データベースシステム 94−24
（1993．7．23）

大規模電子化辞書DB管理プログラム作成のための自然言語インタフェース

森 義和　　落合 尚良
（株）日本電子化辞書研究所

大規模な辞書では、データの維持管理を行うためのシステムは必要不可欠である。システムには、逐次検索をはじめ体系的なチェック、一括した更新などの機能を要求される。これらの機能を扱いやすくするため日常使用している自然言語をインタフェースとして用いることは有望な手段であると考える。また、これらの管理要求の多くは個別の操作手順を組み合わせることにより初めて実行することができる。このため、個別の操作手順を自然言語で指示するだけでなくこれらの操作手順をまとまりとしたプログラムを作成する機能が必要である。

本稿では、辞書管理作業に必要な指示を問い合わせ例文として収集し、この例文集をもとに自然言語解析上の課題と方策を述べる。さらに辞書管理要求における基本操作単位を抽出し、この基本操作単位をもとに設計したDB管理プログラム作成システムとその動作例について述べる。

Natural Language Interface of Database Managing Programming for Electronic Dictionaries

MORI Yoshikazu and OCHIAI Takayosi
Japan Electronic Dictionary Institute,Ltd

A managing system is indispensable for a large scale electonic dictionary to be well maintained. This system requires functions such as independent retrieval, systematical chek, over-all rewite and so on. It is efficient to use natural language as the interface, which brings the advantage of easy and useful usage. The managing requirements are often performed by a combination of several individual procudures. In this system, besides adopting the natural language interface, we designed a program to form a series of procedures which are related with each other from the viewpoint of electronic dictionary managing. This paper presents some problems and thier solutions of natural language processing, which were derived from the example sentences prepared for managing the dictionaries. Some demonstrations are introduced of our database managing programming system to accomplish both the macro operations and their basic individuals.

1. はじめに

大規模電子化辞書の開発には、新規語の登録やエラーの修正、情報の追加、品質の検査等の作業が必須である。これらの作業および組織的な大規模データの維持管理を行うにはデータベースを用いるのが適切であろう。しかしながらデータベースを利用するには、その利用技術やデータベースの専門的な知識を必要とする。辞書管理作業を行う作業者はかならずしもこれらの知識に明るいとは限らない。このようなデータベースに暗い作業者にとって、データベースの利用技術や知識を習得するにはかなりの負担となる。データベースを簡便に利用できるように各種のインターフェースが研究されてきた。SQL[1]などの形式的な言語が主流であるが、データベース上での辞書の構成や形式言語のプログラミングなど知っている必要があり作業者にとって容易なものではない。またQBE[2]にみられる表形式のインタフェースや近年にみられるGUIを用いたインタフェースがある。これらのインタフェースは覚えやすく使いやすい利点があるが、前者においては検索結果が作業者の意図通りとはならい場合があり、後者では利用者によってアイコンの意味を様々に解釈しうる曖昧さが生じる可能性がある。また、複雑な検索においては柔軟に対応できない可能性がある[3]。作業者の負担を軽くするために、データベースのインタフェースとして新たに文法を修得する必要なく、表現の自由度の高い日本語を手段として用いることは有望であると考える[4]。さらに、前述のような辞書管理要求は、いくつかの機能を組み合わせた一連のプログラムに相当する記述となることが多い。従って、本システムで要求される日本語インタフェースは、個々の機能を日本語で操作するだけでなく、一連の作業をプログラミングするのに相当する枠組も備える必要がある。

本稿では、辞書管理作業のための管理システムに問い合わせる問い合わせ例文集をもとに、日本語解析上の課題を抽出し、その解決の方策を述べる。さらに管理作業における基本的な操作単位を抽出し、この操作単位をベースに設計した日本語によるプログラミングのための日本語による辞書管理プログラム作成の支援システムについて述べる。

2.システム設計のための業務の分析

実際に辞書の管理業務を行っている2人の担当者に業務内容を日本語で記述させ、管理業務コーパスの収集を行った。

収集した質問文コーパスより、データベースの応答時間によって即時に答えが分かる即応答性の質問と時間がかかる遅応答性の質問に分類でき、さらにデータベース全体もしくは広い範囲内での網羅的な処理と個別的な処理に分類できることが明らかになった。

