

電話またはキーボードを介した対話に基づく対話データベース ADD の構築†

江 原 暉 将** 小 倉 健 太 郎*** 篠 崎 直 子**
森 元 逞†† 樽 松 明††

計算機による対話処理, 特に自動翻訳電話技術を達成するには, その対象となる話し言葉に関する言語現象を詳細に知ることが必要である。このため, 主に電話またはキーボードを用いた目的指向型の対話に基づいて, 対話データベースを作成した。本データベースは以下のような特徴を持つ。①日本語話者と英語話者の対話または日本語話者同士の対話に基づいている。前者は通訳を介し, 後者は翻訳によって日英2か国語の対訳データを収容している。②録音データを単に文字化だけでなく, 事前に各種の言語分析を加え, 利用に際しての情報抽出を容易にしている。③総延べ語数が日本語だけで100万語以上と大規模である。事前分析の内容は具体的に, 以下の項目から成る。①単語, 句 (文節 日本語のみ), 文などの言語単位を認定する。②単語単位の属性としての品詞や読みなどを付与する。③日本語の単語間の格・係り受け関係を認定し, その属性を付与する。④日本語と対訳の英語で同一の意味内容を持つ部分の対応である日英対応を付与する。事前分析の結果を利用して, 各種言語現象の抽出や, 翻訳のための知識ベースの構築が効率的に行える。本データベースは大規模な英日対話データベースとして唯一のものであり, 対話処理, 対話モデル, 対話システム (特に, 自動翻訳電話) の研究に有効である。

1. はじめに

いかなる技術にしても, その技術を達成するには, 対象にしようとしている自然現象を詳細に知ることが必要であり, 自然言語処理技術の場合も例外ではない。自然言語処理が対象とするものは人間の用いる言語であるから, 言語現象に関する知識を得ることが必要となる。従来, これらの知識は, 主として, 母国語話者の内省*** に頼って構築されてきたが, 内省には

揺れがあり, また定量性に欠ける欠点があった。一方, 言語データベースを利用して, 知識を構築する方法は, 客観性に優れ, 定量的な知識を得ることができる利点がある。

従来の言語データベースとしては, 書き言葉に関するものが多かった。例えば, 英語に関して, Brown コーパス²⁾や Lancaster-Oslo-Bergen コーパス³⁾があり, 日本語では, 国立国語研究所の言語データベース^{4), 5)}などがある。一方, 話し言葉に関して, London-Lund コーパス (英語)⁶⁾, 国立国語研究所の話し言葉データベース (日本語)⁷⁾ があるが, 単一言語によるコーパスである。

本文で述べる言語データベース ADD**** は自然言語処理技術のうち, 対話処理技術, 特に自動翻訳電話の設計・作成・試験のために構築されたもので, 具体的には以下の利用目的を持っている。

- 自動翻訳電話の設計のために各種言語現象を抽出調査する。
- 用例に基づいて翻訳精度の向上を図る, いわゆる用例主導型翻訳に利用する知識ベースを構築するための基礎データとする。
- 自動翻訳電話のうち, 主に機械翻訳部分の試験をするためのテストデータを得る。
- 対話処理, 対話モデル, 目的指向の対話システムの

† Construction of a Dialogue Database (ADD) from Dialogues through Telephone or Keyboard Channels by TERUMASA EHARA (Knowledge and Data Base Department, ATR Interpreting Telephony Research Laboratories), KENTARO OGURA (Knowledge Systems Laboratory, NTT Network Information Systems Laboratories, Nippon Telegraph and Telephone Corporation), NAOKO SHINOZAKI (Development Section 2, Information System Research and Development Center, Sharp Corporation), TSUYOSHI MORIMOTO (Knowledge and Data Base Department, ATR Interpreting Telephony Research Laboratories) and AKIRA KUREMATSU (ATR Interpreting Telephony Research Laboratories).

