TR-I-0153

言語データベースADDの検索システム
A Retrieval System for The ADD (ATR's Linguistic Database)

橋本一男　小倉健太郎†　森元 逞
Kazuo Hashimoto　Kentaro Ogura　Tsuyoshi Morimoto

1990.3

概要

本稿では、言語データベースADD(ATR Dialogue Databe)の検索システムを紹介する。本システムは、ADDに納めされている100以上に及ぶ言語データの中から必要とする情報を効率的に抽出し、自動翻訳技術の研究者に提供することを目的としたシステムである。現在、ATRのホスト計算機ATR-DP上で稼働しており、所内の研究者サイトのダム端末から手軽に利用することができる。

ATR 自動翻訳電話研究所
ATR Interpreting Telephony Research Laboratories

† 現在、NTT情報通信処理研究所

目次

1　はじめに

本稿では自動翻訳技術研究用言語データベースADD(ATR Dialogue Databe)の検索システムを紹介する。ADDには、日本語対話テキストの形態素データを中心として係受けデータや日英対訳対応データなどの多くの言語データが蓄積されている[1)2)]。さらに、これらを組み合せて用いることができるように、言語データ間に様々な観点から関連付けを行っている。

ADD検索システムは、利用者が問い合わせ言語DEFSEARCHを用いて言語データ自身の属性や言語データ間の関連を宣言的に記述するだけで、目的とする言語現象を得られる機能を有するシステムである[3)]。本システムを用いることで、利用者は従来の言語データベースではなしえなかった多様で複雑な言語現象の抽出を効率的に行うことができる。

近年、MMI(マンマシンインタフェース)としては、操作性を重視したオブジェクト指向型システムが開発されてきている。筆者らも本システム開発当初、高機能ワークステーション(Symbolics3600)を利用したオブジェクト指向MMIシステムのプロトタイピングを行い[4)]、多くの知見を得た。DEFSEARCHはその成果である。一方、ADDは今後、ATR以外の研究機関に開放する予定であり、ADDを簡便に使えるような環境を考える必要がある。高機能ワークステーションは低価格化が進んではいるものの、依然、高額で一般的には導入しにくいモノである。そこで本システムでは、ダム端末(具体的にはVT282端末)をMMIのターゲット機として、プロトタイピングで得たオブジェクト指向(風)MMIを実現している。

以下、2章で検索システムの概要を述べる。3章では検索問合せ言語DEFSEARCHを説明し、4章でMMIについて述べる。

2 概要

2.1 ADD

ADDは関係データベースを用いて実現している。具体的には19のテーブルと158のフィールドで構成する。各テーブルのレコードは、フィールド値を介して様々な観点から多重に関連付けてある。図1は、ADDのデータ構成を表している。実線の箱が各テーブルのレコードを表し、点線囲み内の項目がレコード自身の属性を表す(実体属性)す。下線付きの項目がキーである。また矢印は、他のレコードとの関係を示す属性(関連属性)を表している。

例えば「単語レコード」は、実体属性として「出現単語」や「品詞」を持つ。また関連属性としては、直前、直後の単語レコード(の単語ID)を示す「P_IS」と「N_IS」や、上位の文節レコードを示す「P_OF」、対応する英語単語との関連を付ける「単語対応」、他の単語との係受け関係を示す「係受け」などを持つ。「係受け」は1つの単語に複数の係受けが付く可能性があるので、多値属性であり、これは2重の矢印で表している。