2.1即応答性の質問

1)見出し語などに対する属性値の検索

例文1：書くの品詞を教えて下さい。

2)辞書の個別情報に対する更新、追加、削除

例文2：読むで概念ＩＤが016ab0のレコードを削除。

2.2遅応答性の質問

1)属性間の整合性の検証

例文3：品詞が動詞の見出し語を教えて下さい。

例文4：品詞が動詞で左連接属性がJLV1で右連接属性がJRV2の見出し語を昇順にファイルListに出力しなさい。

2)辞書情報に対する一括修正、追加、削除

例文5：ファイルＡの内容を辞書に追加して。

3)辞書全体に対する数量調査

例文6：概念見出しの平均数を教えて。

例文7：最も登録数の多い品詞をあげて。

4)辞書自体に対する処理

例文8：単語辞書のバックアップをとって下さい。

3. 問題の分析

業務の種類を分類したコーパスを言語処理的観点から分析したところ以下のような日本語解析上の問題が明らかになった。

3.1.長文に対する解析精度の問題

利用者が管理要求を完全に表現することは難しい。

また利用者が問い合わせを曖昧性がないように表現すると冗長にならざるをえない。

しかし、冗長な問い合わせ文は一見で理解できるものとはいえない。

また、問い合わせ文が長くなると解析精度が低くなり、係先の曖昧性の増加により構文解析が正しい解析結果を出力できない可能性がある。

例文11：頻度情報が２以上の名詞か動詞で連接属性がJLV1又はJLV2を持つ単語の、概念見出しと文法情報と頻度情報を見出し語毎に表示してください。

上記の例文には係り受け関係にいくつかの曖昧性がある。さらに長文になると、この係り受けの曖昧性が増え、構文解析が正しい解析結果を出力できない事や、ユーザの意図とは別の解を数多く作成してしまい、ユーザがこれらの中から選択できないなどの問題がある。

3.2指示代名詞や接続詞の曖昧性の問題

辞書情報の整合性をチェックする等の条件を一文で表現すると大変長くなるので、読みやすくするため、いくつかの文を指示代名詞や接続詞等の指示代名詞相当表現を用いてつなぐことがある。これらの条件を解釈するためには、入力した複数の文から指示代名詞相当表現で必要な情報を参照できるように保持しなくてはならない。また、指示代名詞相当表現の持つ係り先の特定には、背景知識や文脈情報を持っても正しく特定できないなどの問題がある。

3.3 条件に用いる値に関する解析上の問題

例文9：読みがあいの見出し語を検索

上記の例文では、単語辞書の「読み」という属性名に格納する属性値「あい」を持つ単語の見出し語を検索する文である。この文にある「あい」のような属性値として使用される単語すべてを解析辞書に登録することは不可能である。このため、未登録語になり、解析に失敗する可能性が高くなるという問題がある。また、ユーザが意図しない別の解で解析する等の問題もある。

例文10：読むの品詞を検索

上記の例文では、ユーザの意図としては「読む」は単語辞書の見出しの値であるため、固有名詞として扱っているので、辞書管理のためのコーパスとしては正しい構文である。しかし、この「読む」を動詞として解釈すると誤った構文となる（*書くの論文→書く論文）。したがって、構文解析を失敗する可能性がある。

4.入力文形式の実用的制約

日本語の持つ曖昧性や、解析結果の精度から生じる問題の解消をユーザとの対話を用いて行うことは有効な手段であると考えられている。しかし、前掲例文11で示すような曖昧性を解消するためには、煩雑な会話を人間とシステムの間で行う必要があり実用的で効率的なインタフェース構築の視点からは望ましくない。そのため、日本語の記述において自然さを損なわない程度の制約（規約）を設け、これらの問題を解消する方法を選択した。

4.1属性値を「」で括る入力

属性値を日本語で入力する場合、前掲例文9から生じる問題や、前掲例文10で発生する問題に対応するため、属性値を「」を用いて括ることにより固有名詞として扱うことができる。これらの例文で発生する問題を解決できる。