†† (株)エイ・ティ・アール自動翻訳電話研究所データ処理研究室

††† 日本電信電話(株)NTT 情報通信網研究所知識処理研究部

‡ シャープ(株)技術本部情報技術開発センター第2開発部

‡‡ (株)エイ・ティ・アール自動翻訳電話研究所

* 現在 日本放送協会放送技術研究所
Science and Technical Research Laboratories, Japan
Broadcasting Corporation

** 本論文の内容は著者が(株)エイ・ティ・アール自動翻訳電話研究所に所属していた時の成果である。

*** 自分自身の思考過程の観察。ここでは, ある言語現象が成立するかどうかを被験者の直感に頼って判断すること¹⁾。

**** 本データベースは ATR 対話データベース, ATR Dialogue Database と呼ばれる。以下, ADD と略称する。

研究に基礎データを提供する。

ADD の特徴として、以下の点が挙げられる。

- 主に電話またはキーボードを介した目的指向型の対話に基づいて構築された話し言葉ないしは疑似話し言葉に関するデータベースであり、対話処理の研究にきわめて有効である。
- 日本語話者と英語話者の対話または日本語話者同士の対話に基づいている。前者は通訳を介し、後者は翻訳によって日英2か国語の対訳データを収容している。
- 録音データを単に文字化するだけでなく、事前に各種の言語分析を加え、利用に際しての情報抽出を容易にしている。特に、日英の対応データが存在する。
- 総延べ語数が日本語だけで100万語以上あり、大規模な対話データベースとして唯一のものである。

以下、2章では、データ収集の方法について、3章では事前分析の方法について、4章ではADDの利用例について述べる。

2. データの収集

言語データベースを構築するに当たって、どのようなデータを収集するかが第1に重要である。これは、データベースの利用目的に合ったものでなければならない。ADDは対話処理、特に自動翻訳電話の設計・作成・試験に利用することが目的であるから、その対象となる言語行為に関する偏りのない標本となっていることが望ましい。しかし、ここで「対象となる言語行為」とはどの範囲かということ、「偏りのない標本」をいかにして収集するかということが問題になる。

前者に関しては、以下のように考えた。自動翻訳電話の対象は技術の進歩によって、拡大し得る。初めは、簡単な領域に対する簡単な構文による対話を対象にしていたとしても、技術の進歩によって、対象がより複雑なものへと変化し得る。一方、大規模な対話データベースを構築するのは、困難な作業であり、できるだけ汎用性を持たせることが望ましい。そこで、「対象となる言語行為」として、将来にわたって利用できるように、かなり広範囲に考えた。具体的には、2人の人間（通訳を介した場合は3人）による制約のできるだけ少ない電話による目的指向対話（後述）を主たる収集対象とした。しかし、このような電話対話を近未来における自動翻訳電話の処理対象とすることには、困難性が強いと予想されるので、直近の対象と

なり得る対話として、電話ではなく、キーボード（以下ではキーと略記する）による対話も収集した⁹⁾。また、手紙による対話データも小量収集した。ただし、キーや手紙の場合も対話の方法に関する制約はできるだけ少なくした。

対話の種類として、雑談のように特定の対話目的を持たない「自由展開型」と情報伝達や行動の要求など明確な目的を持った「目的指向型」およびこれらの「混合型」の3種がある⁹⁾。自由展開型の対話の場合は、話題の変化する度合いが大きいことや、対話者間での主題の共有が少ないなど、機械処理での困難性が予想される。このため、かなりの将来を見ても自動翻訳電話の対象としては不適切であると考え、目的指向型の対話のみを対象とした。これは、ちょうど書き言葉の機械翻訳が目的指向型の文書であるマニュアル、論文、新聞記事などを対象にしておき、随筆や詩などを対象にしていないことに対応する。

次に、「偏りのない標本」を収集するには、現実社会で行われている目的指向型の電話やキー対話から無作為に標本を抽出することが望ましいが、通信の秘密を侵さない範囲で、このような抽出をすることは困難である。そこで、われわれは模擬実験対話によるデータ収集^{10)~12)}を主体にし、一部、ラジオ番組で用いられた現実の電話対話も収集した。模擬実験を行うに当たっては、できるだけ対話方法に制約が加わらないように、被験者には、対話内容に関する最小限の情報をプロットとして与えるに止め、話し方（キー対話なら入力の方法）は被験者の自由とした。広範囲な内容について収集するには、話題の領域をできるだけ広くすることが望ましいが、そうすると、各領域での収集会話量が少なくなってしまう、標本としての有効性が薄れてしまう。そこで、目的指向型の対話目的として考えられる、情報の収集、情報の提供、行動の要求、感謝や抗議が、少なくとも含まれるように、模擬実験の領域を次のように設定した。

