TR-I-0150

自動翻訳電話研究用言語データベースの構成
ATR's Linguistic Database Structure

橋本一男　小倉健太郎†　江原暉将　森元 逞
Kazuo Hashimoto　Kentaro Ogura　Terumasa Ehara
Tsuyoshi Morimoto

1990.3

概要

本稿では自動翻訳電話研究用言語データベースADDのデータ構成について解説する。ADDは電話対話のテキストを、19のテーブルと158のフィールドで表現する構造化テキストデータベースである。ここではこれらのテーブルやフィールドの表す意味や特徴について解説を加え、また付録としてADDの詳細なスキーマを添付する。

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、自動翻訳電話研究用の言語データベースADD(ATR Dialogue Database)のデータ構成について述べる。自然言語処理の研究を行う上で、基礎データとなる言語データを収集し整備することの重要性は今さら問うまでもなく、これまでにも多くの物が報告されている。しかし、対象が話し言葉となると、データ収集からすでに多大な労力を要するためか、本格的なものは国立国語研究所のものが唯一であった。これは、会話の内容を広く日常会話中心にとっているため、必ずしも計算言語学的分析に適当ではない。また、電子化辞書研究所でも大規模な言語データベースを作成しているが、これは書き言葉が対象である。

ATRでは、自動翻訳電話の研究資料とすることを目的として、1987年より対話テキストを対象とした言語データベースの構築を進めてきた。本データベースは最終的には収録語数100万語以上の規模となるが、すでに約15万語が利用可能となっている(1990年3月末日現在)。

ADDは従来の単純構造のテキストベースではなく、言語現象を多様な角度から分析できるよう高度に構造化してある。したがって、本データベースを効果的に利用するには、そのデータ構成を把握しておくことが望ましい。そのため本稿では、まず2と3でADDの概要とデータ構成の基本的な考え方について述べる。次いで4から8までで各テーブルの意味や意義そして他のテーブルとの関連について説明する。最後に9で格納についての注意点を述べる。また、付録としてADDのスキーマとコード表を掲載する。

2 ADDの概要

2-1 対象データ

対象とするのは「国際会議や旅行についての問合せや申込み」といった、情報の伝達や行為の要求などの明確な目的を持ってなされる対話のテキストである。ただし、実際の対話を収集することは、通信の守秘という法律上の問題があるので、模擬対話によってテキストを収集している。また、日英間の翻訳技術を研究するため、和文と英文の両方のテキストを収集している。

ADDには、これらのテキストを直接格納するのではなく、表1に示すような作業を事前に行い、そこから得られるデータを格納している(図1)。事前に行う作業の詳細については、別稿で報告する。

表1

作業名	作業内容	得られるデータ
収録	電話あるいはキーボードを用いて、模擬的に行う会話を収録する。英文テキストについては、通訳を用いた同時収録と事後翻訳の2方式を併用する。	模擬会話の設定状況データ 和文テキストデータ 英文テキストデータ
形態素解析 文節組立て	まず、テキストを形態素分析し、形態素単位に属性データを付与する。次いで、形態素から文節単位を組み立てる。	形態素属性データ 文節データ
係受け解析	形態素間の係受け関係を洗い出す。	係受けデータ
日英対訳対応	発話、文、文節、単語(形態素)の各単位毎に、日本語と英語の対訳対応付けを行う。また、テキストの意味的なまとまりを格という単位で取り、これについても日英対訳対応付けを行う。	発話対応データ 文対応データ 文節対応データ 単語対応データ 格対応データ など

2-2 システム構成(図2)

関係データベース管理システム(日本語UNIFY)を用いて、ATRのホストコンピュータ(VAX8820/Ultirix)上に実装している。検索は基本的には、VT282端末(エミュレーション可)から専用検索システムを通して行うが、関係データベースの標準データ操作言語SQLによる利用も可能である。

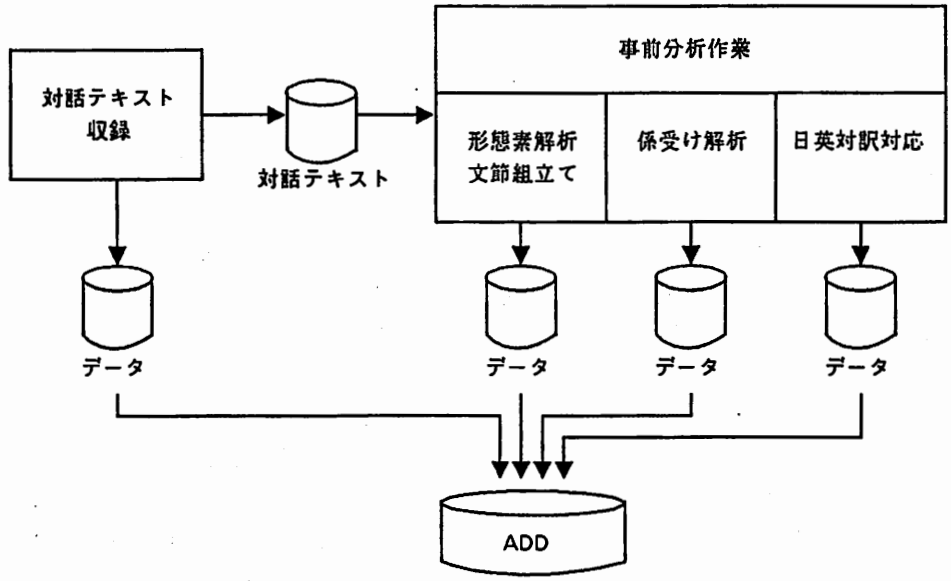


図1　ADD構築作業の概略

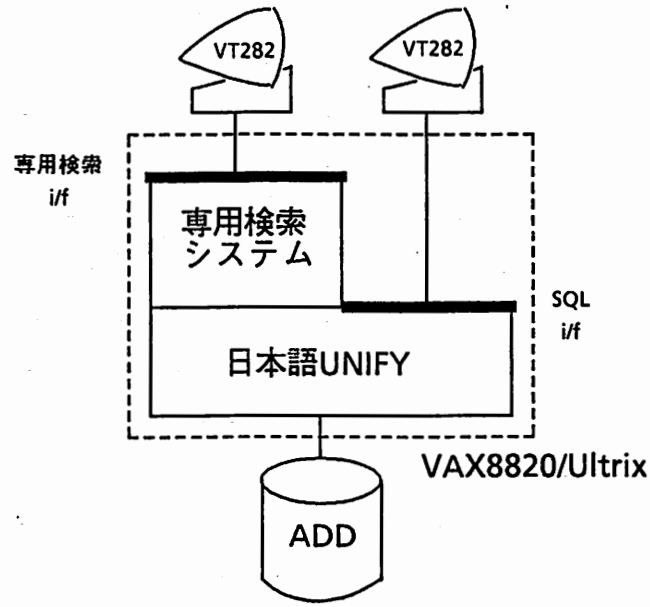


図2　システム構成

3 データ構成の基本的な考え方

対話テキストの事前分析により得られたデータは、その意味や、利用方法などに基づいて19の関係(テーブル)、158の属性(フィールド)に整理した(表2)。

各テーブルのレコードは、属性を介して様々な観点から多重に関連付けてある。したがって、ADDは従来のテキストベースのような平たんなデータ構造ではなく、高度に構造化されたデータ構造になっている。図3は、ADDの全体構成をまとめたものである。点線囲み内の項目はレコード自身の属性(実体属性)であり、下線付き項目がキーである。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。

テーブルの性質は大きく二分することができる。言語要素そのものを表す実体集合としてのテーブルと、言語要素の結びつきを表す関連集合としてのテーブルである。前者は「日本語テキストの分割単位」及び「英語テキストの分割単位」を表すテーブルであり、後者は「日本語あるいは英語の格構造」及び「日本語と英語の対応」そして「(日本語の)単語の係受け」を表すテーブルである。以下の節では、これらのテーブル群について解説を行う。