前掲例文9、10をこの規約を用いて記述した例文を以下に示す。

例文12：読みが「あい」の見出し語を検索

例文13：「読む」の品詞を検索

4.2個別条件の集合による入力

長文解析の種々の問題を回避する為に、文の分割について検討した。その結果、複雑な検索文であっても、それほど不自然な文分割にはならず、場合によっては文意を明確化できることがわかった。これらの多くの管理・検索要求は以下に上げる日本語パターン例文の組み合わせによって表現できる。

・属性名と属性値に対する条件の例

1)Nのみ　見出し語(を検索する)

2)NのN　「読む」の品詞

3)NがNの(である)N　品詞がJVEの単語

4)NでNをV+N　動詞でJLV1を持つ単語

5)NとN　品詞と表層格情報

6)NかN　形容詞か形容動詞

(N=属性名、又は属性値)

・処理に対する条件の例

1)出力形式に対する条件　昇順に出力

2)出力先の条件　ファイルAに出力

このような基本的な操作の単位を個別条件と呼ぶ、これらを用いて前掲例文11で示す文について、分割した例を以下に示す。

例文14：頻度情報が2以上の名詞を検索

例文15：それは動詞で連接属性がJLV1又はJLV2を持つ単語も検索

例文16：それは概念見出しと文法情報と頻度情報を表示する。

例文17：それは見出し語毎に表示します。

このように操作単位の文に分割できば、一文内の構文の曖昧性はほとんど無くなる。したがって、十分に安定した解析結果が得られることが期待できる。また、この個別条件の列で記述することによって、作業者自身にとっても一文で記述するより各条件を認識しやすいというメリットもある。しかし、文の列全体で見れば指示代名詞相当表現による曖昧性が増加しているので、この指示代名詞相当表現の曖昧性を解消が次の課題となる。

4.3指示代名詞相当表現の処理

指示代名詞相当表現の係り先の特定は背景知識や文脈などに依存するため、文の解析結果だけでは大変難しい。しかし、管理作業のコーパスの分析の結果、係り先の文には以下の優先度で推定できることがわかった。

1)入力文の指示代名詞相当表現が属性名（属性値）を修飾している場合で、この属性名（属性値）を持つ文のうち、入力順で現在入力している文に最も近い文。

例文18：その品詞に動詞を持つ単語

→（「品詞」を使用した文で、入力順で現在入力している文に最も近い文）

2)入力文の指示代名詞相当表現が平均値やマクロ名などの予約語を修飾している場合で、この予約語を持つ文のうち、入力順で現在入力している文に最も近い文。

例文19：その平均値より小さい単語を教えて。

→（平均値を求めている文で、入力順で現在入力している文に最も近い文）

3)入力文が検索条件の場合で、条件を記述している文のうち、入力順で現在入力している文に最も近い文。

例文20：それは頻度が２以上です。

→（入力順で現在入力している文に最も近い条件を入力した文）

4)直前に入力した文

例文21：それを見出し語順に表示する。

（1)-4)の数字が小さいほど優先度が高い）

さらに、係り先の文との係り受け関係には以下に上げる４つのパターンに大別することができた。

1)前文にAND条件として加える。

例1：それは属性値を持つ。

例2：その属性名は属性値だ。

2)前文にOR条件として加える。

・該当文なし

3)前文の結果を条件とする。

例3：それと同じ属性名or属性値を持つ。

4)出力形式に関する操作。

例4：それに属性名を表示。

また、この指示代名詞相当表現と係り先との関係を分析するときに収集した、指示代名詞相当表現とその文の構造をもとに、例1～4で示すようなテンプレートを作成した。このテンプレートを使用して入力文を、これら４つのパターンのいずれかであるかを推定することができる。

この２つの推定手法から指示代名詞相当表現の係り先文と係り先との関係の推定を行うことができるが、完全な推定ができない以上最終結果の確認をユーザに求める必要がある。また、ユーザからの求めに応じこれらの関係を修正する機能も必要である。しかし、多くの場合正しく処理できるため、実際に係り先関係の修正を行う回数を少なくすることができる。