- 国際会議の開催や参加に関する事務局員と参加者との対話
 - 旅行に関する旅行会社窓口と客との対話
 - ホテルの予約に関する予約係と客との対話
 - 学校訪問に関する高校生同士の対話
- 上記対話目的を持つ対話の例として、次がある。
- 情報の収集
 - 国際会議参加申込方法の問い合わせ
 - 情報の提供

学校祭への参加の案内

- 行動の要求
ホテルの宿泊予約
- 感謝や抗議、苦情などの感情表現

事故による旅行計画の狂いについての抗議

被験者として、以下の対話者はそれぞれの専門家に担当してもらった。

- 国際会議の事務局員
- 旅行会社の窓口担当
- ホテルの予約係
- 通訳者

それ以外の被験者は非専門家である。模擬会話収集に関する以下の情報を会話情報ファイルにまとめた。これは、被験者に提示したプロットの内容、被験者の氏名、性別、年齢、居住地、生育地、使用言語、役割(事務局員か客かなどの別および、役割としての想定年齢)、収集の日付、対話の領域、対話目的、メディア(電話、キー、手紙の別)、言語パターン(日英または日日)の情報から成る。

一方、ラジオ番組での対話は模擬実験ではない現実の電話対話の標本として以下のものを収集している。

- NHK ラジオ第1放送の番組「今日も元気で・ふるさと通信」における、聞き手と全国各地の通信員などとの電話対話

これは、情報の収集を目的とした対話である。ラジオ番組の対話にも会話情報ファイルを作成している。ラジオ番組の対話は英語に翻訳していない。

電話による模擬対話実験の場合、収録は多く家庭用のカセットテープレコーダとピンマイクを用いて行った。ただし、一部のデータはラジオスタジオを用いて、高品質で収録された。後者のデータは音声処理のためのデータとしても利用し得る。

表 1 ADD のデータ量
Table 1 Word counts of ADD.

メディア	領 域	単 語 数 (日本語部分)
電 話	国際会議	211,628
	旅 行	282,156
	ホテル予約	68,259
	学生生活	11,566
キ ー	ラジオ番組	197,568
	国際会議	203,969
手 紙	旅 行	214,389
	国際会議	12,252
	旅 行	12,437

現在までに収集されたデータ量を表1に示す。これは、データ収集と文字化を行ったデータ量であり、後述する事前分析は一部のデータには施されていない。

3. 事前分析

収集された対話データを単に文字化して記録しておくだけでなく、言語学的分析を事前に加えている。これらの分析は一般に以下のように分類できる。

- 言語単位*を認定する。
- 言語単位に属性を付与する。
- 言語単位間の関係を認定する。
- 言語単位間の関係に属性を付与する。

これらのうち、ADDの利用目的から有効なものを選んで実施した。以下でその内容を、文字化も含めて述べるが、表2に実施した事前分析の内容と、分析結果の利用例についてまとめて示す。表2から分かるように、事前分析を行うことは、データベースの効率的な利用のためにきわめて有効である。なお、以下の作業のうち、単語単位の認定と単語属性の付与は、形態素解析プログラムを利用して半自動で行われたが、それ以外は手作業によった。

3.1 文 字 化

収集された電話対話は事前分析の第1段階として、文字化を行う。日本語は漢字かな混じりで、英語は通常の表記法で文字化を行い、音声字母などは用いていない。文字化により発音が曖昧になる場合、例えば、数字の「0」が、ゼロかレイか(英語では zero または 0) 曖昧になる場合は、〈 〉記号で囲んで読みを

表 2 事前分析の内容と分析結果の利用例
Table 2 Contents of the preliminary analyses and utilization of them.