表2　ADDのテーブル一覧

No.	テーブル名	内容	性質
1	会話	日本語の会話単位を示す。	実体
2	発話	日本語の発話単位を示す。	実体
3	文	日本語の文単位を示す。	実体
4	文節	日本語の文節単位を示す。	実体
5	単語	日本語の単語単位を示す。ADDの最小単位。	実体
6	日本語格対応	日本語の格の単位を示す。	関連
7	英語会話	英語の会話単位を示す。	実体
8	英語発話	英語の発話単位を示す。	実体
9	英語文	英語の文単位を示す。	実体
10	英語単語	英語の単語単位を示す。	実体
11	英語格対応	英語の格の単位を示す。	関連
12	会話対応	日本語の会話単位と英語の会話単位の対応を示す。	関連
13	発話対応	日本語の発話単位と英語の発話単位の対応を示す。	関連
14	文対応	日本語の文単位と英語の文単位の対応を示す。	関連
15	文節対応	日本語の文節単位と英語の単語単位の対応を示す。	関連
16	単語対応	日本語の単語単位と英語の単語単位の対応を示す。	関連
17	ランダム対応	日本語の単語単位と英語の単語単位の対応を示す。	関連
18	格対応	日本語の格単位と英語の格単位の対応を示す。	関連
19	係受け	日本語の単語間の共起(二項)関係を示す。	関連

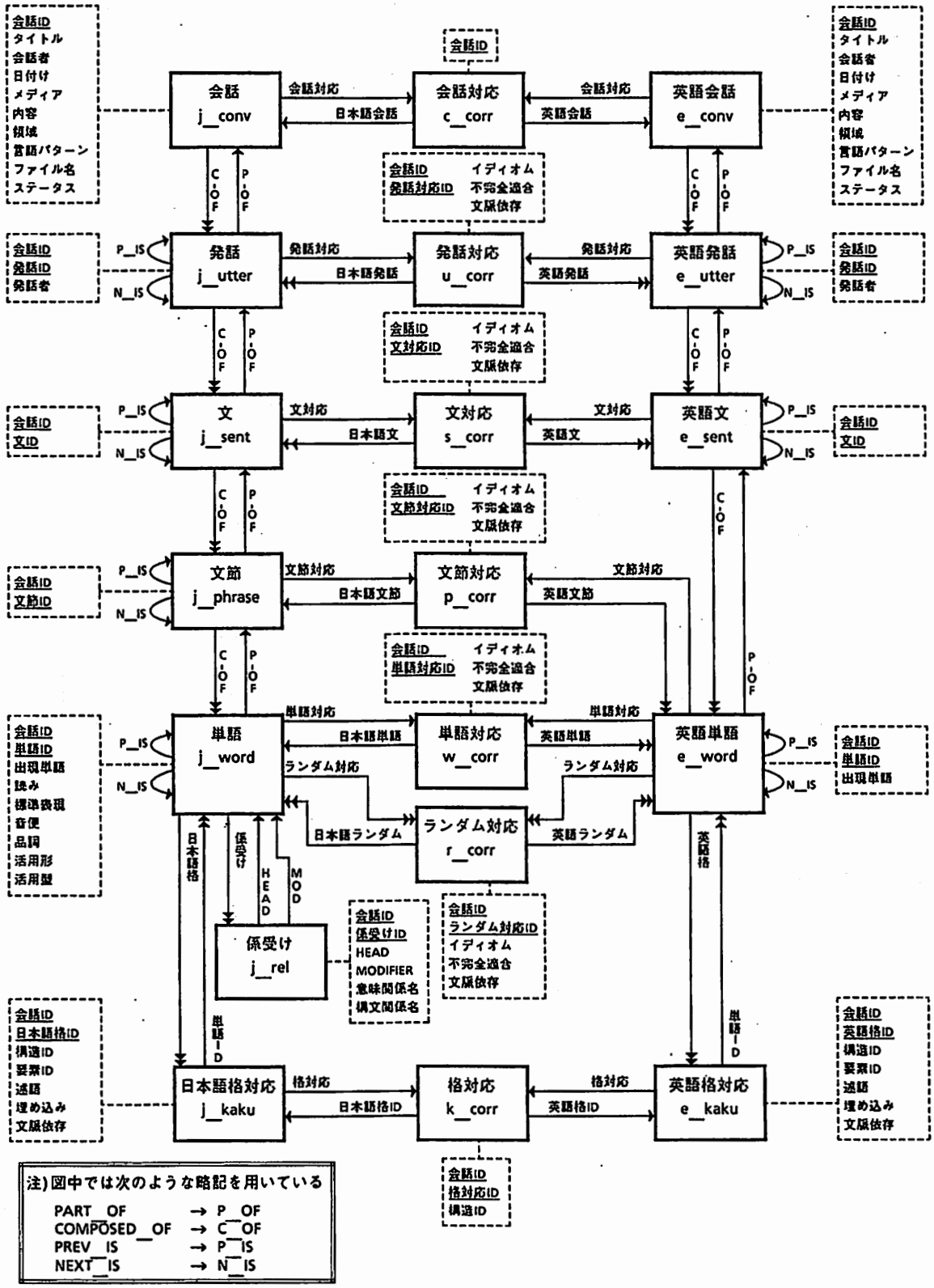

会話ID
タイトル
会話者
日付け
メディア
内容
領域
言語パターン
ファイル名
ステータス
会話
j__conv
会話対応
日本語会話
会話対応
c__corr
英語会話
会話対応
英語会話
e__conv
C-OF
P-OF
会話ID
発話ID
発話者
P__IS
N__IS
発話
j__utter
発話対応
日本語発話
発話対応
u__corr
英語発話
発話対応
英語発話
e__utter
会話ID
文ID
文
j__sent
文対応
日本語文
文対応
s__corr
英語文
文対応
英語文
e__sent
文節ID
文節
j__phrase
文節対応
日本語文節
文節対応
p__corr
英語文節
文節対応
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
日本語格
係受け
HEAD
MOD
係受け
j__rel
係受けID
HEAD
MODIFIER
意味関係名
構文関係名
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
英語格ID
イディオム
不完全適合
文脈依存
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3　ADDの全体構成

4　日本語テキストの分割単位

日本語の対話テキストを「会話」、「発話」、「文」、「文節」、「単語」という5つの単位で分割し、それぞれの単位をRDBのテーブルとした。テキストデータは単語(テーブルの)レコードだけにあり、それ以外のレコードは階層的に単語レコードを参照する。例えば、文節レコードは、その文節を構成する単語レコードを参照し、逆に単語レコードは、属する文節のレコードを参照する(図#で、P__OF、C__OFと書かれている属性がこれである)。また、「発話」、「文」、「文節」、「単語」の各レコードは、時間的順序関係を表すために、前後のレコードを参照する(図3で、P__IS、N__ISと書かれている属性がこれである)。(図4)

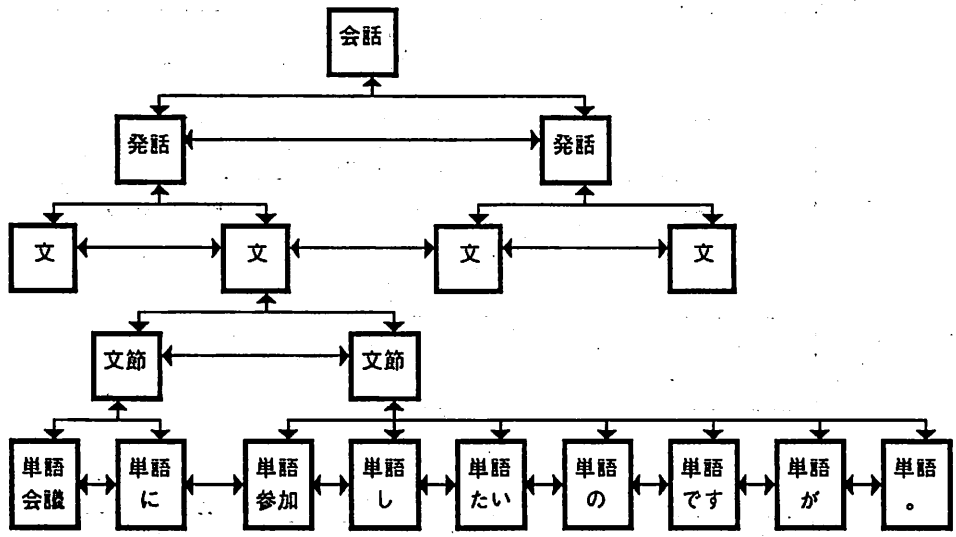


図4　日本語テキストの分割単位

5　英語テキストの分割単位

ADDは、その特徴の一つとして日本語に対応する英語のテキストを持つが、これらの英語テキストも日本語テキストと同様に単位分割して、「英語会話」、「英語発話」、「英語文」、「英語単語」というテーブルにした(ただし、英語では、文節テーブルはない)。レコード間の参照は、日本語テキストと同様である。

