自 然 言 語 処 理 100−13
（1994．3．18）

構文テキストベースの構築と意味分類コードを用いた類似例文検索への応用

兵藤安昭　河田実成　青山典生　浅井泰博　池田尚志

岐阜大学工学部

本稿では，係り受け情報を付与したデータベース（構文テキストベース）を用いた類似例文検索システムと構文テキストベースを作成するための解析支援ツールについて述べる．類似例文検索は，各単語を分類語彙表を用いて意味コード化し，その階層性を利用して類似性を処理するものである．検索パターンに係り受け構造を指定することができるので，ユーザの検索要求に柔軟に対応することができるシステムである．構文テキストベース作成支援ツールは，Ｘｌｉｂ上のＧＵＩとして作成したもので，計算機による形態素・構文解析結果を係り受けの三角表現を用いて表示し，非常に簡単な操作で後編集を行なうことができるようにしたシステムである．

TEXTBASE WITH SYNTAX STRUCTURE AND IT'S APPLICATION TO SIMILAR SENTENCE SEARCH

Yasuaki Hyodo　Yoshinari Kawada　Norio Aoyama　Yasuhiro Asai　Takashi Ikeda

Gifu University

This paper describes a similar sentence serach system over the textbase with syntax strucuture (tree bank), and an analysis support system for building such a textbase. The similar sentence search system encode words by using of semantic classification code. The system searches similar sentences based on syntactic structure and utilizing the hierarchy of the semantic code. The analysis support system has a graphical user interface. It displays a syntactic structure in the trianglar expression and a user can correct very easily a computed morphologic and syntactic structures through it's graphical user interface.

1 はじめに

近年，辞書，新聞記事など大規模な電子化データベースが存在するようになってきた．しかし，その大部分は原テキストのまま電子化されたものである．ＳＧＭＬなどによって構造化されたものも見られるようになったが，テキスト自体が構造化されたものはまだ見られない．

テキストデータベースは多様な検索が可能となることで大きな意味が出てくるわけだが，原テキストだけのデータでは，類似用例検索などの高度な応用に柔軟には対応できない．形態素解析処理を行なって，単語の品詞や活用形などの情報を付与したものはみられるが[1] [2]，文の構造を指定して用例を検索することにまでは対応し難い．

我々は，係り受け解析の情報を付与したデータベース(構文テキストベース)の有効性を確認するため類似例文検索システムを作成し実験した．また，構文テキストベースを作成するための解析支援ツールを構築した．

類似例文検索は，各単語を分類語彙表を用いて意味コード化し，その階層性を利用して類似性を処理するものである．検索パターンに係り受け構造を指定することができるので，従来の検索システムより，ユーザの検索要求に柔軟に対応することができる．

構文テキストベース作成支援ツールは，Ｘｌｉｂ上のＧＵＩとして作成したもので，計算機による形態素・構文解析結果を係り受けの三角表現を用いて表示し，非常に簡単な操作で後編集を行なうことができるようにしたシステムである．前述のような類似例文検索を実用的なものにするためには，正確に解析されたゆらぎのない構文テキストベースの存在が前提となる．しかし現段階では完全自動的に，正確に形態素・構文解析することは難しい．従ってこのような支援ツールの重要性は大きい．

2 構文テキストベースと例文検索システム

2.1 構文テキストベース

我々の類似例文検索システムでは，図1のような係り受け構造を持ったデータベース（構文テキストベース）を検索する．本実験では，新聞記事等から選んだ約３５００文を人手で形態素，係り受け構造を付与したデータベースを使用した[3]．この構文テキストベースは，人手で作成しているため多くの誤りを含んでいた．そこで次章で述べる支援ツールを用いて修正を加えた．

```
1:原テキスト
スイスはたくさん雪が降ります
2:係り受け構造
(@1 ((スイス) は 体用)
 @2 ((雪) が 体用)
 @3 ((たくさん) 副用)
 @4 ((降る @1 @2 @3) ます 用終))
```

