2006－CE－87 (4)
2006／12／9

日本語教材作成のための用例抽出システムの開発
－検索式の改良の仕組みと評価－

于壮飛(ユソウヒ)† 佐野洋‡

東京外国語大学 †大学院地域文化研究科 ‡外国語学部

概要

我々は，日本語教育教材作成のための用例抽出システムの開発を進めている。そして，教育用日本語用例サイトの提供を目指している。本稿は，日本語教育教材作成のための用例抽出システムの概要と，本システムの検索部の特徴について述べる。

本システムは，タグ付された日本語用例(日本語コーパス)から，文型パターンに照合する用例を動的に，且つ大量に提供することを目的にしている。効率的な用例抽出や教材データの共有化など教材作成コストの軽減を実現する。本稿で詳細に説明する用例検索部の特徴は，言語的な抽象化の単位を用例検索に使う点にある。言語的な抽象化の単位の特徴説明と，システム上で改善した点について述べ，用例検索式の比較に基づいた用例検索部の評価を示す。

Development of Sentence Retrieval System for Japanese Teaching Material Production
－Mechanism and Evaluation of Improvement of Search Expression－

YU Zhuangfei† SANO Hiroshi‡

Tokyo University of Foreign Studies †Graduate School of Area and Culture Studies ‡Faculty of Foreign Studies

Abstract

The authors are developing the system for extracting example sentences to support Japanese teaching material production, in order to provide them as a website for educational use. This study describes the summary of the extracting system and the features of the retrieval module in it.

1. はじめに

CNN News Update や BCC Radio NewsPod などのポッドキャストで英語学習に使えるサイトが増え，語学教育分野でも教材の提供メディアの多様化は著しい。時間や場所に制限されない新しい学習形態である。同時に学習ニーズも多様化が進み，会議用英語やプレゼンテーション英語，

電子メールの英語など各種の学習教材が作成されている。経済市場のボーダレス化に伴う英語学習ニーズの急拡大だけでなく，英語以外の様々な言語に接する機会が増えた結果，英語以外の日本語や中国語の学習ニーズも拡大し，多様化している。

ところで，教材メディアの開発が技術指向であるのに対して，教材コンテンツの開発は，基本的に属人的な労働行為であって，作成コストが高いことがしばしば指摘される。一般に教材作成は，教授者の専門分野の知識，教授経験や教授手法とその表現の力量に依存する部分が多いからである。

この問題に対し，筆者等は効果的で効率的な教材コンテンツの作成枠組みの研究を進めている。教材用例の自動抽出の試みもその一つで，2005 年度，筆者等の研究室と(株)小学館コーパスネットワークの共同研究により英語文型教授のための教育用英語用例サイトを開発した。本用例サイトには，BNC(British National Corpus)から，コンピュータを使って自動抽出した 1320 項目の英語文型用例(およそ 80 万例文)がアップされている(2005 年 9 月に公開)。

我々は，英語だけでなく，日本語教育教材作成のための用例抽出システムの開発も進めており，上述とパラレルに教育用日本語用例サイトの提供を目指している。本稿は，日本語教育教材作成のための用例抽出システムの概要と，本システムの検索部の特徴について述べる。

本システムは，タグ付された日本語用例(日本語コーパス)から，文型パターンに照合する用例を動的に，且つ大量に提供することを目的にしている。効率的な用例抽出や教材データの共有化など教材作成コストの軽減を実現する。本稿で詳細に説明する用例検索部の特徴は，言語的な抽象化の単位を用例検索に使う点にある。利用指向の用例検索で，日本語教育に携わる者に使いやすいシステムを目指している。言語的な抽象化の単位の特徴説明と，システム上で改善した点について述べ，用例検索式の比較に基づいた用例検索部の評価を示す。

以下，2 章は，手短に先行研究に言及し，3 章では，日本語教育教材作成のための用例抽出システムの概要を示す。4 章は，本システムの検索部の特徴を説明する。5 章は，用例検索部の評価を示す。

2. 先行研究

2.1. 文型パターンを利用する例文検索システム

文型パターンを利用する例文検索システム[1](山本 2006)は，日本語教育教材(『初級日本語文法解説』[1])で扱われている文型を利用し，日本語文をタグ付けした結果から，文型パターンに照合する用例を抽出することができる。文型パターンは『文法解析』で示されている 334 文型を基に，時制や否定などの下位分類を設け，合計 1698 文型のパターンに拡張している。タグ付けは形態素レベルで，形態素解析システムChasen(茶筌)[2]を使用している。