6　日本語あるいは英語の格構造

上記の単位とは別に、テキストの意味的なまとまりを取るため、日本語、英語それぞれに「格」テーブルを設けた。「格」テーブルのレコードは、格(要素)を構成する「単語」テーブルのレコードを参照する。逆に「単語」テーブルのレコードは、属する格(要素)のレコードを参照する。(図5)

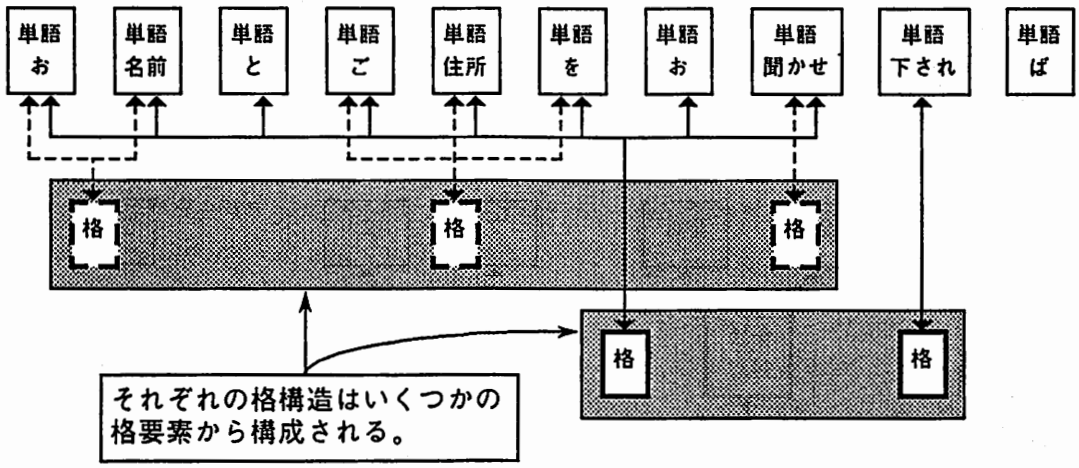


図5　日本語の格構造

7　日本語と英語の対応

日本語に対応する英語を単位毎に求められるよう、「会話対応」、「発話対応」、「文対応」、「文節対応」、「単語対応」、「ランダム対応」、「格対応」というテーブルを設けた。対応レコードは、対応の付く日本語テキストのレコードと英語テキストのレコードをそれぞれ参照する。逆に日英それぞれのレコードは、対応レコードを参照する。日英の対応関係は、会話は一対一、発話、文は多対多である。文節は、英語に存在しないので、日本語文節と英語単語との間で一対多の対応である。単語対応は一対多、ランダム対応は節(clause)や複合名詞などの対応を示し、多対多である。(図6)

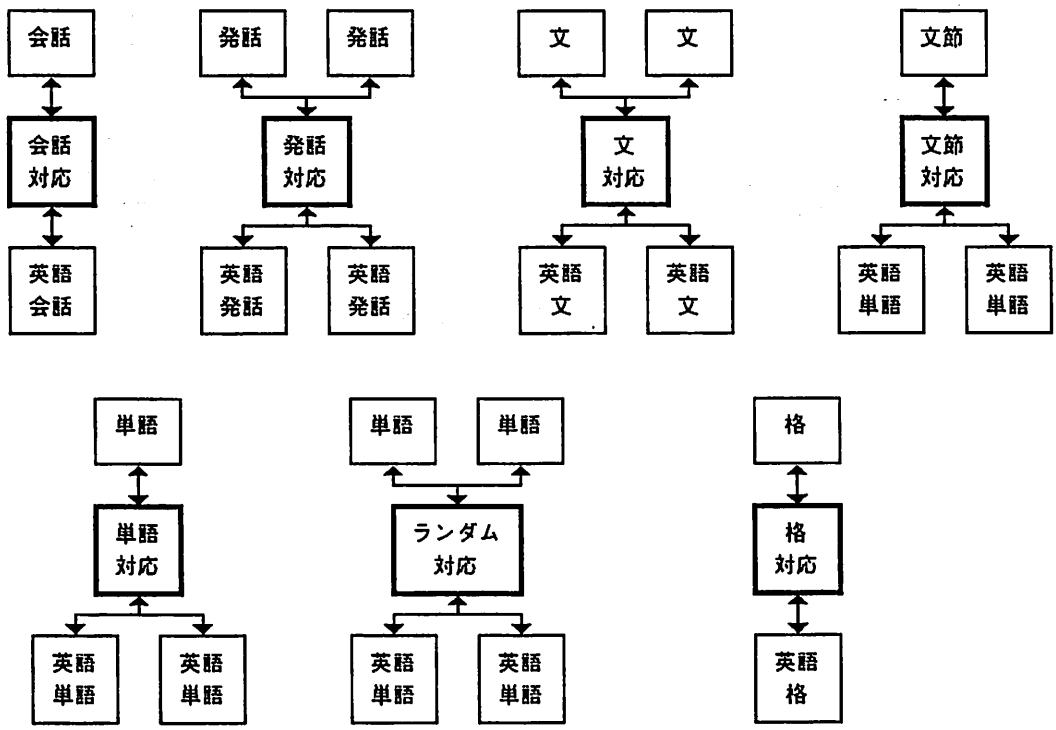


図6　日本語と英語の対応

8　単語間の係受け

単語間の意味的な関係をとらえるため、「係受け」テーブルを設けた。係受けレコードは、二項共起の関係にある単語レコードを参照する。逆に単語レコードは、その単語にどのような係受けが係っているかを示すため、係受けレコードを参照する。(図7)

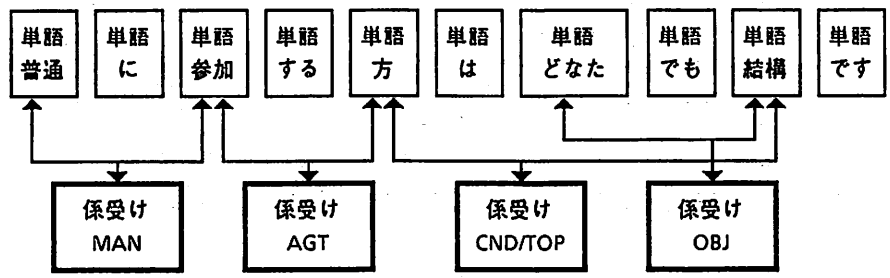


図7　係受け

9 格納方法

ADDはこれまで述べてきたように関係データベースを用いているが、データの格納方法については、二、三注意することがある。

8-1 ID規則

レコードのIDはすべて1会話毎のシーケンス、すなわち1会話内でのみユニークである。したがって、単語テーブルを例に取れば、単語ID＝100のレコードが複数存在することになる。そこで、レコードの一意識別には、会話タグとレコードIDを組合せて用いる。会話タグとは、全レコードの先頭にある属性で、そのレコードがどの会話のレコードであるかを示す。つまり、

レコードキー＝会話タグ(ID)‖レコードID　(‖は連結)

である。(図8)