4.4曖昧性と省略補間

1)自然言語の表現で、その概念が異なるものであっても対象システムに対する要求では同じ概念になるものもある。

例）「読み」の品詞を検索する。

「読み」の品詞を教えて。

データベースに対しては見出し語が「読み」の行を検索し、その品詞を表示するという同じ要求である。

2)質問文の一部が省略されても一義に要求が推測できるものがある。

例）「読み」の品詞は。

例では、見出し語が「読み」の行を検索し、その品詞を表示することと推測できる。

3)システムに依存している要求がある。

例）読みが「あい」の見出し語を出力して。

例では、出力媒体がシステムに依存している。

"出力"の場合はシステムに接続されている媒体すべてがその対象となる。

"表示"の場合は、媒体が画面と推測できる。

"書く"の場合なら、ファイルと推測できる。

このような判断が一般的に言える。

4)データベースの情報に関してその規模（数の大小）を数値で入力することは難しく、曖昧な程度表現を使うことがある。

例）連接続性をたくさん持っている見出し語を表示して。

程度表現などは対応する数値の知識としてシステムに蓄える。

4.3管理要求のマクロ化

複雑な質問文は指示語や省略を用いても繁雑になってくる。一連の処理手順に名前を付け、同一処理の呼び出しや、さらに複雑な処理を記述する場合の部分処理にマクロ名を付けることにより、指示の簡素化が実現できる。

5.実現への方策

5.1システム構成

これらの手法をもとに、日本語による辞書管理

プログラム作成の支援システムを設計した。このシステム構造をシステム構成図1に示す。また、この構成図に添って処理の流れを説明する。始めに、ユーザから入力された入力文は解析部で解析し、意味構造を作成する。解析に曖昧性がある場合、複数の結果を作成する。必要に応じユーザが直接決定する機構も持つ。LQL生成部では、解析が作成した複数の結果から１つを選択し、内部形式のLQL構造（管理要求の意味表現：Logical Quarry Language）に変換する。さらにユーザにその解釈結果の正否を確認するための機構を持つ。また、入力文にある指示代名詞相当表現の認定を行なう。次に、指示代名詞処理部ではLQL生成部で認定された指示相当表現の参照先の曖昧性の解消を行なう。これを内部で保持し指示代名詞相当表現の曖昧性解消などの際に参照する。最終的にユーザからの指示で保持したLQL構造からSQLに変換し実行プログラムとしてファイルに出力する。また、作成したプログラムを直接本システムから実行し、動作の確認を行うことができるシステムとなっている。

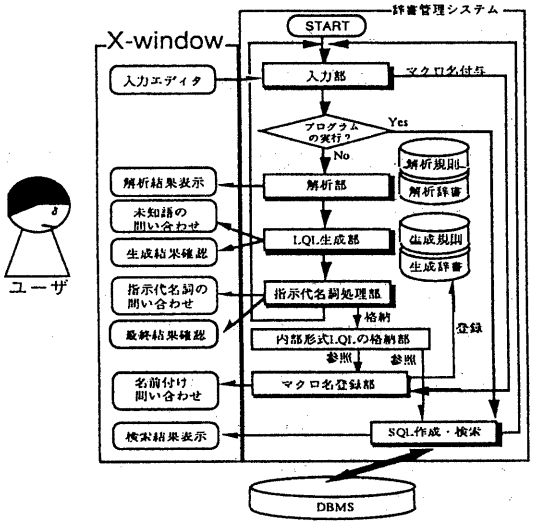

図1.システム構成

5.2日本語の解析機能

日本語の入力文は解析部をへてLQL生成部で検索言語に依存したLQL構造に変換される。このLQL構造には、論理式の構造や、データの持つ曖昧性や実際のデータと表現するための名称との関係、省略時のdefault値等の情報等を保持している。このLQL構造を使って、指示代名詞相当表現の係り先などの曖昧性の解消を行い、管理要求として格納する。以下に具体的な入力文より、どのようなLQL構造に変換し、実行プログラムとしてどのようなSQLとなるか示す。