このシステムでは，Chasen で解析した形態素解析結果をそのままタグに利用しているため，文型照合のためのパターン(検索式)記述が煩雑で書きにくい。つまり，Chasen の日本語分析枠組みを直接利用するので，日本語教育に携わる者にとって分かり難い(日本語教育で利用される日本語分析枠組みと，国語教育で利用される日本語分析枠組みが異なることに起因する)。日本語教育分野における教材作成の支援の観点から見ると，検索式の記述方法に改善が必要である。

---

[1] 東京外国語大学・留学生日本語教育センターで作成された日本語教材の 1 つで，6 分冊の中の文法説明を担う。本センターでの日本語教育に使われており，過去 30 年の教育実績が反映されている。

[2] Chasenについては 4.1 節を参照されたい。

2.2. 日本語形態素解析ツール WinMorph1.2

日本語形態素解析ツールWinMorph1.2[2]は，日本語教育のための日本語分析枠組みを直接利用する形態素解析を有する。用言の活用体系など形態素の分割方法がChasenと異なる。形態素解析エンジンにはBreakfast[3]を使用し，形態素分類データ，形態素辞書データ，形態素連接規則データを，利用者が利用目的に応じてカスタマイズできるという特徴を持つ。

タグ付けツールとして，WinMorph1.2 を利用することで，日本語教育に携わる者にとっても，文型照合のためのパターン(検索式)記述が容易になる可能性がある。しかし，WinMorph1.2 が予め用意する辞書データは規模が小さく，試験的に利用するにはよいが，実用利用上では，解析精度の低さから採用できない。形態素分類データや辞書データ，連接規則などを作成するには高度な専門知識が必要のため，一般，日本語教育に携わる者が辞書規模を大きくするにも難点がある。

WinMorph1.2 で採用されている活用体系が，日本語教育に適している点は評価できる。筆者等は，Chasen 解析性能の高さと，WinMorph1.2 で採用されている活用体系の教育利用の適切さを活かしたシステム開発を行った。Chasen の形態素解析結果の形態素単位を変換することで，直感的で分かりやすい用例検索用のタグ付けを実現した。このタグ付けに従った文型照合のためのパターン(検索式)記述は，日本語教育現場の日本語教育者にとってより分かり易く，その結果，システムが利用しやすくなり，用例抽出に活用ができる。

3. 日本語教育教材作成のための用例抽出システム

現在開発中の日本語教育教材作成のための用例抽出システムは，タグ付された日本語コーパスから，文型パターンに対応した用例を動的に，且つ大量に提供する機能の実現を目指している(図 1)。本システムを利用することで，教材作成者が文型パターンに照合する用例を効率的に見つけることができる。教材作成の負担が軽減すると同時に，教材作成者の言語直感に頼らず，言語の運用事実に基づいた用例を教材に反映することができるだろう。また，日本語学習者が直接用例サイトを利用して自主的な学習を行うことも可能である。システムの開発手順は以下の通りである。

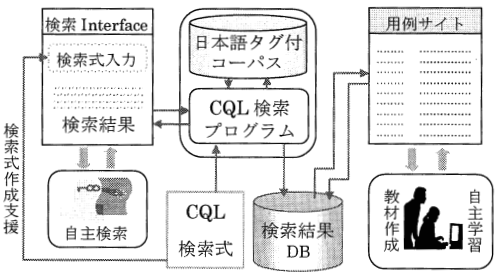

図 1. システムの概念図

[1] 日本語タグつきコーパスを構築する

分野を特定し，その分野から大量に集めた日本語テキストを形態素解析し，形態素分割した文字列にタグ付けを行う。そうして日本語コーパスを作成する(図 2)。形態素解析には Chasen を利用する。しかし，解析システムが付与したタグを直接利用しない。日本語教育に適した言語的な抽象化の単位を用例検索に使うため Chasen の解析結果を元に形態素変換作業を行う。変換作業の必要性とその仕組みについては 4.2 節で詳しく述べる。

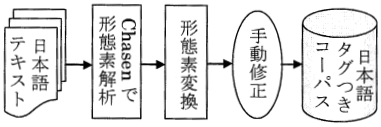

図 2. タグつきコーパスの構築

[3] 富士通研究所が開発した日本語形態素解析システム，フリーで利用できる。

[2] 文型ごとに検索式を作成する

現存の日本語教科書(『初級日本語文法解説』)等で提示されている文型を整理した上，電子処理できるようにパターン化する。文型パターンごとに検索式を作成する。検索式については 4.3 節で述べる。