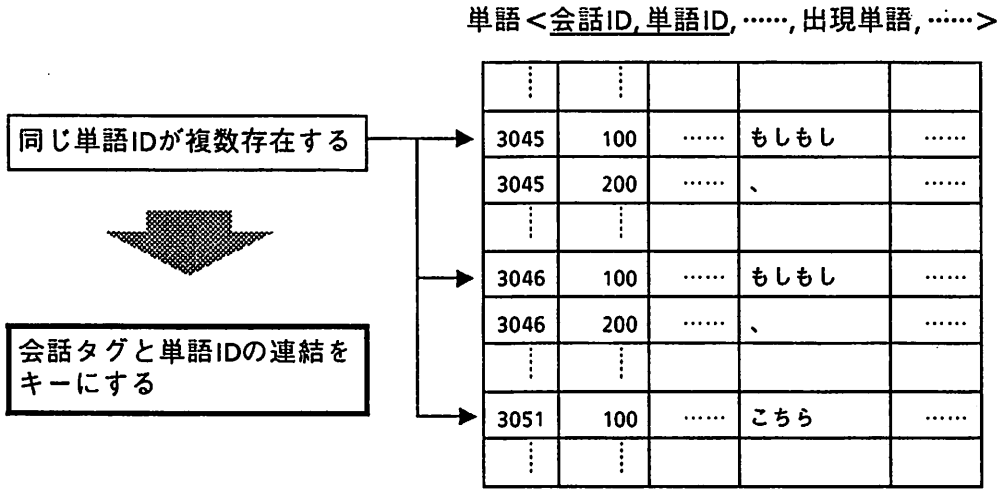


図8

8-2 関連属性

3節で、「ADDのレコードは他のレコードとの関連を示す関連属性を持つ」と述べたが、具体的には、関連するレコードのID値を直接入れている。単語テーブルを例にとると、全体部分関係を示すためPART_OFにはその単語を含む文節のIDが入り、順序関係を示すためPREV_IS、NEXT_ISには直前、直後の単語のIDがそれぞれ入る。ただし順序関係については、文の先頭(および末尾)の単語のPREV_IS(同NEXT_IS)には、'-1'を入れて、文単位の区切りを示す。(図9)

8-3 疑似マルチ型

関係データベースは基本的に、データの一貫性を保つため正規化構造を持つが、そのため複雑なデータモデルを忠実にデータベースに反映しようとする場合、テーブル数が増大し、全体として非常にわかりにくいスキーマになってしまう。ADDではテーブルの多発を防ぎ、分かりやすいスキーマを実現するため、関係データベースを単なる「容れ物」として捉え、一部、非正規構造の疑似マルチ型による格納を行っている。疑似マルチ型とは、複数のデータを記号/で

連結したデータ型で、見かけ上は単一の値(例えば、'300/400/500'のような値)であるため、従来の関係データベースで管理することができる。ただし疑似マルチ型データの論理的な操作を行うには、現状のSQLでは不十分なので、専用検索システムを通して行う。(図9)

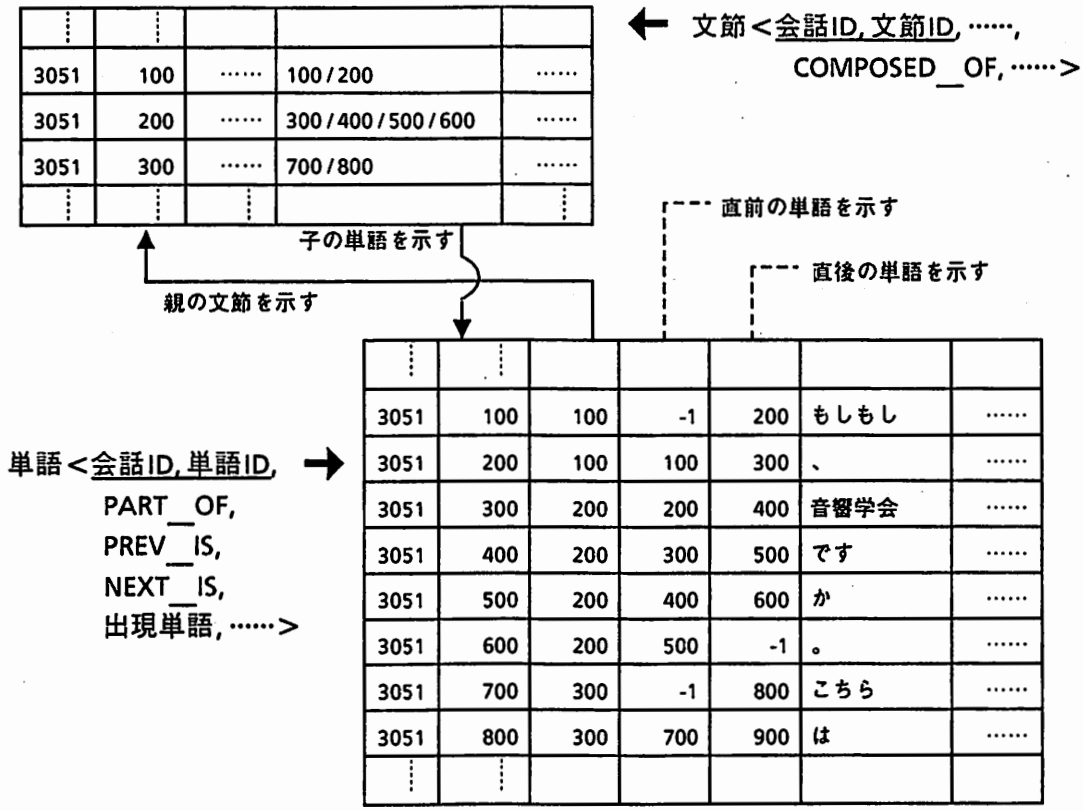


図9

10 まとめ

本稿では、自動翻訳電話研究用の言語データベースADDのデータ構成について述べた。本データベースは、対話テキストの事前分析より得られる数々のデータを整理し統合化したもので、言語現象の効率的な分析を可能にしている。すなわち、単純な文字列検索にとどまらず、文字列の属性や時間的順序関係、意味関係、さらには英語との対応などを考慮した検索ができ、これまで言語データの収集に頭を悩ませていた研究者の負担を大幅に軽減している。

本データベースはすでにATR内で活用されており、これまでに多くの言語現象の分析が行われている。しかし、ADDが持つべきデータに対するユーザの要求は、1987年のデータ収集開始時に比べて幾分変わってきている。そこで、新たに発生したユーザ要求に応えるため、現在のスキーマに新たな属性を付け加えるなどした改訂版ADDを、1990年9月にリリースする予定である。変更点および新スキーマについては、別の機会に報告する。

謝辞

ADD構築はデータ処理研究室および言語処理研究室の方々のご協力により進められました。とりわけ篠崎直子元研究技術員、井ノ上直己研究員、幸山秀雄研究員には言語データ収集の担当者として貴重な時間を多分に割いていただきました。構築に関わる各種作業については(株)インターグループ、(株)東洋情報システム、(株)エムシーワードセンター、(株)日本アイアール、(株)応用技術の各社にご協力いただきました。ここにあわせて感謝いたします。

文献

テキストデータの収集について

- 相沢(1987)「通訳を介した日英会話の収集と分析」、ATRジャーナル第2号
- 工藤、森元(1988)「キーボード会話収録システムについて」、ATRテクニカルレポートTR-I-0046
- 篠崎、小倉、森元(1988)「言語データベース作成のためのシミュレーション会話」、情報処理学会第37回全国大会論文集5B-8
- 森元、小倉、飯田(1988)「自動翻訳電話研究用言語データベースの収集について」、情報処理学会第37回全国大会論文集5B-8

形態素分析について

- 小倉、篠崎、森元(1988)「言語データベース収集支援システム」、情報処理学会第36回全国大会論文集4U-4
- 篠崎、小倉、森元(1988)「言語データベースの品質管理」、情報処理学会第36回全国大会論文集4U-5
- 篠崎、水野、小倉、吉本(1989)「形態素情報利用解説書」、ATRテクニカルレポートTR-I-0077
- 吉本(1987)「日本語品詞の分類」、ATRテクニカルレポートTR-I-0008