○入力文

文1:品詞が形式名詞の単語を出して下さい。

文2:その中で概念見出しを持つものは。

○ユーザの確認後のLQL構造

```
文1:条件文1{
    表示[field(fld=hdwd,tbl=wd,type=char,spl=単語})],
    条件 like(field(fld=gram,tbl=wd,type=char,spl=品詞),
    value(fld=gram,tbl=wd,type=char,spl=形式名詞,
    pat[1] ="%;JN7;%"))
}
文2:指示代名詞文{and, 指示先=1, 単語=,
    isnotnull(field(fld=chwd,tbl=cdh,type=char,
    spl=概念見出し))}
```

文1では、表示fieldと条件を1つ保持する構造となっている。valueは実際に格納されている値と通常使用する属性値の関連を付けるための情報をもつ。文2では、概念見出しについての条件の追加を行っている。ここでは、係り先の情報の中にその追加対象の条件を保持する形式となっている。

○SQLプログラム

```
select distinct wd.hdwd from wd, cdh
    where   wd.gram like '%;JN7;%'
      and (cdh.chwd is not null)
      and  wd.cid = cdh.cid ;
```

[1] 生成辞書から得られる情報で、属性値に対する実際のデータベースに格納されている値（コード）を示す。

5.3マクロ機能

マクロ名の登録は、マクロ名登録部で行い、ユーザからマクロ名を得た後、対象となるLQL構造を生成辞書等に登録している。マクロ名引用時には、LQL生成部で登録したLQL構造を取り出して処理可能としている。

5.4管理作業の内容を出力する機能

入力した管理作業内容をシステムはLQL構造として内部で保持し、これからプログラムへの変換を行う。変換するプログラムはデータベースの操作言語であるSQLを生成する。変換時には、データベーススキーマを用いて条件の補填やデータ間の整合性を取る操作を行う。条件の補填には、データベースのテーブル間のリンクを取るための条件や、副問い合わせなどで使うキーとなる属性名の整合性を取る等の処理を行い、エラーのないプログラムを生成する。

6.インタフェースの実行例

日本電子化辞書研究所（以下EDR）辞書を開発・改良する上で人手による、作業が不可欠である。このため、作業に必要な情報収集や、作業後の整合性の修正と行った管理作業が常に発生する。ここではこのような開発・改良作業に必要な情報収集を例に上げる。この情報収集の対象となるのは、EDRで開発している辞書には単語の読みや品詞等の情報を持った単語辞書や概念を表している単語の持つ個別の意味とそれらの関係等の情報を持った概念辞書等がある[5][6]。単語辞書には、見出し語、読み、連接属性、文法属性、概念IDの項目がある。単語は見出し語、読み、概念の各々の違いにおいて区別される。例えば見出し語と読みが同じであっても概念が異なると別単語としてみなされる。また、概念辞書にはその概念を表す概念見出しと概念IDがある。単語辞書と概念辞書は、両辞書の項目に存在する概念IDによって、関連付けされる。以下に、前述の機能を組み込んだ試作システムの実行例を示す。