[3] 検索式を用いてコーパスから用例を抽出するプログラムを作成する

[2]の検索式ごとに[1]で構築したコーパスから自動抽出できるプログラムを開発する。プログラムには学習者のレベルに応じた語彙フィルタと文型難易度ランクを提示できるようにする。また，[4]の用例サイトのため，[2]の検索式ごとに抽出した例文をデータベース化する。

[4] サービス提供のためのインターフェースを作成する

利用者が自主的に検索・抽出できるためのインターフェースを設計する。また，[3]で作成した例文データベースを利用して用例サイトを構築し，インターネット上で用例を配布する。インターフェースについて，利用者の多くが文系出身者であることを考慮し，ユーザにとっての操作性が容易になるよう設計する予定である(後に報告したい)。

4. 文型パターンと検索式

4.1. Chasen の形態素情報

Chasen は，奈良先端科学技術大学院大学・松本研究室で開発された形態素解析ツールである。解析精度が高く，フリーで利用できる。Unix OS だけでなく Windows OS での利用も可能である優れたツールである。

Chasen の形態素品詞体系は，従来の学校文法に基づいて整理された IPA 品詞体系に準じている。表 1 は Chasen で使われる主な品詞を示す。接頭詞や名詞，動詞などはさらに下位分類を持つ，動詞と形容詞は「自立」「接尾」「非自立」のそれぞれが活用型と活用形ごとの下位分類を持っている。詳細については[6]を参照されたい。

表 2 は Chasen の解析結果の一例を示す。

表 2. Chasen の解析結果

表層	辞書形	品詞	活用型	活用形
手紙	手紙	名詞-一般		
を	を	助詞-格助詞		
書き	書く	動詞-自立	五段・カ行イ音便	連用形
まし	ます	助動詞	特殊・マス	連用形
た	た	助動詞	特殊・タ	基本形

表1. Chasen の品詞(抜粋)

品詞		例
連体詞		「この」
接頭詞		「最」,「ぶっ」
名詞		「大根」,「終了」
動詞	自立	「合う」,「たてつく」
	接尾	「られる」,「させる」
	非自立	「しまう」,「ちゃう」
形容詞	自立	「暑い」,「めでたい」
	接尾	「ったらしい」
	非自立	「がたい」,「よい」
副詞		「たいそう」,「あまり」
接続詞		「けれども」
助詞		「の」,「と」
助動詞		「らしい」,「ござる」
感動詞		「うむ」,「トホホ」
記号		「。」,「？」
そのた		「あ」

4.2. 形態素変換

(日本の)国語教育に採用されてきた学校文法にはいろいろ不備があることは，従来から指摘されてきた[3](彭広陸 2003)。彭によると，そのような文法体系は日本語教育に適応しないどころか，その妨げにさえなってしまっている嫌いがある。さらに彭は，日本語教育の観点から見て，学校文法の欠陥が用言の活用表に集中していると指摘し，理想的かつ合理的な活用表を組み立てるた

めの原則として，「断続」という構文的機能に基づくべきであると述べている。

例えば，彭が提案した形容詞活用表では「高かったり(例示形)」「高いと(ト条件形)」「高くたって(タッテ逆条件形)」のように解釈する。学校文法では動詞に後続する「助詞」や「助動詞」として扱う形態素を動詞の活用語尾の一部として考えるのである。

日本語教育では，このような考え方は，形態素単位が明瞭で扱いやすく，学習者にも理解しやすく覚えやすいと考えられる。

本研究では，彭が提案している考え方や，WinMorph1.2 で使われた日本語分析枠組みを基に，形態素として電子化処理の都合を考慮して独自の用言活用体系を整理した。用言を変化しない語幹と活用語尾に 2 分し，さらに動詞に後続する助詞や，「ぬ」「まい」など活用形の持たない「助動詞」を活用語尾と結合させ，「活用助辞」とした。表 3 は本研究で使用した活用助辞の一部を例示したものである。