係受け分析について

- 井ノ上、小倉、森元(1988)「言語データベース用格.係受け意味体系」、ATRテクニカルレポートTR-I-0029

日英対訳対応について

○ 篠崎、小倉、森元(1988)「言語データベース作成のための日英対訳対応付け」、ATRテクニカルレポートTR-I-0043

格納と管理について

○ 小倉、橋本、森元(1988)「言語データベース統合管理システム」、情報処理学会研究会資料NL69-4

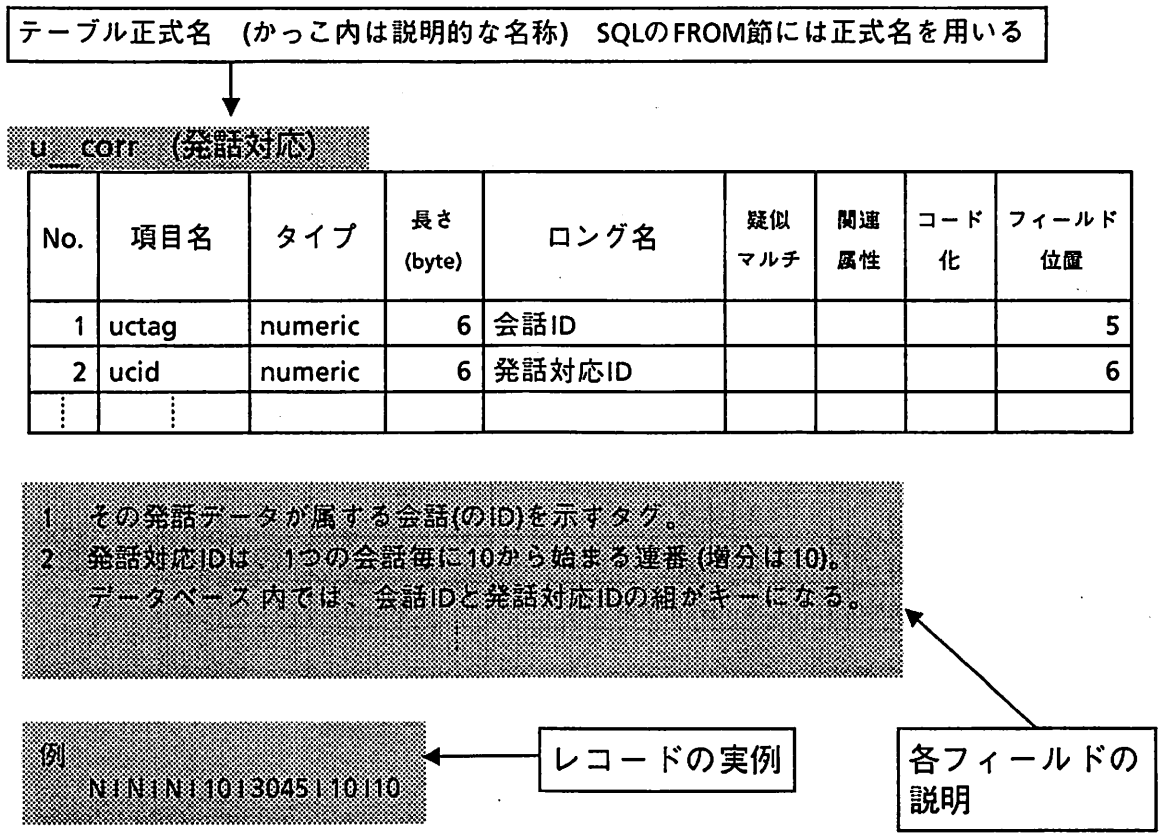
○ 橋本、小倉、森元(1989)「フレーム表現による検索機能を有する言語データベース管理システム」、情報処理学会アドバンスト・データベース・システム・シンポジウム

その他の言語データベースについて

○ 国立国語研究所(1987)、CL研究第1号

付録1　ADDスキーマ

表の見方



No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化	フィールド位置
1	uctag	numeric	6	会話ID				5
2	ucid	numeric	6	発話対応ID				6
⋮	⋮							

各列の意味は次のとおり

項目名
　フィールドの正式名称。ADD内で一意に定まっている。
タイプ
　フィールドのデータタイプで、次の4種類ある。
　numeric - 整数、date - 日付、string - 文字列、text - テキスト
長さ
　numeric とtext は、その長さ以下の不定長。dateとstringは、その長さの固定長。
ロング名
　説明的なフィールドの名称。SQLのSELECT節にはロング名を用いる。
疑似マルチ
　○ならば、疑似マルチ型フィールド。
関連属性
　○ならば、関連属性のフィールド。空白なら実体属性のフィールド。
コード化
　○ならば、コード化されたデータである。
[フィールド位置]
　実際の格納における先頭からのフィールド位置。
　この列がない時は、表の記述通りに格納されている。

j_conv (会話)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化
1	jcid	numeric	6	会話ID			
2	jccompo	text	256	COMPOSED_OF	○	○	
3	jctitle	text	100	タイトル			
4	jcspeak	text	100	会話者	○		
5	jcdate	date	8	日付け			
6	jcmed	string	1	メディア			○
7	jcsub	string	1	内容			○
8	jcdom	string	1	領域			○
9	jcfile	text	100	ファイル名			
10	jcsts	string	1	ステータス			○
11	jcpat	string	2	言語パターン			○
12	jccorr	numeric	6	会話対応		○	

1 区分コード: 0～2999 - 電話会話、3000～4999 - キーボード会話
2 子発話レコード(のID)を示す。疑似マルチ型。
例) 10/20/30/……/290
3 会話のタイトル (省略あり)。
4 次のような形式で模擬会話の役割設定を示す。xxxx、yyyy は演者の氏名(省略もあり)。
国際会議 - 事務局:xxxx/質問者:yyyy
旅行 - 担当者:xxxx/申込者:yyyy
5 会話を収録した日付(MM/DD/YY)
6 会話に用いたメディア
1 - 電話、2 - キーボード
7 会話の内容
1 - 問合せ、2 - 相談·交渉、3 - 依頼·説得、4 - 感情表現
8 会話の領域
1 - 国際会議、2 - 旅行
9 オリジナルテキストファイルのディレクトリ
10 LDBSHで利用可能かどうかの判別フラグ
I - 利用可能、空白 - 利用不可
11 模擬会話に使った言語パターンを示す。矢印(→)の出元が発呼者、出先が被呼者。
JE - 日本語→通訳→英語、EJ - 英語→通訳→日本語、JJ - 日本語→日本語、EE - 英語→英語
je - 日本語→通訳→英語(同時通訳)、ej - 英語→通訳→日本語(同時通訳)
12 会話対応レコード(のID)を示す。

例

3045|10/20/30/40/50/60/70/80/90/100/110/120| |通訳者:/事務局:|07/08/88|2|1|1|
key-152-01|I|EJ|3045

j__utter (発話)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化
1	jutag	numeric	6	会話ID			
2	juid	numeric	6	発話ID			
3	jupart	numeric	6	PART__OF		○	
4	jucompo	text	128	COMPOSED__OF	○	○	
5	juprev	numeric	6	PREV__IS		○	
6	junext	numeric	6	NEXT__IS		○	
7	juspeak	text	20	発話者			
8	jucorr	numeric	6	発話対応		○	

1 その発話データが属する会話(のID)を示すタグ。
2 発話IDは、1つの会話毎に10から始まる連番 (増分は10)。
データベース内では、会話IDと発話IDの組がキーになる。
3 親会話レコード(のID)を示す(jutag と等しい)。
4 子文レコード(のID)を示す。疑似マルチ型。
5 直前の発話レコード(のID)を示す。ただし、会話の先頭は'-1'。
6 直後の発話レコード(のID)を示す。ただし、会話の末尾は'-1'。
7 発話者の役割
例) 事務局:、通訳者:、質問者: など
8 発話対応レコード(のID)を示す。