これらの情報を組み合せて用いれば、例えば次のような検索が可能になる。

(1) 単純な検索
- 単語ID、表記、読み、標準表現、品詞を求める。
- 固有名詞を頻度付きで求める。

(2) 発話構成要素間の全体-部分関係を用いた検索
- 「教える」という語を含む文を求める。

(3) 発話構成要素間の時間的順序関係を用いた検索
- 助詞「の」を含む文節とその次の文節を求める。

(4) 係受け関係を用いた検索
- 「する」と係受け関係にある単語を意味カテゴリとともに求める。

(5) 日英対訳対応関係を用いた検索
- 文を対応する英文とともに求める。

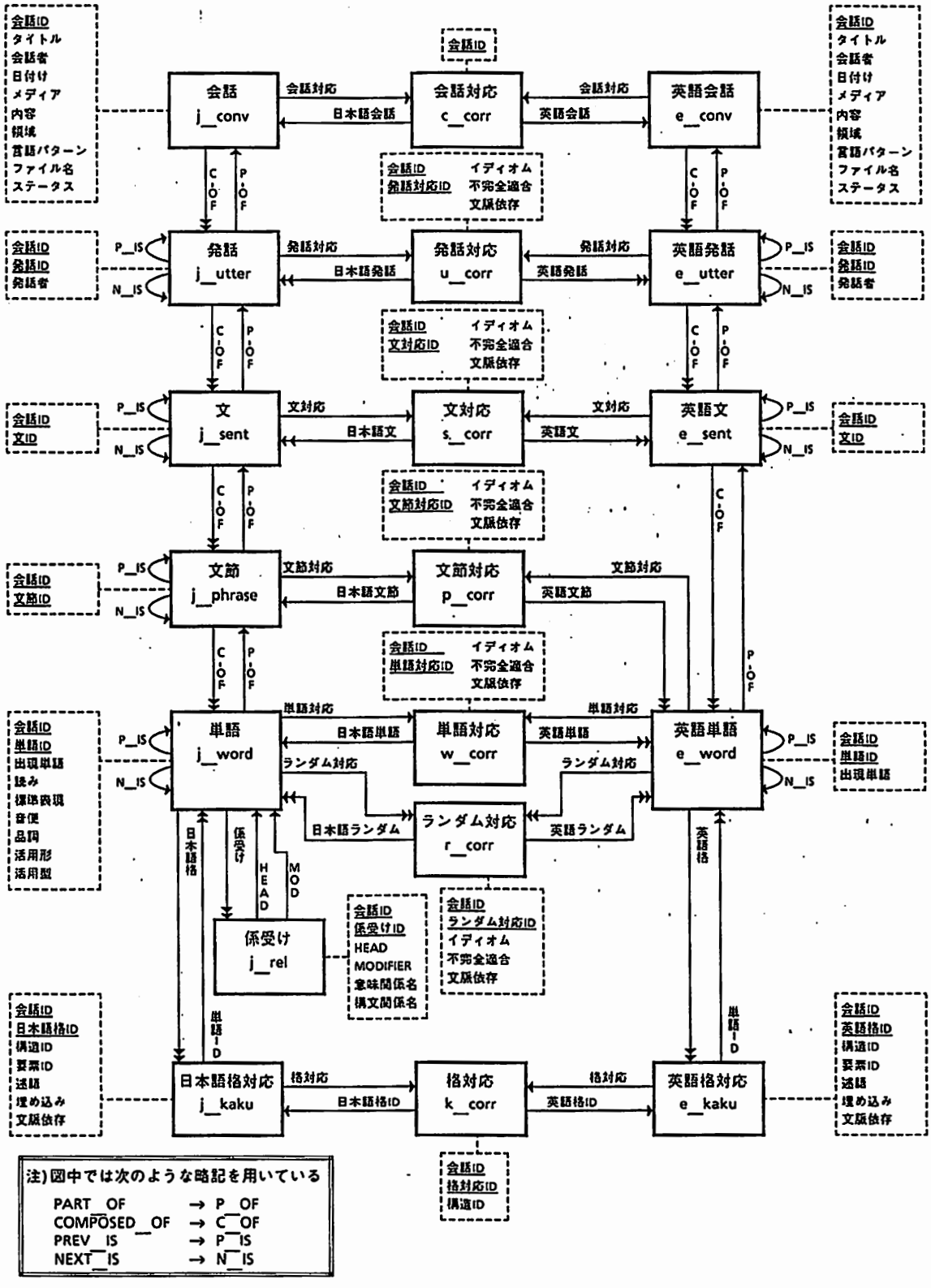
会話ID
タイトル
会話者
日付け
メディア
内容
領域
言語パターン
ファイル名
ステータス
会話
j__conv
会話対応
日本語会話
会話対応
c__corr
英語会話
会話対応
英語会話
e__conv
C-OF
P-OF
会話ID
発話ID
発話者
P__IS
N__IS
発話
j__utter
発話対応
日本語発話
発話対応
u__corr
英語発話
発話対応
英語発話
e__utter
会話ID
文ID
文
j__sent
文対応
日本語文
文対応
s__corr
英語文
文対応
英語文
e__sent
会話ID
文節ID
文節
j__phrase
文節対応
日本語文節
文節対応
p__corr
英語文節
文節対応
英語文節
会話ID
単語ID
出現単語
読み
標準表現
音便
品詞
活用形
活用型
単語
j__word
単語対応
日本語単語
単語対応
w__corr
英語単語
単語対応
英語単語
e__word
ランダム対応
日本語ランダム
ランダム対応
r__corr
英語ランダム
ランダム対応
英語ランダム
日本語格
係受け
HEAD
MOD
係受け
j__rel
会話ID
係受けID
HEAD
MODIFIER
意味関係名
構文関係名
会話ID
ランダム対応ID
イディオム
不完全適合
文脈依存
単語-ID
会話ID
日本語格ID
構造ID
要素ID
述語
埋め込み
文脈依存
日本語格対応
j__kaku
格対応
日本語格ID
格対応
k__corr
英語格ID
格対応
英語格ID
e__kaku
会話ID
格対応ID
構造ID
注)図中では次のような略記を用いている
PART__OF → P__OF
COMPOSED__OF → C__OF
PREV__IS → P__IS
NEXT__IS → N__IS