図 1: 構文テキストベース

2.2 類似例文検索システム

2.2.1 システム概要

システム構成図を図2に示す．検索対象となる構文テキストベース中のテキストデータは，次節に述べる方法であらかじめコード化してある．検索指示文節構造を入力すると，同様のコード化を行ない，検索対象データとの照合を行なう．照合は，文節コードと構造コードの照合に分けられる．まず，文節コードに一致する文章を検索する．文節コードでの検索では，意味的曖昧性を考慮した検索がなされる．次に，文節コードの照合をパス

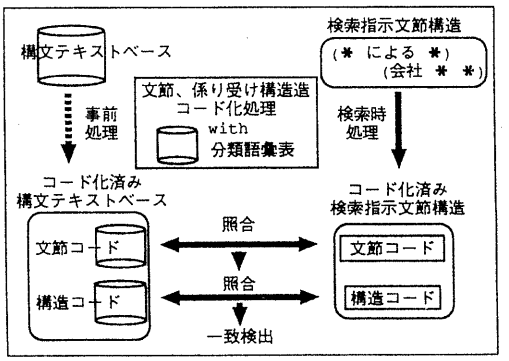

図 2: システム構成図

した文章に対して，構造も一致するかの検査を行なう．これら，２つのコードで一致を得ると検索結果として出力される．検索結果は，曖昧一致レベルと共に一致箇所を反転し強調して出力する．

2.2.2 構文テキストベースのコード化

構文テキストベースのコード化によって，文節コード，構造コードの２つのコードが作成される．

文節のコード化 文節中の自立語を図3のように分類語彙表[8]による意味分類を用いた番号付けによってコード化する．このコード化によって単純な前方一致検索をするだけで，その一致位置によって次のように，完全一致を含めた３つのレベルの曖昧検索が可能になる．

1 A,B,C が一致： 曖昧一致レベル 0 (完全一致)
2 A,B のみ一致： 曖昧一致レベル 1
3 A のみ一致： 曖昧一致レベル 2

機能語と文節カテゴリは単純な番号づけによってコード化している．

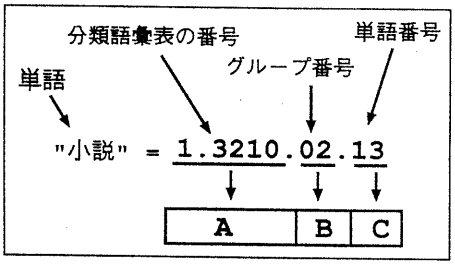

図 3: 自立語のコード化

係り受け構造のコード化 係り受け構造は，構文テキストベース中の@に続く文節番号で表現されている．このような係り受け構造のリスト表現を，図4のようにコード化した．ここで，文節番号 i の係り先は，dest[i] となっている．

2.2.3 検索アルゴリズム

検索のプロセスは以下のようである．

(1) **検索指示データのコード化** 入力インターフェースから検索指示文節構造を受けとり，それ

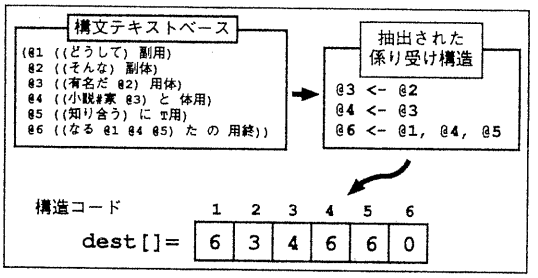

図 4: 係り受け構造のコード化

に検索対象データ作成時と同様のコード化処理を施して，コード化済み検索指示文節を得る．これは，データ構造的には検索対象データと全く同じものである．

(2) 文節コードの照合 次に，各文の文節コード化データに対し，検索指示文節構造内の文節がすべて含まれるかを調べる．文節内の自立語，機能語，カテゴリの３要素が，その情報量（データの種類の多さ）の順に，自立語＞機能語＞カテゴリと，なっている．そこで，文節毎の照合では，自立語を始めに照合して，その一致を確認後，機能語とカテゴリを検索することによって，無駄な途中までの一致を極力回避するようにしてある．指定文節がすべて含まれる文章に対しては，各指定文節に一致のレベルと文節番号を付与する．一致のレベルは検索結果出力時に表示され，一致した文節番号は，次の構造検索で用いる．