例

3045|80|3045|2400/2500/2600|70|90|事務局:|80

j__sent　(文)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化
1	jstag	numeric	6	会話ID			
2	jsid	numeric	6	文ID			
3	jspart	numeric	6	PART__OF		○	
4	jscompo	text	256	COMPOSED__OF	○	○	
5	jsprev	numeric	6	PREV__IS		○	
6	jsnext	numeric	6	NEXT__IS		○	
7	jscorr	numeric	6	文対応		○	

1　その文データが属する会話(のID)を示すタグ。
2　文IDは、1つの会話毎に100から始まる連番 (増分は100)。
　データベース 内では、会話IDと文IDの組がキーになる。
3　親発話レコード(のID)を示す。
4　子文節レコード(のID)を示す。疑似マルチ型。
5　直前の文レコード(のID)を示す。ただし、会話の先頭は'-1'。
6　直後の文レコード(のID)を示す。ただし、会話の末尾は'-1'。
7　文対応レコード(のID)を示す。

例

3045|100|10|100/200/300/400|-1|200|100

j__phrase(文節)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化
1	jptag	numeric	6	会話ID			
2	jpid	numeric	6	文節ID			
3	jppart	numeric	6	PART__OF		○	
4	jpcompo	text	128	COMPOSED__OF	○	○	
5	jpprev	numeric	6	PREV__IS		○	
6	jpnext	numeric	6	NEXT__IS		○	
7	jpcorr	numeric	6	文節対応		○	

1　その文節データが属する会話(のID)を示すタグ。
2　文節IDは、1つの会話毎に100から始まる連番 (増分は100)。
　データベース 内では、会話IDと文節IDの組がキーになる。
3　親文レコード(のID)を示す。
4　子単語レコード(のID)を示す。疑似マルチ型。
5　直前の文節レコード(のID)を示す。ただし、文の先頭は'-1'。
6　直後の文節レコード(のID)を示す。ただし、文の末尾は'-1'。
7　文節対応レコード(のID)を示す。

例

3045|100|100|100/200|-1|200|100

j_word (単語)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化
1	jwtag	numeric	6	会話ID			
2	jwid	numeric	6	文節ID			
3	jwpart	numeric	6	PART_OF		○	
4	jwprev	numeric	6	PREV_IS		○	
5	jwnext	numeric	6	NEXT_IS		○	
6	jwcurr	text	50	出現単語			
7	jwpron	text	100	読み			
8	jwregu	text	50	標準表現			
9	jweup	string	2	音便			○
10	jwhin	string	2	品詞			○
11	jwpat	string	2	活用形			○
12	jwfo	string	2	活用型			○
13	jwrel	text	60	係受け	○	○	
14	jwccorr	text	128	日本語格	○	○	
15	jwwcorr	text	128	単語対応		○	
16	jwwcorr	text	128	ランダム対応	○	○	

1 その単語データが属する会話(のID)を示すタグ。
2 単語IDは、1つの会話毎に100から始まる連番 (増分は100)。
データベース 内では、会話IDと単語IDの組がキーになる。
3 親文節レコード(のID)を示す。
4 直前の単語レコード(のID)を示す。ただし、文の先頭は'-1'。
5 直後の単語レコード(のID)を示す。ただし、文の末尾は'-1'。
6 表記。
7 読み
8 標準表現(正規表現)
9 音便コード表参照。
10 品詞コード表参照。
11 活用形コード表参照。
12 活用型コード表参照。
13 係受けレコード(のID)を示す。疑似マルチ型。
14 日本語格レコード(のID)を示す。疑似マルチ型。
15 単語対応レコード(のID)を示す
16 ランダム対応レコード(のID)を示す。疑似マルチ型。

例

3045|700|400|600|800|し|し|する|-1|19|01|04| |600| |4

j__kaku (日本語格対応)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化	フィールド位置
1	jktag	numeric	6	会話ID				5
2	jkid	numeric	6	日本語格ID				6
3	jkstid	numeric	6	構造ID		○		3
4	jkelmid	numeric	6	要素ID		○		4
5	jkkaku	numeric	6	格対応		○		9
6	jkword	text	128	単語ID	○	○		8
7	jkpred	string	1	述語			○	7
8	jkemb	string	1	埋め込み			○	1
9	jkcntxt	string	1	文脈依存			○	2

1 その格データが属する会話(のID)を示すタグ。
2 日本語格IDは、1つの会話毎に100から始まる連番 (増分は100)。
データベース 内では、会話IDと英語格IDの組がキーになる。
3 格構造は、複数の格要素で構成される。
構造IDは、1つの会話毎に100から始まる連番 (増分は100)。
4 1つの格要素が、1つの日本語格対応レコードになる。
要素IDは、1格構造毎に1から始まる連番 (増分は1)。
(会話ID, 英語格ID) は、(会話ID, 構造ID, 要素ID) と同値。
5 格対応レコード(のID)を示す。ただし、対応がない場合は、'-1'。
6 その格に含まれる単語レコード(のID)を示す。疑似マルチ型。
7 述語フラグ。
Y-述語の特性を持つ、N-持たない
8 埋め込みフラグ。
Y-埋め込み文の特性を持つ、N-持たない
9 文脈依存フラグ。
Y-英語の格要素と文脈に依存して対応する、N-持たない

例

Y|N|100|1|3045|100|N|300/400/500|400

e__conv (英語会話)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化	フィールド位置
1	ecid	numeric	6	会話ID				1
2	eccompo	text	256	COMPOSED_OF	○	○		2
3	ectitle	text	100	タイトル				3
4	ecspeak	text	100	会話者	○			4
5	ecdate	date	8	日付け				5
6	ecmed	string	1	メディア			○	6
7	ecsub	string	1	内容			○	7
8	ecdom	string	1	領域			○	8
9	ecpat	string	2	言語パターン			○	9
10	ecfile	text	100	ファイル名				11
11	ecsts	string	1	ステータス			○	12
12	eccorr	numeric	6	会話対応		○		10

データ項目は「会話(j__conv)」と同じ

1 区分コード: 0～2999 - 電話会話、3000～4999 - キーボード会話
2 子発話レコード(のID)を示す。疑似マルチ型。
例) 10/20/30/……/290
3 会話のタイトル (省略あり)。
4 次の形式で模擬会話の役割設定を示す。xxxx、yyyy は演者の氏名(省略あり)。
国際会議 - 事務局:xxxx/質問者:yyyy
旅行 - 担当者:xxxx/申込者:yyyy
5 会話を収録した日付(MM/DD/YY)
6 会話に用いたメディア
1 - 電話、2 - キーボード
7 会話の内容
1 - 問合せ、2 - 相談･交渉、3 - 依頼･説得、4 - 感情表現
8 会話の領域
1 - 国際会議、2 - 旅行
9 模擬会話に使った言語パターンを示す。矢印(→)の出元が発呼者、出先が被呼者。
JE - 日本語→通訳→英語、EJ - 英語→通訳→日本語、JJ - 日本語→日本語、EE - 英語→英語
je - 日本語→通訳→英語(同時通訳)、ej - 英語→通訳→日本語(同時通訳)
10 オリジナルテキストファイルのディレクトリ
11 LDBSHで利用可能かどうかの判別フラグ
1 - 利用可能、空白 - 利用不可
12 会話対応レコード(のID)を示す。

例

3045 | 10/20/30/40/50/60/70/80/90/100/110/120 | | 通訳者: / 事務局: | 07/08/88 | 2 | 1 | 1 | 1 | EJ | 3045 | key-152-01 | |

e__utter (英語発話)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化
1	eutag	numeric	6	会話ID			
2	euid	numeric	6	発話ID			
3	eupart	numeric	6	PART__OF		○	
4	eucompo	text	128	COMPOSED__OF	○	○	
5	euprev	numeric	6	PREV__IS		○	
6	eunext	numeric	6	NEXT__IS		○	
7	euspeak	text	20	発話者			
8	eucorr	numeric	6	発話対応		○	

