞〕ば〔助詞－接続助詞〕化す〔動詞－自立／五段・サ行〕乾く〔動詞－自立／五段・カ行イ音便〕
920304001-23	こと〔名詞－非自立－一般〕だ〔助動詞／特殊・ダ〕ない〔助動詞／特殊・ナイ〕にとって〔助詞－格助詞－連語〕は〔助詞－係助詞〕は〔助詞－係助詞〕決して〔副詞－一般〕好ましい〔形容詞－自立／形容詞・イ段〕

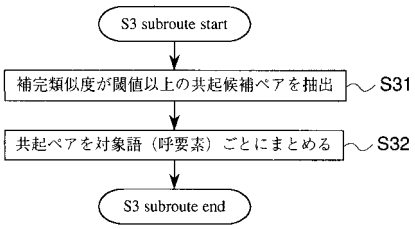
【図７】



【図８】

補完類似度	共起候補ペア	
0.325851	て〔助詞－接続助詞〕	いる〔動詞－非自立／一段〕
0.155690	と〔助詞－格助詞－引用〕	いう〔動詞－自立／五段・ワ行促音便〕
0.142344	する〔動詞－自立／サ変・スル〕	を〔助詞－格助詞－一般〕
0.128615	だ〔助動詞／特殊・ダ〕	ある〔助動詞／五段・ラ行アル〕
0.107968	する〔動詞－自立／サ変・スル〕	た〔助動詞／特殊・タ〕
0.105680	て〔助詞－接続助詞〕	くる〔動詞－非自立／カ変・クル〕
0.101710	と〔助詞－接続助詞〕	よる〔動詞－自立／五段・ラ行〕
0.098688	する〔動詞－自立／サ変・スル〕	て〔助詞－接続助詞〕
：	：	：
0.019765	ない〔助動詞／特殊・ナイ〕	決して〔副詞－一般〕
：	：	：
0.008747	ば〔助詞－接続助詞〕	さえ〔助詞－係助詞〕
：	：	：
0.001602	ない〔形容詞－自立／形容詞〕	決して〔副詞－一般〕
：	：	：
0.001169	はず〔名詞－非自立－一般〕	きっと〔副詞－一般〕
：	：	：
0.000082	普及〔名詞－サ変接続〕	畜産〔名詞－サ変接続〕
0.000082	普及〔名詞－サ変接続〕	装用〔名詞－サ変接続〕
0.000082	扶養〔名詞－サ変接続〕	別居〔名詞－サ変接続〕
0.000082	怖い〔形容詞－自立／形容詞〕	笑える〔動詞－自立／一段〕
0.000082	府〔名詞－接尾－地域〕	留守〔名詞－サ変接続〕

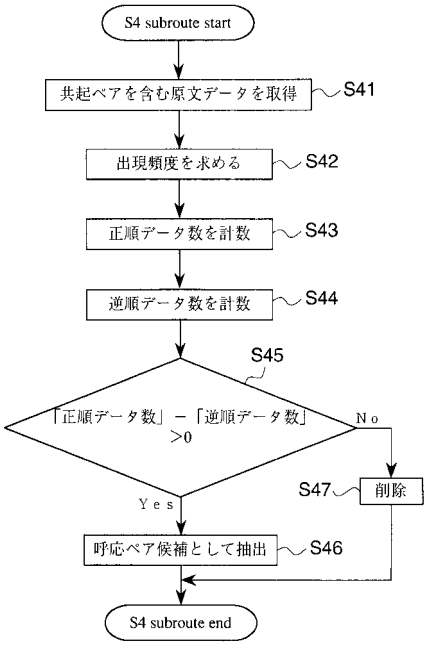
【図９】



【図１０】

呼要素候補（対象語）	応要素候補
きっと（副詞－一般）	う（助動詞／不変化型）
きっと（副詞－一般）	だ（助動詞／特殊・ダ）
きっと（副詞－一般）	と（助詞－格助詞－引用）
きっと（副詞－一般）	て（助詞－接続助詞）
きっと（副詞－一般）	思う（動詞－自立／五段）
きっと（副詞－一般）	です（助動詞／特殊・デス）
きっと（副詞－一般）	ない（助動詞／特殊・ナイ）
きっと（副詞－一般）	ます（助動詞／特殊・マス）
きっと（副詞－一般）	の（名詞－非自立－一般）
きっと（副詞－一般）	も（助詞－係助詞）
きっと（副詞－一般）	はず（名詞－非自立－一般）
きっと（副詞－一般）	くれる（動詞－非自立／一段）
きっと（副詞－一般）	いる（動詞－非自立／一段）
きっと（副詞－一般）	ば（助詞－接続助詞）
きっと（副詞－一般）	よ（助詞－終助詞）
きっと（副詞－一般）	なる（動詞－自立／五段・ラ行）
きっと（副詞－一般）	から（助詞－接続助詞）
きっと（副詞－一般）	ん（名詞－非自立－一般）
きっと（副詞－一般）	こと（名詞－非自立－一般）
きっと（副詞－一般）	たち（名詞－接尾－一般）
きっと（副詞－一般）	違い（名詞－ナイ形容詞語幹）
きっと（副詞－一般）	か（助詞－副助詞／並立助詞／終助詞）
きっと（副詞－一般）	に（助詞－格助詞－一般）
きっと（副詞－一般）	さん（名詞－接尾－人名）
きっと（副詞－一般）	ね（助詞－終助詞）
きっと（副詞－一般）	ある（動詞－自立／五段・ラ行）
きっと（副詞－一般）	で（助詞－接続助詞）
きっと（副詞－一般）	いい（形容詞－自立／不変化型）
きっと（副詞－一般）	くる（動詞－非自立／カ変・クル）
きっと（副詞－一般）	言う（動詞－自立／五段）
きっと（副詞－一般）	この（連体詞）
きっと（副詞－一般）	よう（名詞－非自立－助動詞語幹）
：	：