分 析 内 容	利 用 例
日本語形態素解析	単語パターンの抽出 文節パターンの抽出 文パターンの抽出
格・係り受け解析	係り受け知識ベースの作成
英語形態素解析	単語パターンの抽出 文パターンの抽出
日 英 対 応	日英での対応する表現の抽出 用例主導型翻訳のための知識ベースの作成

* 言語を分析するに際して分析の基準となる単位のこと、なんらかの成分、要素、材料となる¹⁾。例えば、単語、文節、文、発話、会話など。

添えた。

句点（英語ではピリオド）は文の切れ目に挿入した。文の認定は文献 14) を参考にし、次の 3 つの基準を利用して行った。

第 1 の基準が主な基準である。

- 発話内容の 1 まとまりを 1 文とする。
- 文法形式（例えば終助詞や接続助詞）を用いて、文の切れ目かどうかを判断する。
- 韻律情報（ポーズ、イントネーション）を用いて、文の切れ目かどうかを判断する。

文の切れ目かどうか、判断しがたいときは、文の途中であるとした。文献 14) の文認定法と大きく異なる点は、発話の先頭の間投詞を 1 文としていない点である。これによって、14) と比較して、1 文の平均長さがかなり長くなっている。

韻律情報のうち、1 文内のポーズに関しては、その長さを作業者が判断して、読点（英語ではコンマ）または 3 点リーダー（…）で表した。前者より後者の方が長いポーズを表す。イントネーションについては、文末昇調による疑問表現を“?” 記号で表した場合がある。また、話し言葉に特有な以下の項目は文字化したデータの中で括弧記号を使って囲むことで表示した。

- 間投詞は [] で囲む。
- 言い直し、言い淀みは () で囲む。

表 3 日本語の品詞
Table 3 Parts of speech of Japanese.

品 詞 名	品 詞 名
記号	終助詞
形容詞	接尾語
普通名詞	接頭語
サ変名詞	補助動詞
代名詞	固有名詞
数詞	形容名詞
副詞	本動詞
連体詞	間投詞
接続詞	準体助詞
感動詞	並立助詞
助動詞	係助詞
副助詞	慣用句
接続助詞	その他
格助詞	

担：はい、こちら国際コンピューター会議事務局です。

申：[え] トロント大学のノダと申しますが、[えーっと] 十月の[えー] 会議のチケットが届いたんですけども、[えー] どうもありがとうございました。[あー] それですね、ちょっと帰りのチケットの変更をちょっとお願いしたいんですけども、そちらでよろしいでしょうか。

担：[あっ、あー] ノダ先生でございますね、[えーっと] ロボットの[あの] 担当していただくことになっておるノダ先生でよろしいわけですね。[あのですね、えーっと] チケットの変更といたしますと、[あー] 旅行の予定が変わられたということでしょうか。

申：[あっ] はい、[あー] 最初はもうそのまま[えー] トロントに帰るつもりだったんですけど、ちょっとホンコンの方に寄りたいと思うので、[えー] 最終日の三十日の[えー] 夜ぐらいにホンコンに行って、で、[ん] 十一月の二日ぐらいに[えー] (東) ホンコンから東京へ戻って、で、東京からトロントに帰りたいんですけども。

担：[あ] さようでございますか。

担：事務局員

申：参加者

図 1 文字化された電話対話例

Fig. 1 Example of transcribed telephone dialogue.

- ある発話者の 1 文中の発話に相手の発話が割り込む、いわゆる「相づち」は { } で囲む。
- 読みの困難な場合、後に < > で囲んで読みを添える。

図 1 に文字化された例を示す。キーボード対話および手紙による対話はデータ収集時点から文字データである。

3.2 言語単位の認定

文字化されたデータは、言語単位の認定が行われる。ここで用いた言語単位としては、次のものがあり、記号<の左の単位は右の単位より小さい言語単位

表 4 英語の品詞
Table 4 Parts of speech of English.

品 詞 名	品 詞 名
名詞	副詞
数名詞	疑問詞
代名詞	関係副詞
疑問代名詞	THERE
関係代名詞	否定詞
仮主語	前置詞
固有名詞	TO 不定詞の TO
再帰代名詞	等位接続詞
動詞	副詞節従属接続詞
助動詞	名詞節従属接続詞
BE 動詞	間投詞
HAVE 動詞	文中の終止符
形容詞	コンマ
数形容詞	スラッシュ
疑問形容詞	終止符
関係形容詞	記号
限定詞	その他
数量詞	