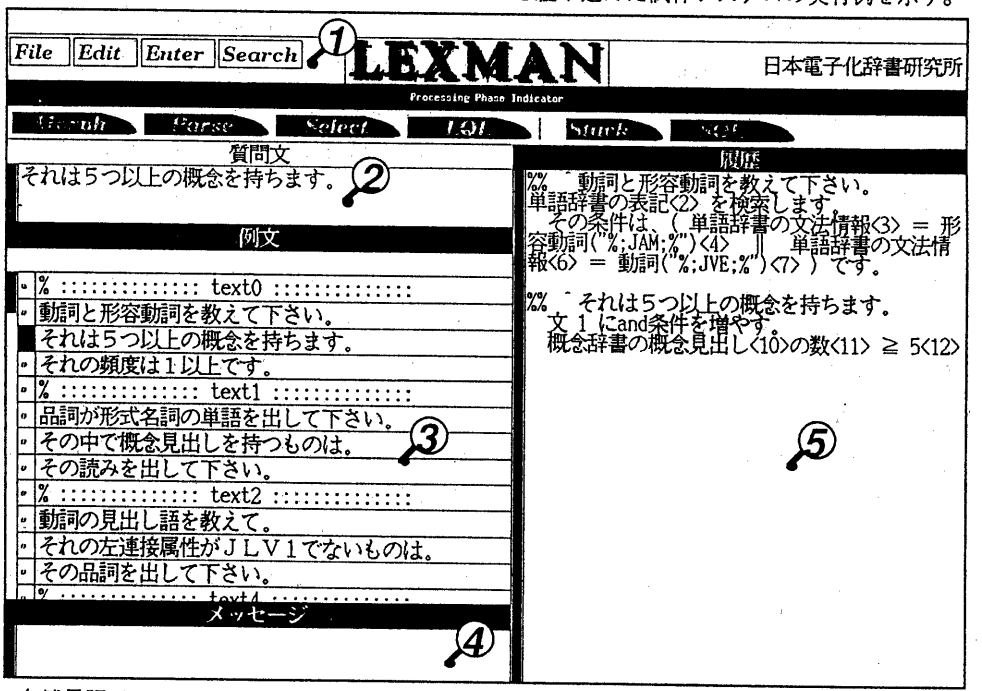

図2　自然言語インタフェース操作ウィンドウ

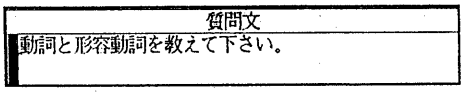

図３　質問文

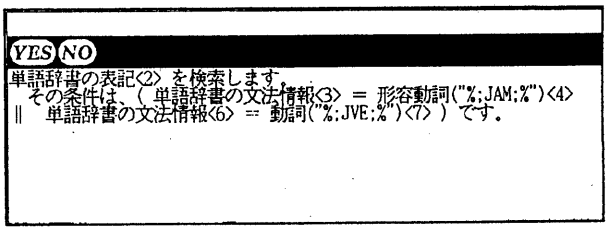

図４　LQL確認１

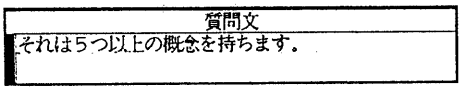

図５　質問文２

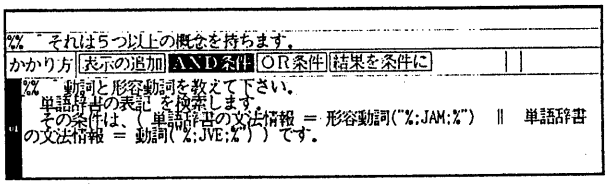

図６　係り先確認

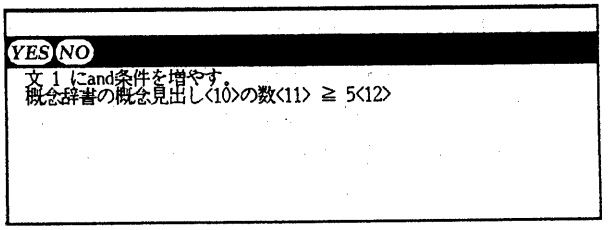

図７　LQL確認２

操作ウィンドウ（図２）は、次の5つのウィンドウから構成されている。

1）コマンドウインドウ
システムの終了や解析の開始、DBMSへの検索の指示などを行う。

2）入力ウインドウ
管理要求を日本語を用いて入力するウインドウ。

3）例文表示ウィンドウ
デモンストレーションを効率よく行うためのウィンドウ。

4）メッセージウィンドウ
システムからユーザに提示するメッセージを表示するウィンドウ。

5）履歴ウィンドウ
入力した管理要求の履歴をLQL構造で表示するウィンドウ。

この操作ウィンドウの入力ウインドウから管理要求の文を入力する(図３)。まずはじめにユーザは入力ウィンドウに質問文を直接入力するか、例文ウインドウより質問文を選択し入力する。質問文の入力後、Enterボタンを押すことにより、入力された質問文は解析部を経てLQL生成部でLQL構造に変換される。そして、ユーザに対して自然言語に近いかたちで質問文の解釈の確認がおこなわれる（図４）。