図1　ADDの構成[2)]

2.2 システム構成

本システムは図2に示すように、ホスト計算機VAX8820を用いた集中処理型システムで、Ultrix(DEC社のUNIX)環境下で稼働する。全体として、次の4つのソフトウェアで構成される。

(1) データ管理部UNIFY

ADDの管理にはRDBMS(関係データベース管理システム)UNIFYを用いる[5]。

(2) MMI(マンマシンインタフェース)部LDBSH

ADDに対する非定型検索操作のインタフェースを提供する。主な機能として検索範囲の指定、検索問合せの作成、検索の起動がある。検索問合せの作成には、本システム専用のデータ操作言語DEFSEARCHを用いるが、利用者には必要最小限の記述ですませることのできるようにパターンエディタを提供する。

また、UNIFYは独自の利用者インタフェースとして対話型SQLを持つが、本システムではLDBSHを通してこれを利用することができる。

(3) 検索処理部LDBMSD4

LDBSH上で作成された検索要求を解析し、その検索処理と加工処理を行う。一般的にSQLでは、一つのテーブルに対する選択や射影の操作は高速に処理されるが、他のテーブルの要素も合わせた検索処理の場合、テーブルの結合やSQLのネストが多発するため、実用的な性能が得られないことがある。LDBMSD4は、LDBSHから検索問合せを受け取ると解析を行い、UNIFYが高速に処理可能な検索問合せに分割して渡す。UNIFYからの結果が返されると、利用者が指定した出力形式を満足する形に加工し、必要ならば再び検索問合せを生成してUNIFYに渡す。例えば単語条件の記述から英語文を検索する場合は、

単語テーブル-(part-of)→文節テーブル-(part-of)→文テーブル-(文対応)→
文対応テーブル-(英語文)→英文テーブル-(composed-of)→英単語テーブル

というパスで順次、処理を行う。

(4) 統計情報システム

言語現象に関する定型的な統計情報を定期的にADDから抽出しファイリングする。利用者はUNIXのcatやmoreなどのコマンドでファイルを閲覧することができる。本稿では統計情報システムについては述べない。