(3) 構造コードの照合 (2) を満足した文章は，一致した文節番号の情報を基に，指定された構造と同じかどうかを調べる．ここで調べなければならないことは，一致した２ノード間の構造のみである．木構造データに対しては，一般にルートノードから目的のノードを探す検索が考えられるが，ここでは上記の目的を実行可能でありさえすればよいので，それを実現する単純なアルゴリズムを用意した．

図5で具体例を示す．検索指示文節構造が，図のように＃に続く番号で表現されているとすると，文節コードでの検索において，すでに＃１と@５，＃２と@４が一致していることが判明している．したがって，ここでは@５ — @４間の構造

が，＃1 — ＃2間の構造と同じであれば，構造が一致したといえる．（ここでは単に，＃1の下に＃2があるのと同様に，@5の下に@4があるかどうかを調べることに相当する．）文節コード，構造コードの両方について一致した文章を構文テキストベースから出力する．

(4) **検索結果の出力**　(3)を満足した文章を検索結果として出力する．

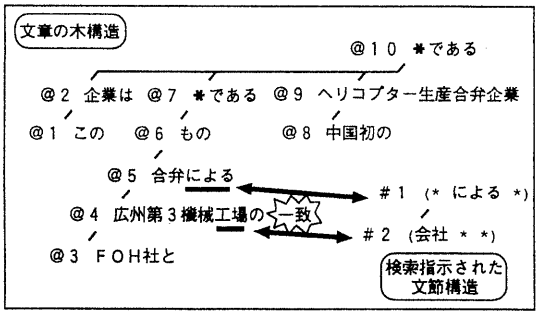

図 5: 検索アルゴリズム

2.2.4　入出力インターフェース

検索システムのインターフェースには，日本語emacsを用いた．インターフェースは図6のように３つのウィンドウを持ち，上から，検索指定文入力ウインドウ，検索結果表示ウインドウ，構造表示ウインドウとなっている．検索指定文入力ウィンドウで，検索指示文節構造を入力する．検索パターンは，図6の例にあるように構文テキストベースの書式で，自立語，機能語，カテゴリの順に入力する．係り受け構造を指定する時にはタブを用い，図6の例では「英語」が「学ぶ」に係る例文を検索することを表現している．また，”＊”を用いてワイルドカード指定も可能である．検索結果表示ウィンドウには，検索された文章と文章番号を表示する．ここで指定した文章について，その文章の構造を構造表示ウィンドウで確認することもできる．

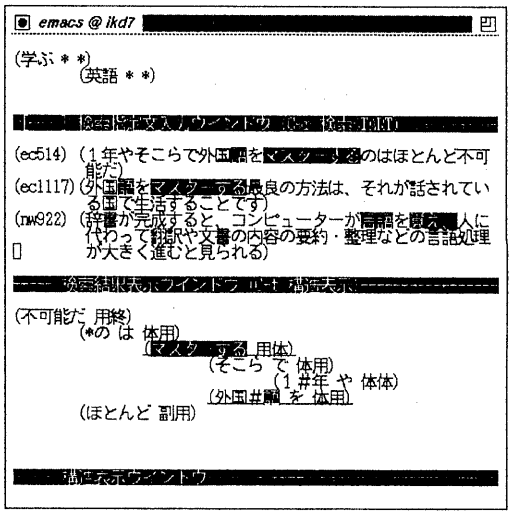

図 6: インターフェース

2.2.5　検索例

以下に，いくつかの検索例を示す．

(1)
(移す * *)
(* に *)
(* を *)

(ec115)　(１０分も車を走らせれば美術館に着きます)

(ec414)　(彼は時を移さず香港に向けて出発した)

(nw265)　(合併後は社名を「日立ツール」に変更し，本社も東京都江東区に移転する)