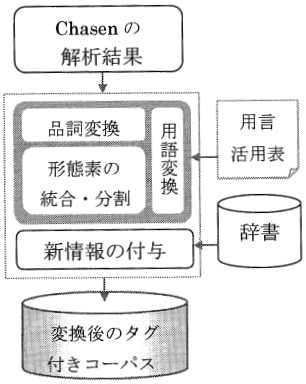

図 3. 変換作業の概念図

表3. 活用助辞の例

		現在	完了	意志	推量	命令	条件-バ	条件-ト1	順序	完了中立	否定-ヌ	否定-ズ	連接-i	連接-a	…
動詞	カ行	く	いた	こう	いたろう	け	けば	くと	いたり	いて	かぬ	かず	き	か	
	カ変	る	た	よう	たろう	い/よ	れば	ると	たり	て	ぬ	ず	ϕ	ϕ	
	⋮	⋮													
形容詞		い	かった	かろう	かったろう	×	ければ	いと	かったり	くて	からぬ	からず	く	ϕ	
⋮		⋮													

Chasen の解析結果を上記の活用体系にしたがって変換するため，一括変換を行うプログラムを作成した。図 3 は変換処理の概念を示す。また変換前と変換後の比較例は表 4 で示す。

表4. 変換前と変換後の例

変換前	書い	て	い	まし	た
	動詞-連用タ接続	接続助詞	動詞-連用	助動詞-連用	助動詞-特殊
変換後	書	いて	い	ϕ	ました
	動詞	活用助辞-完了中立	動詞	活用助辞-連接	活用助辞-完了

本研究で使用した活用体系は，本研究の目的である例文検索処理を円滑に行うために整理したものであり，組織付けの基準や活用形のネーミングなど必ずしも一貫していないところがある。ここで言語学的な意味で一つの文法体系として樹立しようとするものではないことを断っておきたい。あくまで日本語教育に適した区分を実現することを目的としている。この活用体系を用いることで文型照合が容易になっている。詳細は 5 章で述べる。

4.3. 検索式の定義

文型パターンを用いてコーパスから用例を検索する場合，文型パターンごとに検索式を作成する必要がある。検索式は，CQL (Corpus Query Language)式と呼ばれ，独自の記述方法がある。本節では CQL の記述規則を概略的に説明する。

(1) 語の単位毎に検索パターン指定し，検索パターンは必ず波括弧で括る。一つの語の検索パターンを項と呼ぶ。パターンは，キーとその値を等号記号で結びつけた組み合わせで表現する。項は空白で区切ることで複数個書くことができる。項には，ワイルドカード(任意個数の語)が含まれる。

```
'{W="たり"} [0,4] {W="たり"}'
```

(2) キーで語の種類を指定する。キーには，W(表層形)，L(レマ)，P(品詞＋活用型＋活用形)の 3 種がある。L，W，P は，必ず大文字で指定する。キーの並びには制限があり，W L P の順でなければならない。2 つまでキーの省略が可能である。

```
'{W="でし" L="です" P="助動詞-連用形" }'
'{ W="でし" L="です"}'
```

(3) 値の指定は，キーを明示してから値を二重引用符で括り，等号で結ぶ。W キーは，語の表層形態を値として指定する。L キーは，いわゆる語の辞書形態を指定することができる。P キーは，品詞分類名を指定する。この品詞分類は，Chasen で形態素解析した結果をもとに，変換プログラムで定められた品詞分類名である。

```
W="言い"
L="言う"
P="動詞-連用形"
```

(4) W，L，P の各キーは，その値を OR 条件で結合することができる。OR 条件は"|"記号で表現する。

```
'{W="ます|ました|ません"}'
```

(5) ワイルドカードは任意語数の語の連鎖を表現する。ブラケットで n と m を順に指定する。n は 0 以上で，m は 1 以上である。上記の表現の意味は，「前後の項の間に，最低 n 語から最大 m 語の任意の語を含む」ということである。なお，P キーの値に対してのみ"*"，"."を使ったワイルドカードを利用できる。"*"は，ゼロ個以上の任意の文字列である。". "は，任意の 1 文字を表す。さらに OR 条件と組合せることも可能である。

```
WC ::= [n,m]
'{P="動詞.*"}'
'{P="動詞.*|形容詞.*"}'
```

5. 検索式の比較と評価

本稿では，Chasen の解析結果をそのまま利用して作成した検索式を従来方式と呼び（以下略して【従】），Chasen の解析結果を変換プログラムによって変換したあとの形態素を利用して作成した検索式を本提案方式呼ぶ（以下略して【本】）。本章は，文型の具体例を挙げながら，【従】と【本】の両検索式を比較する。