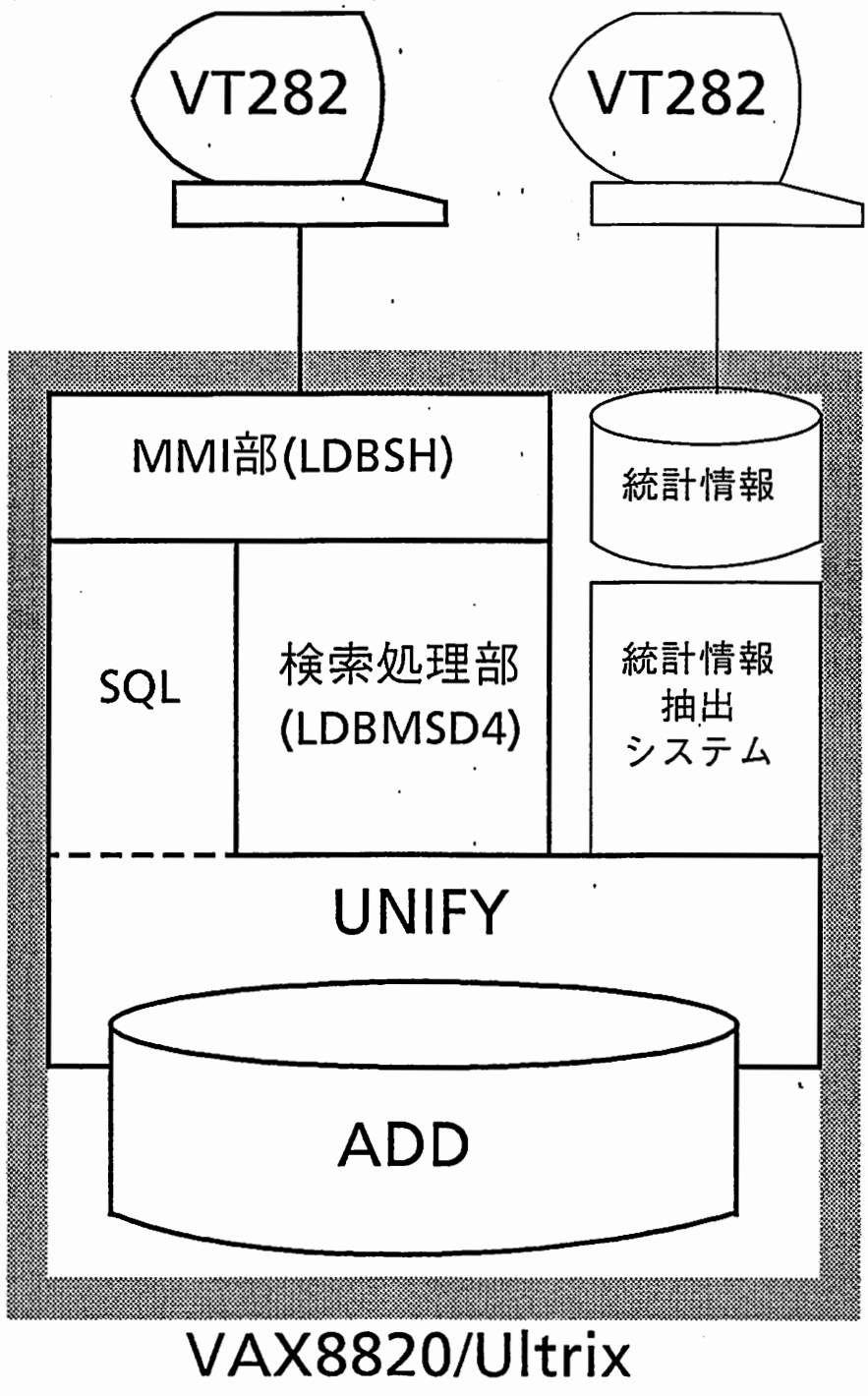

図2　ADD検索システムの構成

3 問合せ言語DEFSEARCH

3.1 概要

ADDの検索機能として第一に要求されるのは、処理の即時性よりも検索要求の多様性である。特に、特定の言語データに着目して、それに関連する他の言語データも合わせて観察できるような検索機能が要求される。本システムでは、これを満足するために、次の形式の問合せ言語DEFSEARCHを提供する。以後、DEFSEARCHを用いて記述された検索問合せを検索パターンと呼ぶ。

(DEFSEARCH　<検索パターン名>
　<言語データ記述>
　<出力形式>)

<検索要求名>は、利用者が任意につける名称である。
<言語データ記述>は検索の対象となる言語データの仕様記述部である。
<出力形式>は、検索した言語データの出力仕様記述部である。

<言語データ記述>::=<オブジェクト一記述>+
<オブジェクト一記述>::=
　　(OBJECT-PATTERN　<オブジェクト変数>
　　((INSTANCE-OF　テーブル名)
　　<スロット条件>))

<出力形式>::=(RETURN　<オブジェクト変数> +)
ただし、+は0回以上の繰り返しを表す

テーブル名には対象とするデータのテーブル名を、<スロット条件>には実体属性や関連属性についてデータが満たすべき条件を指定する。<オブジェクト変数>は検索パターン内でデータを識別するための仮称であり、データ間の関連条件や出力形式の指定に用いる。検索パターンの具体例を3.2、3.3、3.4で示す。

3.2 単純な属性検索

単語テーブルから「品詞＝助動詞」のデータを選択し、その表記、標準表現、活用形を返す。

検索パターン

```
(OBJECT-PATTERN  (VAR  単語1)
     ((INSTANCE-OF     単語)
     (表記             (VAR  x)
     (品詞             助動詞)
     (標準表現         (VAR  y))
     (活用形           (VAR  z))))
(RETURN  (VAR  x)  (VAR  y)  (VAR  z))
```

結果

```
1.   ます      ます      連体
2.   れ        れる      連用
3.   ます      ます      終止
4.   た        た        連体
                :
```

3.3 全体･部分関係、時間的順序関係を用いた検索

･ 会話テキストの分析では、ある特定の要素を含む上位の要素を見ることや、同一レベル要素の連鎖を見ることが多い。そのための検索では、PART-OF、COMPOSED-OFという全体･部分関係、PREV-IS、NEXT-ISという時間的順序関係を用いる。下例では「名詞＋を＋する」という単語連鎖に対し、単語の時間的順序関係、文節、文との全体･部分関係を示し、RETURNに(VAR 文1)を与えることで、文レベル表記の表示を行う(図3)。このようにDEFSEARCHでは、品詞列や文字列といった同一属性による連鎖だけでなく、品詞と標準表現の組合せのような異なる属性の連鎖も記述できる。