(nw369)　(富士通は今年四月に一・一四ギガＦＬＯＰＳのＶＰ４００を「実測で世界最高速」と銘打って発表しており，こうしたことから日電が異例の出荷前の実機公開に踏み切った)

(nw1710)　(広州二十五日発新華社電によると，中米合弁のヘリコプター生産企業が一年間の準備を経て，来月生産を開始する)

(tb334)　(こうした複雑かつ深刻な状況の下で，現代のデモクラシーは，その制度と運用の両面に わたって新しいありかたを模索する必要に迫られているといえよう)

従来までの単なる複数キーを含む検索は，このように係り受け構造を指定しないことで実現できるが，これではただ単に「移す」，「に」，「を」が含まれているだけの文章が検出されてしまう．

(2)
```
(移す * *)
    (* に *)
    (* を *)
```

(ec197) (この荷物を 2階に 運ぶのを手伝ってくれませんか)

(nw678) (このため，興銀が直接，ディーラーに 行員を 派遣する 考えはなく，買収資金で依頼があれば応援するだけ，という)

(nw1668) (日立造船は十三日までに，来月一日付で新たな子会社七社を設立し，本社から約千人をそちらに 移す 合理化計画をまとめた)

(tb359) (政治がこのようなはたらきをするためには，社会の成員の行動に 影響を およぼし，意図した効果を生みだす力を必要とする)

「～に～を移す」という係り受け構造をを指定することによって，その構造を満たす文章だけを検出することができる．構造を指定しない(1)の場合は，８５文検出したが，この例では６文に絞り込むことができた

(3)
```
(提携 * *)
    (企業 * *)
```

(nw73) (複合経営をめざす新日本製鉄が米国のベンチャー 企業との 提携や合弁会社を相次いで進めている)

(nw192) (鈴木は新工場の建設や 生産 技術で 協力し，日商岩井は機械設備のあっせんをする)

(nw1077) (横浜ゴムと東洋ゴム工業は十四日，西独の大手タイヤメーカーのコンチネンタル社（本社・ハノーバー）との三社間で，技術交換や 生産 に関して 提携することに合意した，と発表した)

(4)
```
(学ぶ * *)
    (英語 * *)
```

(ec514) (１年やそこらで外国 語を マスターする のはほとんど不可能だ)

(ec1117) (外国 語を マスターする 最良の方法は，それが話されている国で生活することです)

(nw922) (辞書が完成すると，コンピューターが 言語を 覚え，人間に代わって翻訳や文書の内容の要約・整理などの言語処理が大きく進むと見られる)

(3)，(4)で見るように意味コード化による曖昧検索によって，広範囲な類似例文の検索が可能である．

3 構文テキストベース作成支援システム

3.1 システム概要

図7に本システム構成，図8に使用例を示す．本システムは形態素解析処理部，構文解析処理部，インターフェース部からなる．インターフェースはＸｌｉｂを用いて構築し，形態素解析後編集ウインドウ，構文構造表示後編集ウインドウ，辞書登録・文節カテゴリ訂正ウインドウを設けた．形態素・構文解析システムはＬｉｓｐ言語を用い，解析サーバとしてネットワーク上に実現されている．