例(1)　N1 は　N2 です

```
【従】{P="名詞.*"}[0,5]{W="は"}{P="名詞.*"}[0,5]{W="です" L="です" P="助動詞-基本形"}
【本】{P="名詞.*"}[0,5]{W="は"}{P="名詞.*"}[0,5]{W="です" L="です" P="活用助辞-現在形"}
```

例 1 では【従】と【本】は「です」に対する品詞名以外，ほとんど同じである。これは，形態素変換プログラムが主に活用する動詞，助動詞，形容詞などいわゆる「用言」に対して行われるもので，例 1 のような文型ではその違いがあんまり現れないためである。

例(2)　N を　V ませんでした

```
【従】{P="名詞.*"}{W="を"}[0,5] {P="動詞-連用形"}{L="ます" P="助動詞-未然形"}{W="ん" P="助動詞-基本形
"}{L="です" P="助動詞-連用形"}{W="た" P="助動詞-基本形"}
【本】{P="名詞.*"}{W="を"}[0,5]{P="動詞"}{P="活用助辞-連接形"}{L="ます" P="活用助辞-否定形"}{L="です"
P="活用助辞-完了形"}
```

まず，検索式の項数において，【従】では 8 項に対し，【本】では 7 項である。これは形態素変換ツールで Chasen の形態素結果を統合・分割したためである。例 2 では，「ませ」＋「ん」→「ません」，「でし」＋「た」→「でした」のように形態素の統合によって数が減少している。なお，本稿で提唱した用言の形態素分割方法では，すべての動詞を「語幹」と「活用語尾(助辞)」の 2 つに分けたため，例 2 のような「動詞＋ます」の場合は「動詞＋活用助辞＋ます」となり，形態素の数が一つ増えている。しかし，「活用助辞」を独立させることによって，動詞の語幹部分について従来の学校文法で説明している「連用形」や「未然形」といった活用形をなくし，すべての用言を一貫して扱うことができる。

さらに，品詞や活用形の名前について，変換前では「まし」→「助動詞"ます"の未然形」，「でし」→「助動詞"です"の連用形」，「た」→「助動詞"た"の基本形」となっている。それに対して変換後では，「ません」→「活用助辞－否定形」，「でした」→「活用助辞－完了形」のように，非常に分かりやすくなっていることが分かる。日本語教育の現場おいてこのような品詞名は教授者にも，学習者にも使用しやすく，理解しやすいといえる。

例(3)　N　に　N が　います｜いません｜いました｜いませんでした

【従 1】{P="名詞-Place"}{W="に" P="助詞.*"}{P="名詞"}{W="が" P="助詞"}{P="動詞-連用形"}{L="ます" P="助動詞-基本形"}

【従 2】{P="名詞-Place"}{W="に" P="助詞.*"}{P="名詞"}{W="が" P="助詞"}{P="動詞-連用形"}{L="ます" P="助動詞-未然形"}{W="ん" P="助動詞-基本形"}

【従 3】{P="名詞-Place"}{W="に" P="助詞.*"}{P="名詞"}{W="が" P="助詞"}{P="動詞-連用形"}{L="ます" P="助動詞-連用形"}{W="た" P="助動詞-基本形"}

【従 4】{P="名詞-Place"}{W="に" P="助詞.*"}{P="名詞"}{W="が" P="助詞"}{P="動詞-連用形"}{L="ます" P="助動詞-未然形"}{W="ん" P="助動詞-基本形"}{L="です" P="助動詞-連用形"}{W="た" P="助動詞-基本形"}

【本 1】{P="名詞-Place"}{W="に" P="助詞.*"}{P="名詞"}{W="が" P="助詞"}{P="動詞"}{P="活用助辞-連接形 i"}{L="ます" P="活用助辞-現在形|否定形｜完了形｜"}

【本 2】{P="名詞-Place"}{W="に" P="助詞.*"}{P="名詞"}{W="が" P="助詞"}{P="動詞"}{P="活用助辞-連接形 i"}{L="ます" P="活用助辞-否定形}{"L="です" P="活用助辞-完了形"}