「名詞＋を＋する」という単語を持つ文を検索する検索パターン

```
(OBJECT-PATTERN  (VAR  単語1)
     ((INSTANCE-OF     単語)
     (NEXT-IS          (VAR  単語2))
     (品詞             名詞)))
(OBJECT-PATTERN  (VAR  単語2)
     ((INSTANCE-OF     単語)
     (NEXT-IS          (VAR  単語3))
     (PREV-IS          (VAR  単語1))
     (標準表現         を)))
(OBJECT-PATTERN  (VAR  単語3)
     ((INSTANCE-OF     単語)
     (PART-OF          (VAR  文節1)
     (NEXT-IS          (VAR  単語3))
     (標準表現         する)))
(OBJECT-PATTERN  (VAR  文節1)
     ((INSTANCE-OF     文節)
     (PART-OF          (VAR  文1)
     (COMPOSED-OF   *  (VAR  単語3)  *)))
(OBJECT-PATTERN  (VAR  文1)
     ((INSTANCE-OF     文)
     (COMPOSED-OF   *  (VAR  文節1)  *)))
(RETURN  (VAR  文1))
```

ただし、*は0以上のワイルドカード

結果

1. ....という情報誌の記者をしております....
2. ....の取材をするようにと....
3. ....活動をして頂きたいと思います.
4. ....工学の教授をしております....

:

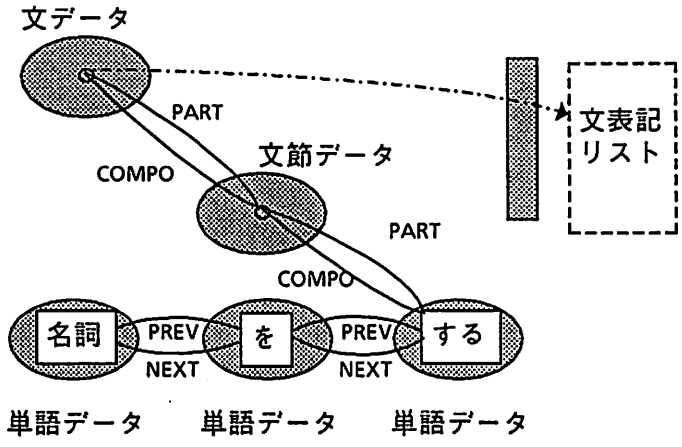

図3　3.3の検索パターンの概念

3.4　係受け関係を用いた検索

文節間の係受けや動詞の格関係は文の意味構造を捉える上で重要である。言語データベースでは便宜上、文節を代表する単語間に係受けがふられているが、全体･部分関係を用いることで本例のような検索が可能になる(図4)。

「する」に関する係受けの検索パターン

```
(OBJECT-PATTERN   (VAR   係受け1)
      ((INSTANCE-OF          係受け)
       (MODIFICANT           (VAR   単語1)
       (MODIFIER             (VAR   単語2))
       (意味関係名           (VAR   x))))
(OBJECT-PATTERN   (VAR   単語1)
      ((INSTANCE-OF          単語)
       (PART-OF     (VAR   文節1)
       (標準表現    する)
       (係受け      *   (VAR   係受け1)   *)))
(OBJECT-PATTERN   (VAR   文節1)
      ((INSTANCE-OF          文節)
       (COMPOSED-OF   *   (VAR   単語1)   *)))
(OBJECT-PATTERN   (VAR   単語2)
      ((INSTANCE-OF          単語)
       (PART-OF     (VAR   文節2)
       (係受け      *   (VAR   係受け1)   *)))
(OBJECT-PATTERN   (VAR   文節2)
      ((INSTANCE-OF          文節)
```

```
        (COMPOSED-OF  *  (VAR  単語2)  *)))
    (RETURN  (VAR  x)  (VAR  文節2)  (VAR  文節1))
```

結果

1.	対象	記者を	しておりますけれども.
2.	対象	取材を	するようにと
3.	条件	今度	するようにと
4.	意味並列	取り、	して頂きたいと

:

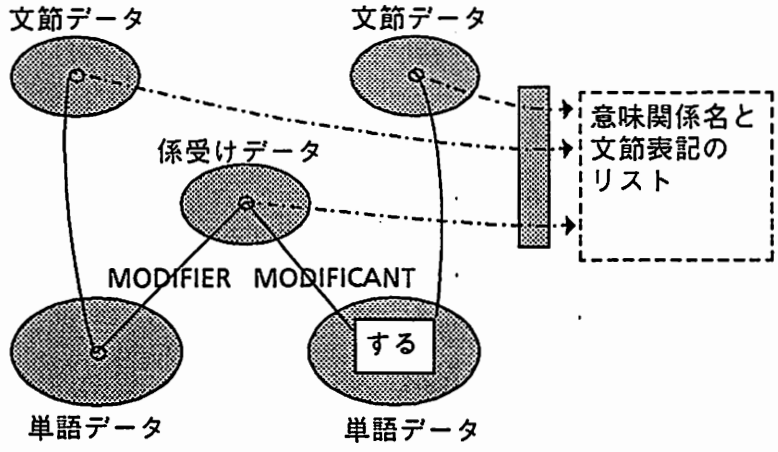

図4　3.4の検索パターンの概念

4 MMI

4.1 概要

ADD検索システムのMMI機能は、そのサブシステムであるLDBSHが果たしている。LDBSHはシェルの一種であり、操作は表1に示すコマンドを用いて行う。また図5の例のように、UNIXコマンドを!でエスケープしての実行や、リダレクション(>と>>)を用いた検索結果のファイル格納、検索のバックグラウンド実行指定が可能である。典型的な操作の流れは次のようになる。

(1) 起動(ldbsh)
(2) 検索対象の選択(target)
(3) 検索結果の出力形式の指定(mode)　　※不要の場合もある
(4) 検索パターンの作成･編集と実行(editpとpsearch)
　　あるいは単語条件列検索の実行(wsearch)
(5) 検索結果の評価(UNIXコマンド)
(6) 必要に応じて(2)以降を繰り返す
(7) 終了(end)

端末としては簡便性を重視し、VT282を想定しているが、エミュレーションができれば機種は問わない(例えばPC-9801でも可)。

以下では主要コマンドとしてtargetとeditpを解説する。

```
ldbsh:  psearch  patfile  patname  >  result  &
検索(result)が終了しました。
ldbsh:  !cat  result
1.  はい、登録用紙は既にお持ちでしょうか。
2.  それでは、こちらからお送りしますので、お名前とご住所を‥‥
3.  参加料はいるのでしょうか。
          ⋮
ldbsh:  end
%
```

図5　操作例

表1　コマンド一覧

コマンド名	機能
ldbsh	起動(B/Cシェルから)。
editp	検索パターンの作成、編集を行う。
editw	一時メモリにある最新の単語条件列を編集する。
end	終了。
help	コマンド、テーブル、フィールド名などの表示。
mode	検索結果の出力形式を指定する。ただし、検索パターンで RETURN-FORM句を使用したときのみ、有効である。
psearch	ファイルに格納された検索パターンを実行する。
restartw	一時メモリにある最新の単語条件列を実行する。
savew	一時メモリにある最新の単語条件列をファイルに格納する。
sql	UNIFY提供のSQLを起動する。
startw	ファイルに格納された単語条件列を実行する。
status	検索実行状態を報告する。
target	検索対象の選択を行う。
wsearch	引数として与えられた単語条件列を実行する。

4.2 検索対象の選択(targetコマンド)

検索の実行に先立ち、検索対象の会話を選択する。選択対象は、次の選択操作が行われるまでは保持される。選択には会話属性指定と会話番号指定の2通りがある。

(1) 会話属性指定

会話属性指定では、会話メディア、領域、内容、言語パターンという4つの会話属性値を指定の結果、該当した会話を検索対象とする。操作は会話属性選択画面上で行う(図6)。

```
ldbsh:  target
範囲指定方法を選択して下さい。
 0. 検索対象表示
 1. 会話属性指定
 2. 会話番号指定(初期化)
 3. 会話番号指定(前回の継続)
 4. 会話番号指定(属性指定の結果を使用)
 ==>1
```

図6 会話属性指定の開始

矢印キーでカーソルを動かし、リターンキーで属性値を選択すると、該当する会話数が画面右上に表示される。

```
                    <<会話属性選択>>
                                  対象会話数: 28

<属性名>          <属性>
メディア          :キーボード  電話
領域              :国際会議
内容              :問合せ  依頼･説得  相談･交渉
言語パターン      :英:日  日:英  日:日  英:日(同)
```