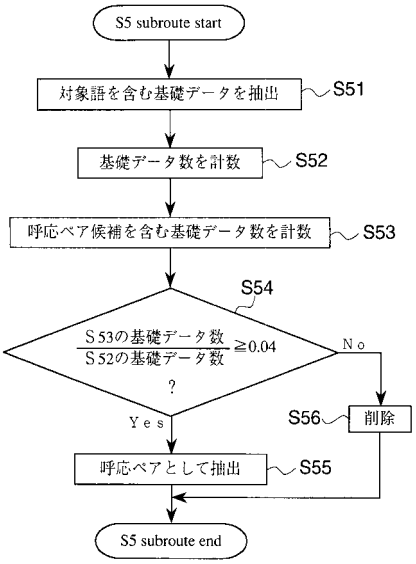
【図１１】



【図１２】

呼要素候補（対象語）	応要素候補	頻度	「呼」「応」の順	「応」「呼」の順	判定
きっと（副詞－一般）	う（助動詞／不変化型）	4276	3882	394	○
きっと（副詞－一般）	だ（助動詞／特殊・ダ）	10340	6771	3569	○
きっと（副詞－一般）	と（助詞－格助詞－引用）	6986	5785	1201	○
きっと（副詞－一般）	て（助詞－接続助詞）	11215	7035	4180	○
きっと（副詞－一般）	思う（動詞－自立／五段）	2384	2115	269	○
きっと（副詞－一般）	です（助動詞／特殊・デス）	2449	1943	506	○
きっと（副詞－一般）	ない（助動詞／特殊・ナイ）	3723	2397	1326	○
きっと（副詞－一般）	ます（助動詞／特殊・マス）	2428	1959	469	○
きっと（副詞－一般）	の（名詞－非自立－一般）	2981	2122	859	○
きっと（副詞－一般）	も（助詞－係助詞）	4905	1776	3129	×
きっと（副詞－一般）	はず（名詞－非自立－一般）	1293	1244	49	○
きっと（副詞－一般）	くれる（動詞－非自立／一段）	1293	1098	195	○
きっと（副詞－一般）	いる（動詞－非自立／一段）	4631	3196	1435	○
きっと（副詞－一般）	ば（助詞－接続助詞）	1534	171	1363	×
きっと（副詞－一般）	よ（助詞－終助詞）	1232	942	290	○
きっと（副詞－一般）	なる（動詞－自立／五段・ラ行）	2514	1716	798	○
きっと（副詞－一般）	から（助詞－接続助詞）	1107	604	503	○
きっと（副詞－一般）	ん（名詞－非自立－一般）	1142	799	343	○
きっと（副詞－一般）	こと（名詞－非自立－一般）	2371	1405	966	○
きっと（副詞－一般）	たち（名詞－接尾－一般）	1046	395	651	×
きっと（副詞－一般）	違い（名詞－ナイ形容詞語幹）	694	676	18	○
きっと（副詞－一般）	か（助詞－副助詞／並立助詞／終助詞）	1532	905	627	○
きっと（副詞－一般）	に（助詞－格助詞－一般）	10478	5451	5027	○
きっと（副詞－一般）	さん（名詞－接尾－人名）	1369	420	949	×
きっと（副詞－一般）	ね（助詞－終助詞）	747	550	197	○
きっと（副詞－一般）	ある（動詞－自立／五段・ラ行）	1673	1017	656	○
きっと（副詞－一般）	で（助詞－接続助詞）	815	558	257	○
きっと（副詞－一般）	いい（形容詞－自立／不変化型）	534	382	152	○
きっと（副詞－一般）	くる（動詞－非自立／カ変・クル）	827	564	263	○
きっと（副詞－一般）	言う（動詞－自立／五段）	609	401	208	○
きっと（副詞－一般）	この（連体詞）	958	337	621	×
きっと（副詞－一般）	よう（名詞－非自立－助動詞語幹）	944	540	404	○
：	：	：	：	：	：