1 その発話データが属する会話(のID)を示すタグ。
2 発話IDは、1つの会話毎に10から始まる連番 (増分は10)。
データベース内では、会話IDと発話IDの組がキーになる。
3 親会話レコード(のID)を示す(eutag と等しい)。
4 子文レコード(のID)を示す。疑似マルチ型。
5 直前の発話レコード(のID)を示す。ただし、会話の先頭は'-1'。
6 直後の発話レコード(のID)を示す。ただし、会話の末尾は'-1'。
7 発話者の役割
例) 事務局:、通訳者:、質問者: など
8 発話対応レコード(のID)を示す。

例
3045 | 30 | 3045 | 400/500/600/700/800/900 | 20 | 40 | 質問者: | 30

e__sent (英語文)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化	フィールド位置
1	estag	numeric	6	会話ID				1
2	esid	numeric	6	文ID				2
3	espart	numeric	6	PART__OF		○		4
4	escompo	text	256	COMPOSED__OF	○	○		5
5	esprev	numeric	6	PREV__IS		○		6
6	esnext	numeric	6	NEXT__IS		○		7
7	escorr	numeric	6	文対応		○		3

1 その文データが属する会話(のID)を示すタグ。
2 文IDは、1つの会話毎に100から始まる連番 (増分は100)。
データベース 内では、会話IDと文IDの組がキーになる。
3 親発話レコード(のID)を示す。
4 子単語レコード(のID)を示す。疑似マルチ型。
5 直前の文レコード(のID)を示す。ただし、会話の先頭は'-1'。
6 直後の文レコード(のID)を示す。ただし、会話の末尾は'-1'。
7 文対応レコード(のID)を示す。

例
3045 | 300 | 200 | 20 | 2300/2400/2500/2600/2700 | 200 | 400

e__word (英語単語)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化	フィールド位置
1	ewtag	numeric	6	会話ID				2
2	ewid	numeric	6	文節ID				3
3	ewpart	numeric	6	PART__OF		○		4
4	ewprev	numeric	6	PREV__IS		○		5
5	ewnext	numeric	6	NEXT__IS		○		6
6	ewcurr	text	50	出現単語				7
7	ewccorr	text	128	英語格	○	○		9
8	ewpcorr	text	128	文節対応		○		1
9	ewwcorr	text	128	単語対応		○		8
10	ewwcorr	text	128	ランダム対応	○	○		10

1 その単語データが属する会話(のID)を示すタグ。
2 単語IDは、1つの会話毎に100から始まる連番 (増分は100)。
データベース 内では、会話IDと単語IDの組がキーになる。
3 親文レコード(のID)を示す。
4 直前の単語レコード(のID)を示す。ただし、文の先頭は'-1'。
5 直後の単語レコード(のID)を示す。ただし、文の末尾は'-1'。
6 表記。
7 英語格レコード(のID)を示す。疑似マルチ型。
8 文節対応レコード(のID)を示す。
9 単語対応レコード(のID)を示す。
10 ランダム対応レコード(のID)を示す。疑似マルチ型。

例

500 | 3045 | 900 | 100 | 800 | 1000 | inquiry | 400 | 600/500

e_kaku (英語格対応)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化	フィールド位置
1	ektag	numeric	6	会話ID				5
2	ekid	numeric	6	英語格ID				6
3	ekstid	numeric	6	構造ID		○		3
4	ekelmid	numeric	6	要素ID		○		4
5	ekkaku	numeric	6	格対応		○		9
6	ekword	text	128	単語ID	○	○		8
7	ekpred	string	1	述語			○	7
8	ekemb	string	1	埋め込み			○	1
9	ekcntxt	string	1	文脈依存			○	2

1 その格データが属する会話(のID)を示すタグ。
2 英語格IDは、1つの会話毎に100から始まる連番 (増分は100)。
データベース 内では、会話IDと英語格IDの組がキーになる。
3 格構造は、複数の格要素で構成される。
構造IDは、1つの会話毎に100から始まる連番 (増分は100)。
4 1つの格要素が、1つの日本語格対応レコードになる。
要素IDは、1格構造毎に1から始まる連番 (増分は1)。
(会話ID, 英語格ID) は、(会話ID, 構造ID, 要素ID) と同値。
5 格対応レコード(のID)を示す。ただし、対応がない場合は、'-1'。
6 その格要素を構成する単語レコード(のID)を示す。疑似マルチ型。
7 述語フラグ。
Y - 述語の特性を持つ、N - 持たない
8 埋め込みフラグ。
Y - 埋め込み文の特性を持つ、N - 持たない
9 文脈依存フラグ。
Y - 英語の格要素と文脈に依存して対応する、N - しない

例

Y | N | 100 | 3 | 3045 | 700 | N | 1000/1100/1200/1300/1400 | 400

c__corr　(会話対応)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化
1	ccid	numeric	6	会話ID			
2	ccjpn	numeric	6	日本語会話		○	
3	cceng	numeric	6	英語会話		○	

1　会話ID。
2　日本語会話レコード(のID)を示す。
3　英語会話レコード(のID)を示す。

例

3045 | 3045 | 3045

u_corr (発話対応)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化	フィールド位置
1	uctag	numeric	6	会話ID				5
2	ucid	numeric	6	発話対応ID				6
3	ucjpn	text	128	日本語発話	○	○		4
4	uceng	text	128	英語発話	○	○		7
5	ucidm	string	1	イディオム				1
6	ucun	string	1	不完全適合				2
7	uccntxt	string	1	文脈依存				3

1 その発話データが属する会話(のID)を示すタグ。
2 発話対応IDは、1つの会話毎に10から始まる連番 (増分は10)。
データベース 内では、会話IDと発話対応IDの組がキーになる。
3 日本語発話レコード(のID)を示す。疑似マルチ型。
4 英語発話レコード(のID)を示す。疑似マルチ型。
5 イディオム (未使用)
6 不完全適合 (未使用)
7 文脈依存 (未使用)

例
N I N I N I 10 I 3045 I 10 I 10

s_corr (文対応)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連	コード化	フィールド位置
1	sctag	numeric	6	会話ID				5
2	scid	numeric	6	文対応ID				6
3	scjpn	text	128	日本語文	○	○		4
4	sceng	text	128	英語文	○	○		7
5	scidm	string	1	イディオム				1
6	scun	string	1	不完全適合				2
7	sccntxt	string	1	文脈依存				3

1 その文対応データが属する会話(のID)を示すタグ。
2 文対応IDは、1つの会話毎に100から始まる連番 (増分は100)。
データベース 内では、会話IDと文対応IDの組がキーになる。
3 日本語文レコード(のID)を示す。疑似マルチ型。
4 英語文レコード(のID)を示す。疑似マルチ型。
5 イディオム (未使用)
6 不完全適合 (未使用)
7 文脈依存 (未使用)

例

N I N I N I 20 I 3045 I 200 I 200/300

p_corr (文節対応)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化
1	pctag	numeric	6	会話ID			
2	pcid	numeric	6	文節対応ID			
3	pcjpn	numeric	6	日本語文節		○	
4	pceng	text	128	英語単語	○	○	
5	pcidm	string	1	イディオム			○
6	pcun	string	1	不完全適合			○
7	pccntxt	string	1	文脈依存			○

1 その文節対応データが属する会話(のID)を示すタグ。
2 文節対応IDは、1つの会話毎に100から始まる連番 (増分は100)。
データベース内では、会話IDと文節対応IDの組がキーになる。
3 日本語文節レコード(のID)を示す。
4 英語単語レコード(のID)を示す。疑似マルチ型。
5 イディオムフラグ。
Y-日英の文節がイディオムとして対応する、N-しない
6 不完全適合フラグ。
Y-日英の文節が部分的に対応する、N-しない
文脈依存とは共存しない
7 文脈依存フラグ。
Y-日英の文節が文脈に依存して対応する、N-しない