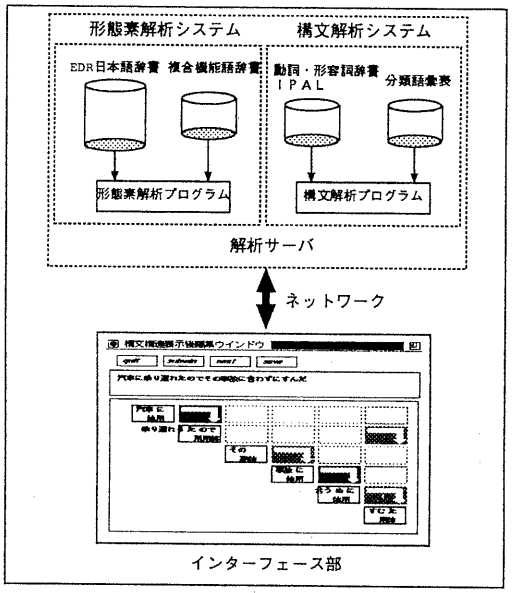

図 7: システム構成図

ユーザが解析テキストの選択を行なうと，計算機により自動的に形態素・構文解析処理を行い，解析結果がインターフェース上に表示される．単語や文節の認定を誤っている場合は形態素解析後編集ウインドウ上で文節の切り直しを行ない，再度構文解析を行なう．構文解析結果は係り受けの三角表現を用いてウインドウ上に表示する．構文解析の後編集は三角表上で正しい係り先をマウスで指示するという簡単な操作で行なうことができる．

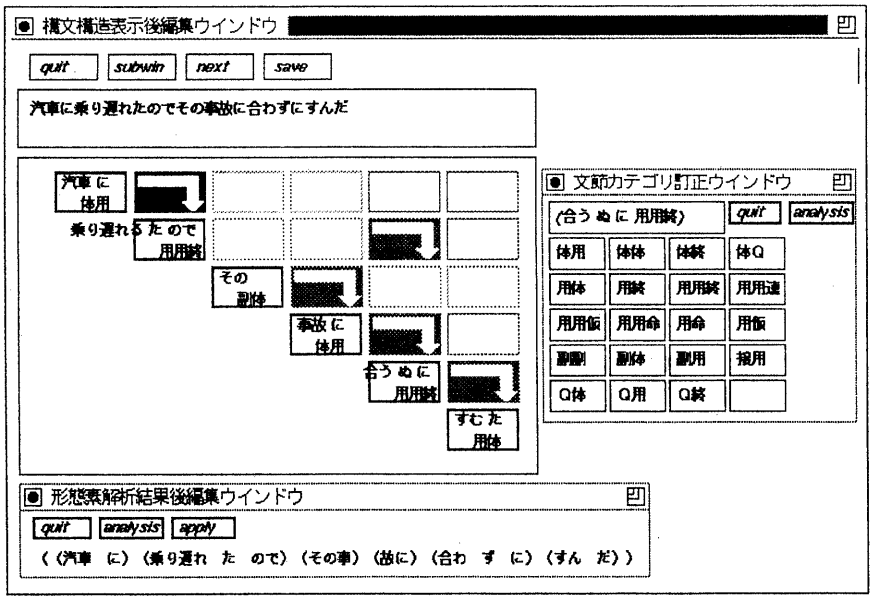

図 8: システム使用例

3.2 形態素解析後編集システム

3.2.1 辞書および解析アルゴリズム

形態素解析に用いた辞書は，自立語としてＥＤＲ日本語辞書 [4] より単語の見出し，品詞情報のみを取りだし，機能語については我々が実際のテキストベースから収集し拡張，整理した辞書 (見出し数約１５００語) を用いた [5].

形態素解析のアルゴリスムは，すべての解析の可能性を後続部分に伝搬させることはせず，最長の文節のみを次の解析につなげていき，全体としては１つの解析結果のみを後編集の対象として提示するという方法を用いている.

3.2.2 解析結果後編集，辞書登録機能

形態素解析された結果はウインドウ上に表示し，そこで誤って切っている箇所，誤って切ってない箇所にマウスを合わせボタンを押すという簡単な操作で形態素解析結果の後編集を行なうことができる (図 8). また未知語が出現した場合は，ウインドウ上で品詞名を選択することにより辞書登録され，辞書登録された単語はこの後の解析に直ちに利用される (図 9).

3.2.3 文節カテゴリ後編集機能

形態素解析処理が施されたテキストには，文節カテゴリが付与されている. 文節カテゴリは，各文節自身のタイプと係りうる文節のタイプによりカテゴライズされたもので，構文解析を行なう時に利用される. 文節カテゴリの訂正はメニュー上で正しいカテゴリを選択することで行う (図 8).