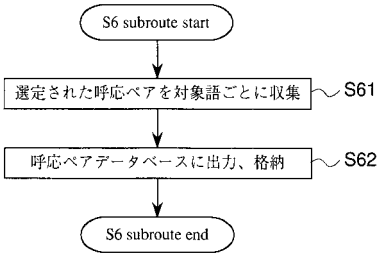
【図１３】



【図１４】

信頼度	呼要素候補（対象語）	応要素候補	判定
0.297814	きっと（副詞－一般）	う（助動詞／不変化型）	○
0.519448	きっと（副詞－一般）	だ（助動詞／特殊・ダ）	○
0.443805	きっと（副詞－一般）	と（助詞－格助詞－引用）	○
0.539701	きっと（副詞－一般）	て（助詞－接続助詞）	○
0.162255	きっと（副詞－一般）	思う（動詞－自立／五段）	○
0.14906	きっと（副詞－一般）	です（助動詞／特殊・デス）	○
0.18389	きっと（副詞－一般）	ない（助動詞／特殊・ナイ）	○
0.150288	きっと（副詞－一般）	ます（助動詞／特殊・マス）	○
0.162792	きっと（副詞－一般）	の（名詞－非自立－一般）	○
0.095435	きっと（副詞－一般）	はず（名詞－非自立－一般）	○
0.084235	きっと（副詞－一般）	くれる（動詞－非自立／一段・クレル）	○
0.245186	きっと（副詞－一般）	いる（動詞－非自立／一段）	○
0.072267	きっと（副詞－一般）	よ（助詞－終助詞）	○
0.131646	きっと（副詞－一般）	なる（動詞－自立／五段・ラ行）	○
0.046337	きっと（副詞－一般）	から（助詞－接続助詞）	○
0.061297	きっと（副詞－一般）	ん（名詞－非自立－一般）	○
0.107787	きっと（副詞－一般）	こと（名詞－非自立－一般）	○
0.05186	きっと（副詞－一般）	違い（名詞－ナイ形容詞語幹）	○
0.069428	きっと（副詞－一般）	か（助詞－副助詞／並立助詞／終助詞）	○
0.418182	きっと（副詞－一般）	に（助詞－格助詞－一般）	○
0.042194	きっと（副詞－一般）	ね（助詞－終助詞）	○
0.078021	きっと（副詞－一般）	ある（動詞－自立／五段・ラ行）	○
0.042808	きっと（副詞－一般）	で（助詞－接続助詞）	○
0.029306	きっと（副詞－一般）	いい（形容詞－自立／不変化型）	×
0.043268	きっと（副詞－一般）	くる（動詞－非自立／カ変・クル）	○
0.030763	きっと（副詞－一般）	言う（動詞－自立／五段）	×
0.041427	きっと（副詞－一般）	よう（名詞－非自立－助動詞語幹）	○
：	：	：	：

【図１５】



【図１６】

呼要素候補（対象語）	応要素候補
きっと（副詞－一般）	う（助動詞／不変化型）
きっと（副詞－一般）	だ（助動詞／特殊・ダ）
きっと（副詞－一般）	と（助詞－格助詞－引用）
きっと（副詞－一般）	て（助詞－接続助詞）
きっと（副詞－一般）	思う（動詞－自立／五段・ワ行促音便）
きっと（副詞－一般）	です（助動詞／特殊・デス）
きっと（副詞－一般）	ない（助動詞／特殊・ナイ）
きっと（副詞－一般）	ます（助動詞／特殊・マス）
きっと（副詞－一般）	の（名詞－非自立－一般）
きっと（副詞－一般）	はず（名詞－非自立－　一般）
きっと（副詞－一般）	くれる（動詞－非自立／一段・クレル）
きっと（副詞－一般）	いる（動詞－非自立／一段）
きっと（副詞－一般）	よ（助詞－終助詞）
きっと（副詞－一般）	なる（動詞－自立／五段・ラ行）
きっと（副詞－一般）	から（助詞－接続助詞）
きっと（副詞－一般）	ん（名詞－非自立－一般）
きっと（副詞－一般）	こと（名詞－非自立－一般）
きっと（副詞－一般）	違い（名詞－ナイ形容詞語幹）
きっと（副詞－一般）	か（助詞－副助詞／並立助詞／終助詞）
きっと（副詞－一般）	に（助詞－格助詞－一般）
きっと（副詞－一般）	ね（助詞－終助詞）
きっと（副詞－一般）	ある（動詞－自立／五段・ラ行）
きっと（副詞－一般）	で（助詞－接続助詞）
きっと（副詞－一般）	くる（動詞－非自立／カ変・クル）
きっと（副詞－一般）	よう（名詞－非自立－助動詞語幹）
きっと（副詞－一般）	た（助動詞／特殊・タ）
きっと（副詞－一般）	に（助詞－副詞化）
きっと（副詞－一般）	を（助詞－格助詞－一般）
きっと（副詞－一般）	と（助詞－格助詞－一般）

フロントページの続き

(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

審査官　和田　財太

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 17/27-17/28