図7 会話属性選択画面

qを押すと、選択が終了し、対象会話が確定する。

```
対象会話:3045 3046.3048.3049.3050.…
ldbsh:
```

図8 会話属性指定の終了

(2)　会話番号指定

会話番号指定は検索する会話の番号を直接選択する方法であり、①新たな番号指定、②前回の番号指定の修正、③前回の属性指定の修正、の3通りがある(図9)。操作は会話番号選択画面上で行う。

```
ldbsh:  target
範囲指定方法を選択して下さい。
 0. 検索対象表示
 1. 会話属性指定
 2. 会話番号指定(初期化)
 3. 会話番号指定(前回の継続)
 4. 会話番号指定(属性指定の結果を使用)
 = = > 2
```

図9　会話番号指定の開始

会話番号選択画面が現れたら、矢印キーでカーソルを移動し、リターンキーで番号を選択する(選択された番号は反転表示される)(図10)。ページ送りには<ctrl>+vと<esc>+Vを用いる。

<< 会話番号選択 >>

会話ID	発話数	文数	単語数
81	14	29	662
82	15	47	1204
83	10	28	575
84	17	48	1191
85	21	47	1259
86	9	26	727
87	17	96	3689
88	11	64	2443
89	12	55	2163

図10　会話番号選択画面

qを押すと、選択が終了し、対象会話が確定する(図11)。

```
対象会話:3045 3046.3048.3049.3050.…
ldbsh:
```

図11　会話番号指定の開始

4.3 検索パターンの作成(editpコマンド)

検索パターンの作成にはパターンエディタを用いる。パターンエディタは、オブジェクトウィンドウと呼ばれる表に記入する形式で、検索パターンを作成できる簡易ツールである。その基本的な仕様は、利用者がオブジェクトウィンドウの適当な場所に検索条件を書き込むことによってDEFSEARCH検索パターンを生成するというものである。

たとえば、「固有名詞」を求めるという問合せを考えることにする。まず、editpコマンドを入力すると図12のようなメッセージを得る。次に問合せに対する答えが単語テーブルの中で求められることを知っているとして、「単語」と入力すると、図13のようなメインウィンドウが表示される。ここでカーソルを最下段の$<単語1>行に移動し、eキーを押すことによって単語テーブルのフィールドが表示されたオブジェクトウィンドウが開かれる(図14)。利用者はここで図#のように記述することで、その問合せの検索パターンを作成することができる。

```
ldbsh:  editp test
ファイル testが存在しません。
新規作成で継続しますか?(Y/N)?y
ベースのテーブル名を入れて下さい[デフォルト:単語]:
```

図12　パターンエディタの起動

```
        <<メインウインドウ>>

パターン名        :
コメント          :
リターン形式      :[レベル] 文
-------------------------------------------------
単語              :$<単語1>
```

図13　パターンエディタのメインウィンドウ

```
        <<オブジェクトウインドウ>>

パターン名        :EXAMPLE1
コメント          :固有名詞
リターン形式      :[変数並び]$<単語1>
-------------------------------------------------
単語              :$<単語1>
AFTER             :
BEFORE            :
文字数            :
PART-OF           :
PREV-IS           :
NEXT-IS           :
出現単語          :
読み              :
標準表現          :
音便              :
品詞              :固有名詞
活用形            :
活用型            :
係受け            :
単語対応          :
ランダム対応      :
日本語格          :
```

図14　パターンエディタのオブジェクトウィンドウ

1つのオブジェクトウィンドウは1つのオブジェクトパターンを表す。オブジェクトパターン間の関連を設定するような検索では、必要なだけのオブジェクトウィンドウを開き、それらの間の関連を各オブジェクトウィンドウの適当な場所に記入する。たとえば3.3で用いた例題「名詞＋を＋する」という検索パターンを作成する場合、下図のように5つのオブジェクトウィンドウを開いて記述を行う。