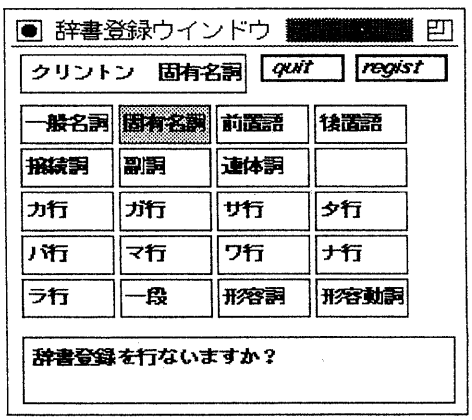

図 9: 辞書登録機能

3.3 構文解析後編集システム

3.3.1 辞書および解析アルゴリズム

本システムでは，解析用辞書として計算機用日本語基本動詞・形容詞辞書ＩＰＡＬ[6, 7]，分類語彙表[8]を用いた．ＩＰＡＬ辞書には各用言がどのような格をとり，それらの格にどのような名詞が入りうるかが記述してある．そこで用言に依存する体言の依存先を決定するために，辞書中に記述してある名詞の例と実際の名詞との類似度を分類語彙表を用いて求め，評価値とした[9]．また分類語彙表の見出し語すべてに，ＩＰＡＬ辞書で使われている意味素性を付与したので，上記の方法で評価値があたえられない場合は，意味素性が等しい場合にも特定の評価値を与えるようにしている．

解析アルゴリズムは，考えられる係り受け構造の候補を求め，それらに対して評価値の計算を行ない，評価値が最大のものを解とする方法を用いた．ただし解析結果が複数ある場合は，近い文節に係るものを多く含んでいるものを解としている．係り受け構造の候補を求める方法としては，文節カテゴリを用いて各文節がどの文節に依存可能かを調べ，非交差条件を利用しながら考えられるすべての係り受け構造を求めている．

3.3.2 構文構造の表示方法

係り受け構造の表示方法として三角表現を用いた．三角表現を用いることにより，ユーザは簡単に非交差条件を画面上で確認することができ，後編集作業を容易に行なうことができる．また各文節の依存可能な文節と不可能な文節が区別できるように色を変えて表示しているので，後編集を行なう時に間違って選択しないようにもなっている．

3.3.3 構文構造の縮退，復元表示

文の係り受け関係をウインドウ上に表示する際には，長い文の表示方法が問題となる．長文の構造を把握するためには，文全体の構造を一度に見ることができるＧＵＩが必要となるが，本システムでは特定のスコープ内にある文節を１つに縮退して表示する方法を用い，文全体の構造を表示できるようにした．この方法は非交差条件を用い，図10の文節１が４に係る時は文節２，３は４より先の文節には係ることができないので文節１から４までの文節を１つの文節として表現している．またユーザが縮退させた文節の中身は別のウインドウに表示することも可能である(図11)．