確認の結果内容に問題がない場合、「YES」のボタンを押すことにより一文の入力が終了し履歴ウインドウに入力した文が表示され最初の操作ウインドウの状態に戻る。さらに条件を追加するために続けて入力ウインドウに条件文を入力する。(図5)次の質問文は質問文１と同様に解析後、LQLに変換され、ユーザに対して指示語の係先ならびに条件項目の選択についての確認がおこなわれる（図６）。

さらに文１と同様に質問文の問い合わせが行われ（図７）、確認の後履歴ウィンドウにそのＬＱＬが追加される。以上で質問文の入力が終わるとコマンドウィンドウのSearchボタンをクリックすることにより、システム内で保持されていた質問文のLQLがSQLに変換される。この時、この処理全体にマクロ名を付けることができる。（図８）そして、データベースに対してSQLが発行される（図９）。

検索ウィンドウ（図９）には、履歴ウィンドウに表示されていたＬＱＬ構造とデータベースシステムで実行するＳＱＬプログラムと検索結果が表示される。検索中に次の管理要求を平行して入力することのできるシステムとなっている。

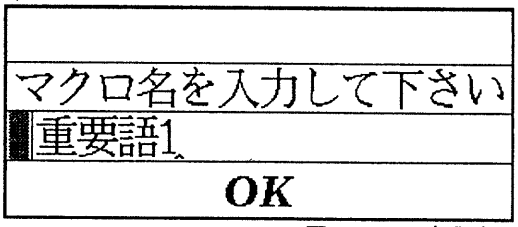

図８マクロ名入力

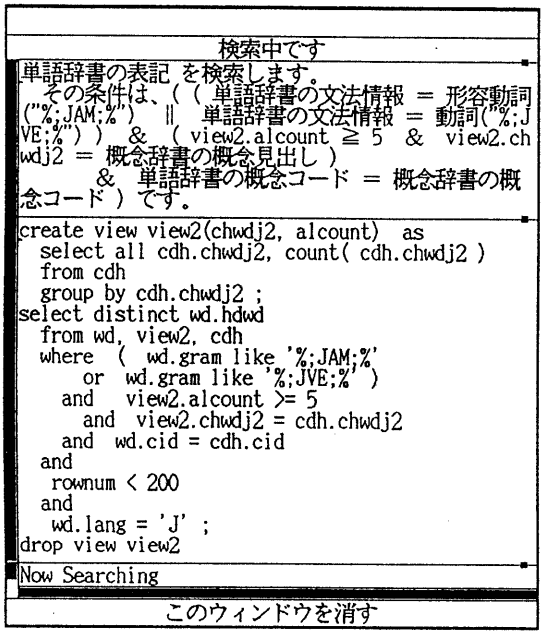

図９　検索ウィンドウ

6.おわりに

本システムの解析では基本文型を解析出来うる文法の他、用言の省略など実際の入力に則した文法も組み込んでいる。生成では、副問い合わせや数量表現などにも対応し多様な条件を処理することが出来る。これにより、実作業を元に収集した数十種類の管理要求に対しプログラム作成を行なうことが出来るシステムとなっている。このシステムにより自由度の高い問い合わせ文の入力が可能となった。今後は、本システムの有用性を評価すると共に、ユーザの操作性を考え、使用頻度が高く曖昧性の少ない操作に関してはメニューシステム等の手法を取り入れた拡張を行なっていく。

[参考文献]

[1]Zoolf, M.M.: Query by Example, Proc. AFIPS National Computer Conf., pp. 19-22, May 1975

[2]芝野耕司:データベース言語SQL, 情報処理, Vol.29, No.3, pp.208-214, 19988年3月

[3]上林, 天野:擬自然言語データベースインタフェースにおける質問作成補助システム,情報処理学会データベース・システム研究会報告69-3

[4]谷　幹也:自然言語インタフェースによるユーザインタラクション,情処43回全大論文集1H-1

[5]EDR電子化辞書： EDR,TR-016 1989

[6]EDR電子化辞書： EDR, TR-018 1989