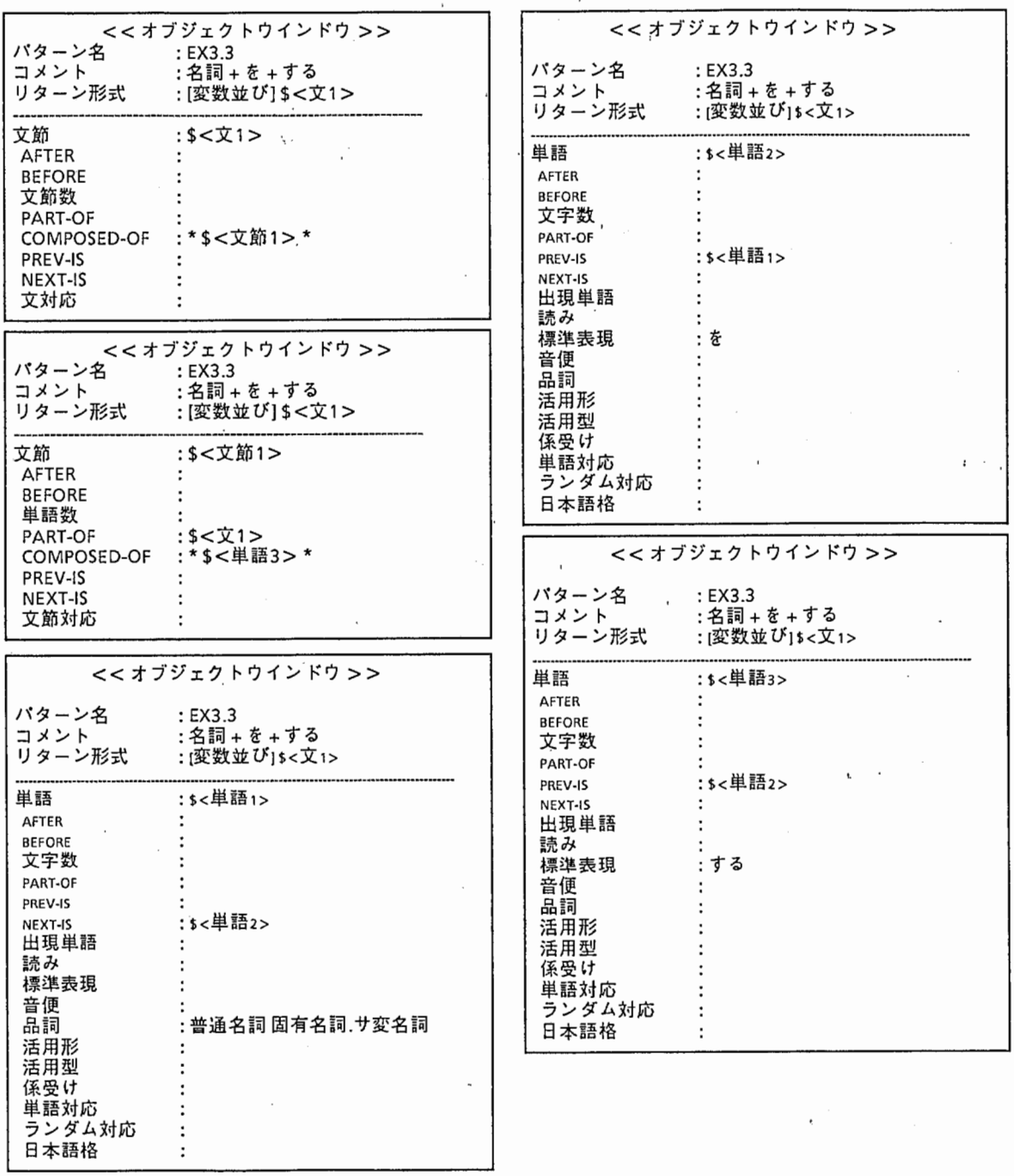

図15 「名詞＋を＋する」という検索パターン作成のためのウィンドウ

5　おわりに

言語データベースADDの検索システムについて述べた。本システムを用いることにより、利用者はダム端末からADDを効率的に利用することができる。利用方法の詳細については文献6)を参照されたい。今後、本システムが広く利用され、有意義な研究が数多く行われることを期待する。

謝辞

本研究の機会を与えて下さいました樽松明社長に感謝いたします。データ処理研究室および言語処理研究室の皆さまには数多くのご討論をいただきました。特に保坂順子研究員、長谷川敏郎研究員には操作性に関して貴重なコメントを多々いただきました。あわせて感謝いたします。最後に、本システムは(株)応用技術殿のご協力により実現したことを付記します。

参考文献

1) 江原、小倉、森元(1990)「電話対話データベースの構築」、情報処理学会第40回全国大会6F-1
2) 橋本、小倉、江原、森元(1990)「自動翻訳電話研究用データベースの構成」、ATRテクニカルレポートTR-I-0150
3) 橋本、小倉、森元(1989)「フレーム表現による検索機能を有する言語データベース管理システム」、情報処理学会アドバンスト・データベース・シンポジウム
4) 橋本、小倉、森元(1988)「言語データベース統合管理システムのマンマシンインタフェース」、情報処理学会第37回全国大会2C-4
5) UNIFY Release4.0 Reference MANUAL
6) 橋本(1990)「ADD検索システム利用解説書」、ATRテクニカルレポートTR-I-0154