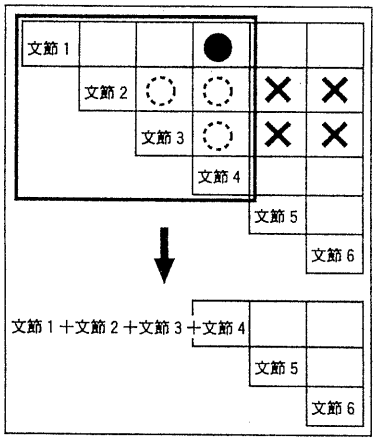

図 10: 非交差条件を用いた係り受け構造の縮退

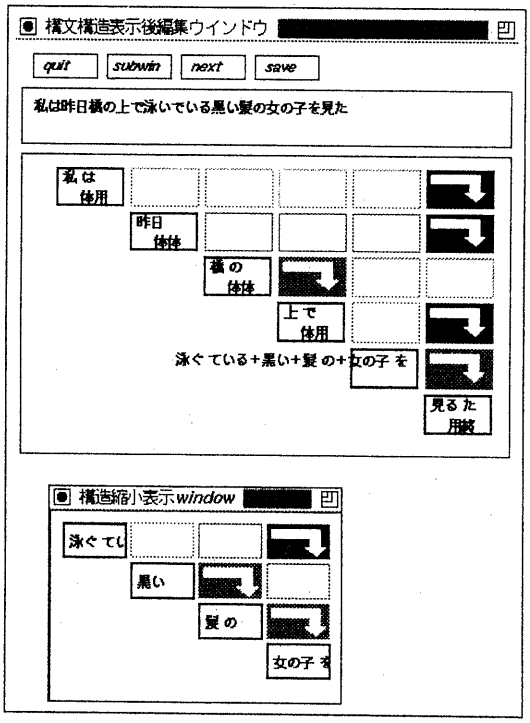

図 11: 構文構造の縮退表示

3.3.4 構文構造の後編集機能

係り受け関係が誤って解析された文節は，正しい係り先をマウスでクリックすることにより後編集を行なうことができる．ただし，各文節の係りうる文節のタイプは文節カテゴリにより決まっているので，係ることのできない文節は色を変えてユーザーに分かるようにし，選択することもできないようになっている．また非交差条件によりその時点では係ることのできない文節も選択することができない．

現在，講談社和英辞書[電子技術総合研究所で電子化されたもの]の用例文(約6万用例)を対象として本システムを用いて，構文テキストベースの構築を行なっている．１００用例(１用例，約３０語)を解析するのに約１時間を要した．

4 おわりに

構文テキストベースを用いた類似例文検索システムを構築，実験し，また構文テキストベースを作成するための解析支援ツールを構築した．

我々が以前構築した例文検索システム[10]では，シソーラス辞書[11]を用いて，名詞部分のみ同義語と下位語に展開し，検索を行なっていた．これに比べて，今回のシステムは意味コード化の方法を用いることによって検索速度が向上し，また，より広範囲に語を展開する分類語彙表を用いたことにより，語の意味で幅広く例文をとらえることができた．一方で，文章構造が指定できるので，文の構造を絞り込んだより厳密な検索が可能である．このような結果から，本システムでは検索者が引きたいと考えた文をより正確に検索結果に反映する手段を得ることができた．今後さらに，構造指定の柔軟さや，意味分類コードの適切な割当などについて検討したい．

構文テキストベース作成支援システムは，係り受け構造を表示する方法として三角表現を用いたことにより，非常に簡単に解析結果を後編集することができるツールを構築することができた．上で述べた例文検索システムを実用的なものにするために，このシステムを用いることにより，より大規模な構文テキストベースの構築が可能になると考えている．今後は，構文解析用辞書の学習機能の導入など構文解析精度の向上を検討していきたい．

{形態素解析システムでは，「日本語単語辞書評価版第２．０版©株式会社日本電子化辞書研究所」を使用した．}

参考文献

[1] 隅田，堤：翻訳支援のための類似用例の実用的検索法，信学論(D-II) ７４Ｄ-II，１０，ｐｐ．１４３７－１４４７，１９９１

[2] 三吉，小渕，濱田，秋山：構造化キワードを用いた用例検索システムの試作，情報処理学会自然言語処理研究会，７０－７，１９８９

[3] 池田，他：中間言語および用例集の開発に関する報告書及び資料集，電子技術総合研究所，１９９０

[4] 日本電子化辞書研究所：日本語単語辞書評価版第２．０版，１９９３

[5] 兵藤，池田：スロット表現による複合機能語の処理，情報処理学会第４５回全国大会，１９９２

[6] 情報処理振興事業協会：計算機用日本語基本動詞辞書ＩＰＡＬ，１９８７

[7] 情報処理振興事業協会：計算機用日本語基本形容詞辞書ＩＰＡＬ，１９９０

[8] 国立国語研究所：分類語彙表

[9] 黒橋，長尾：格構造解析への評価関数の導入による統語的曖昧性の解消，情報処理学会自然言語処理研究会，９２－９，１９９２

[10] 兵藤，池田：構文指定による用例検索システムＴＷＩＸ，情報処理学会第４３回全国大会，１９９１

[11] 荻野：現代日本語の名詞シソーラスの作成，文部省特定研究報告集，１９８９