4.3.1 節の CQL 式の定義により，項の選言(OR)は使えない。そのため項数が違う場合一つの式にまとめることができない。例 3 の【本】では「います」「いません」「いました」は同じ 2 項であるため一つの CQL 式にまとめることができる。「いませんでした」は 3 項のため，別の検索式を作成した。一方，【従】では上記のように 4 つそれぞれに対して単得に検索式を作らなければならない。つまり，否定や過去を表す文末表現だけが異なる文型に対して【本】はより少ない検索式で表現することが可能である。

例(4)　V1 たり，V2 たり　し　ます|ました|ましょう

【従 1】{P="動詞-連用形|連用タ接続"}{W="たり|だり" P="助詞-並立助詞"}{P="記号-読点"}[0,5]{P="動詞-連用形|連用タ接続"}{W="たり|だり" P="助詞-並立助詞"}{L="する" P="動詞-連用形"}{L="ます" P="助動詞-基本形"}

【従 2】{P="動詞-連用形|連用タ接続"}{W="たり|だり" P="助詞-並立助詞"}{P="記号-読点"}[0,5]{P="動詞-連用形|連用タ接続"}{W="たり|だり" P="助詞-並立助詞"}{L="する" P="動詞-連用形"}{L="ます" P="助動詞-連用形"}{W="た" P="助動詞-基本形"}

【従 3】{P="動詞-連用形|連用タ接続"}{W="たり|だり" P="助詞-並立助詞"}{P="記号-読点"}[0,5]{P="動詞-連用形|連用タ接続"}{W="たり|だり" P="助詞-並立助詞"}{L="する" P="動詞-連用形"}{L="ます" P="助動詞-未然ウ接続"}{W="う" P="助動詞-基本形"}

【本】{P="動詞"}{P="活用助辞-順序形"}{P="記号-読点"}[0,5]{P="動詞"}{P="活用助辞-順序形"}{L="する" P="動詞"}{P="活用助辞-連接形 i"}{L="ます" P="活用助辞-現在形|完了形|意志形|"}

例 4 では，例 3 で述べた文末表現に対する処理方法の違いのほか，「たり｜だり」のような日本語の音便変化に対する処理方法の違いも見られる。【従】では，{P="助詞-並立助詞"}だけで「たり｜だり」を特定できないので，{W="たり|だり"}のように音便変化を考慮した上，表層形を指定しなければならない。これに対して，【本】では{P="活用助辞-順序形"}のみで「たり｜だり」を指定することができ，検索式を作成する際に音便変化を考慮する必要がなくなる。

以上の例から見た新【従】の比較から，【本】の特徴を以下のようにまとめる。

(1) 項数が少ないため，記述がよりシンプルになる
(2) 品詞や活用形の名前は分かりやすく，日本語教育の事情に適合している
(3) 文末表現のみ異なる文型をまとめて検索式を作成することが可能である
(4) 音便変化への考慮を省くことができる

6. まとめ

研究開発中の用例抽出システムを概略的に紹介した上，本システムの用例検索部における形態素変換のプロセスおよびシステム的に改善した点について述べた。日本語教育に携わる人達に利用してもらうために，日本語用言活用体系を新たに整理し，それに基づき Chasen の形態素解析結果の変換プログラムを作成した。そして，用例検索式の比較に基づき，システムの評価を示した。

今後，さらに検索式の作成枠組みを改良した上，用例抽出プログラム部分，用例提供サイトおよび検索インターフェースの作成を予定している。

参考文献

[1] 山本樹,「文型パターン解析を利用した日本語教材作成支援システム」東京外国語大学大学院修士論文，2006

[2] 佐野洋,「ドメイン分析を応用した日本語研究ルールの教育ソフトウェアへの適用研究ー形態素解析ツールにおける開発事例」，学術情報処理研究編集委員会，No.4, 2000

[3] 彭広陸,「日本語教育における新しい文法体系の構築のために--用言の活用表を中心に」『国文学解釈と鑑賞』巻号 68(7) , p51-61, 2003

[4] 掛川淳一，中村宏，関谷政則，伊丹誠，伊藤鉱二,「自然言語処理を用いて日本語教育のための例文検索を支援するシステム」，日本教育工学会雑誌，券号 25(2), p85-94, 2001

[5] 東京外国語大学留学生日本語教育センター『文法解説』　凡人社　2001

[6] 浅原正幸，松本裕治，「ipadic version2.6.3 ユーザーズマニュアル」2003 http://chasen.aist-nara.ac.jp/chasen/doc/ipadic-2.6.3-j.pdf (last check 2006/11/07)