例

3045 | 1000 | 1000 | 3300/3400/3500/3600/3700 | N | Y | N

w__corr (単語対応)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化
1	wctag	numeric	6	会話ID			
2	wcid	numeric	6	単語対応ID			
3	wcjpn	numeric	6	日本語単語		○	
4	wceng	text	128	英語単語	○	○	
5	wcidm	string	1	イディオム			○
6	wcun	string	1	不完全適合			○
7	wccntxt	string	1	文脈依存			○

1 その単語対応データが属する会話(のID)を示すタグ。
2 単語対応IDは、1つの会話毎に100から始まる連番 (増分は100)。
データベース内では、会話IDと単語対応IDの組がキーになる。
3 日本語単語レコード(のID)を示す。
4 英語単語レコード(のID)を示す。疑似マルチ型。
5 イディオムフラグ。
Y-日英の単語がイディオムとして対応する、N-しない
6 不完全適合フラグ。
Y-日英の単語が部分的に対応する、N-しない
文脈依存とは共存しない
7 文脈依存フラグ。
Y-日英の単語が文脈に依存して対応する、N-しない

例

3045 | 1000 | 1800 | 1900 | N | N | N

r_corr　(ランダム対応)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化
1	rctag	numeric	6	会話ID			
2	rcid	numeric	6	ランダム対応ID			
3	rcjpn	text	128	日本語ランダム	○	○	
4	rceng	text	128	英語ランダム	○	○	
5	rcidm	string	1	イディオム			○
6	rcun	string	1	不完全適合			○
7	rccntxt	string	1	文脈依存			○

1　そのランダム対応データが属する会話(のID)を示すタグ。
2　ランダム対応IDは、1つの会話毎に100から始まる連番 (増分は100)。
　データベース 内では、会話IDとランダム対応IDの組がキーになる。
3　日本語単語レコード(のID)を示す。疑似マルチ型。
4　英語単語レコード(のID)を示す。疑似マルチ型。
5　イディオムフラグ。
　　Y - 日英の単語群がイディオムとして対応する、N - しない
6　不完全適合フラグ。
　　Y - 日英の単語群が部分的に対応する、N - しない
　　文脈依存とは共存しない
7　文脈依存フラグ。
　　Y - 日英の単語群が文脈に依存して対応する、N - しない

例
　3045 | 100 | 800/900/1000 | 300/400/500/600 | Y | N | N

k_corr (格対応)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化	フィールド位置
1	kctag	numeric	6	会話ID				4
2	kcid	numeric	6	格対応ID				5
3	kcstid	numeric	6	構造ID				1
4	kcjpn	numeric	6	日本語格ID		○		2
5	kceng	numeric	6	英語格ID		○		3

1 その格対応データが属する会話(のID)を示すタグ。
2 格対応IDは、1つの会話毎に100から始まる連番 (増分は100)。
データベース 内では、会話IDと格対応IDの組がキーになる。
3 日本語格対応レコードの構造IDを示す。
4 日本語格対応レコード(のID)を示す。
5 英語格対応レコード(のID)を示す。

例

100 | 300 | 300 | 600 | 3045 | 300

j_rel (係受け)

No.	項目名	タイプ	長さ (byte)	ロング名	疑似マルチ	関連属性	コード化
1	jreltag	numeric	6	会話ID			
2	jrelid	numeric	6	係受けID			
3	jrelhead	numeric	6	HEAD		○	
4	jrelfier	numeric	6	MODIFIER		○	
5	jsemrel	string	7	意味関係名			○
6	jsynrel	string	2	構文関係名			○

1 その係受けデータが属する会話(のID)を示すタグ。
2 係受けIDは、1つの会話毎に100から始まる連番 (増分は100)。
データベース 内では、会話IDと係受けIDの組がキーになる。
3 係元。主に動詞。
4 係先。主に名詞。
5 意味関係コード表参照。
6 構文関係コード表参照。

例
3045 | 1000 | 4100 | 4000 | PRD | 01

付録2　コード表

本付録では、下記のコードについての一覧を示す。これ以外のコードについては、付録1を参照すること。

j__eup　(音便コード表)

No.	項目名	タイプ	表示長 (byte)	ロング名
1	jeupcode	string	2	音便コード
2	jeupname	string	10	音便名

1　音便コード
2　音便名

音便コード	音便名
-1	NIL
00	撥音便
01	促音便
02	イ音便
03	ウ音便

j__patt　(活用形コード表)

No.	項目名	タイプ	表示長 (byte)	ロング名
1	jpatcode	string	2	活用形コード
2	jpatname	string	10	活用形名

1　活用形コード
2　活用形名

活用形コード	活用形名
-1	NIL
00	未然
01	連用
02	終止
03	連体
04	仮定
05	命令
06	語幹

j__form (活用型コード表)

No.	項目名	タイプ	表示長 (byte)	ロング名
1	jfocode	string	2	活用型コード
2	jfoname	string	10	活用型名

1 活用型コード
2 活用型名

活用型コード	活用型名
-1	NIL
00	変則型
01	五段
02	上一
03	下一
04	サ変
05	カ変
06	特殊
11	文語四段
12	文語上二
13	文語下二
14	文語ラ変
15	文語ナ変
21	形容詞ク変

j__parts (品詞コード表)

No.	項目名	タイプ	表示長 (byte)	ロング名
1	jhincode	string	2	品詞コード
2	jhinname	string	10	品詞名

1 品詞コード
2 品詞名

品詞コード	品詞名	品詞コード	品詞名
-1	NIL	19	補助動詞
00	記号	20	自動詞
01	形容詞	21	他動詞
02	形容動詞ダ	30	固有名詞
03	形容動詞タ	31	形容名詞
04	普通名詞	32	本動詞
05	サ変名詞	33	間投詞
06	代名詞	34	準体助詞
07	数詞	35	並立助詞
08	副詞	36	係助詞
09	連体詞	80	慣用句
10	接続詞	99	その他
11	感動詞		
12	助動詞		
13	副助詞		
14	接続助詞		
15	格助詞		
16	終助詞		
17	接尾語		
18	接頭語		

j__sem　(係受け意味関係コード表)

No.	項目名	タイプ	表示長 (byte)	ロング名
1	jsemcode	string	2	意味関係コード
2	jsemname	string	10	意味関係名

1　意味関係コード
2　意味関係名

意味関係コード	意味関係名	意味関係コード	意味関係名	意味関係コード	意味関係名
ACC	随伴	GAU	原因、理由	REC	受け手
ADD	追加	GOA	状態変化の終点	REL	関係の基準
AGT	動作主	INS	挿入	RNG	範囲規定、関係
APP	同格	MAN	方式、様態	ROL	役割
ATT	属性	MAT	材料	SEL	選択
AUT	作成者	MES	比喩実体	SPA	場所
AVO	対象属性値	MET	比喩	SPF	場所、起点
CIR	付帯状況	NOM	命名	SPR	場所、通過
CMP	補文標識	OAT	属性対象	SPT	場所、終点
CNC	譲歩	OBJ	対象	SRC	状態変化の始点
CND	条件	OPP	相手2	TMA	時点
CNE	連接	ORG	与え手	TMD	時、継続
CON	内容	ORR	選言	TMF	時、起点
COR	比較	OTH	その他	TMT	時、終点
DGR	程度1	PAR	部分	TOO	道具、手段
DLM	限定	PMC	準備文標識	TOP	話題提示
EVA	価値判断	POS	所有者	UVA	無意志動作主
EXM	例示	PRD	陳述	UVS	無意志主体
EXP	経験者	PRL	意味並列	VAL	属性値
EXT	程度2	PRP	目的	VIE	観点
GAI	外格	PRT	相手1	WHL	全体

j__syn (係受け構文関係コード表)

No.	項目名	タイプ	表示長 (byte)	ロング名
1	jsyncode	string	2	構文関係コード
2	jsynname	string	10	構文関係名

1　構文関係コード
2　構文関係名

構文関係コード	構文関係名
01	連用格
02	連体修飾
03	連体格
04	述語連用修飾
05	文連用修飾
06	並列
07	複合名詞内修飾
08	補助