である。

単語 < 句 (文節, 日本語のみ認定) <
節 (格構造, 日本語のみ認定) < 文 <
発話 < 会話

日本語の単語単位としては, 短単位⁴⁾にはほぼ一致する単位を採用した。ただし, 固有名詞に関しては, 長単位⁴⁾に近い。融合形で, 分け難いものは, 1単位とした。例えば, 「してはいけないよ」が「しちゃいけないよ」になる場合, 「ちゃ」は1単位とした。英語

の場合は, スペースやコンマなどで区切られた部分を単語とした。ただし, 短縮形は元の形に復元したものによって単語を認定している。例えば, “I’ll” は “I” と “ll” の2単位とする。

句単位は日本語の文節単位であり, 原則として

接頭語*+自立語+接尾語*+付属語*

で定義される。*は要素の0個以上の連続を表す。複合名詞は1つの自立語とみなした。付属語は, 助詞, 助動詞, 補助動詞で構成される。英語では句単位を認定していない。

節単位は1つの述語にいくつかの格要素名詞句が係った日本語の構造であり, 格構造単位とも言う。述語が連体修飾をしている場合も, 1つの格構造としている。この場合, 通常, 被修飾格要素が述語の後方に存在する。述語部分に助動詞や補助動詞が存在するときは, それらを含む部分を別の格構造単位とした。例えば,

「申込用紙を 送ら せ て頂きます」

の下線の部分はそれぞれ異なる格構造として認定した。このため, 格構造単位は, 埋め込み構造を持ち得る。英語では節単位を認定していない。

文単位の認定法は前述した。発話単位は, 話者の交替によって認定した。ただし相づちは話者の交替とはみなさなかった。会話単位は, 対話の開始から終了までとした。

3.3 言語単位の属性

認定された言語単位のうち, 単語単位と会話単位に各種属性を付与した。後者に関しては, 前述した会話情報ファイルの内容である。単語単位の属性は以下のものである。

日本語の単語単位の属性とし, 次のものを付与し

表 5 英語の屈折情報

Table 5 Inflections of English.

品 詞	屈 折 情 報
動 詞	原形
	現在形 一人称単数
	現在形 一人称複数
	現在形 二人称単数
	現在形 二人称複数
	現在形 三人称単数
	現在形 三人称複数
	過去形 一人称単数
	過去形 一人称複数
	過去形 二人称単数
	過去形 二人称複数
	過去形 三人称単数
	過去形 三人称複数
	過去分詞形 完了
	過去分詞形 形容
	過去分詞形 受動
	現在分詞形 進行
	現在分詞形 形容
	現在分詞形 動名詞
助 動 詞	現在形 過去形
形 容 詞 副 詞	比較級 最上級 その他
名 詞 固 有 名 詞	所有格 単数 所有格 複数 所有格 その他 その他 単数 その他 複数 その他 その他
関係代名詞	主格 目的格 所有格
代 名 詞	主格 単数 主格 複数 目的格 単数 目的格 複数

表 6 係り受け関係の構文情報

Table 6 Syntactic attributes for word dependency relations.

実線: 係り元, 2重線: 係り先

関 係 名	例
連 用 格	太郎が <u>参加</u> する
連 体 修 飾	太郎の <u>論文</u>
連 体 格	<u>参加</u> した 人
述語連用修飾	<u>正確</u> に <u>登録</u> する
文連用修飾	電車が <u>遅</u> れたので 会議に出席 <u>でき</u> ない
並 列	会議に <u>参加</u> の 人

た。

- 表記 (キーおよび手紙対話は被験者の表記そのもの、電話対話は文字化作業者の表記)
- 読み (ひらがなによる読みの記述)
- 標準表現 (活用を終止形にしたり、異表記を統一した表現 融合形や言い直し・言い淀みで復元できる場合はそれを用いた)
- 品詞 (表 3 に示す 27 個の品詞)
- 活用型 (5 段活用、形容詞型活用など)
- 活用形 (未然形、連用形など)
- 音便形 (イ音便形、促音便形、撥音便形など)

表 3 で「形容名詞」とは助動詞「だ」や「たる」を伴って形容動詞となる名詞である。「サ変名詞」に「する」が付加した場合の「する」は補助動詞である。

英語の単語単位の属性として、次のものを付与した。

- 表記 (キーおよび手紙対話は被験者の表記そのもの、電話対話は文字化作業者の表記)
- 標準表現 (変化形を原形に戻したものや短縮形などを回復したもの 例えば、「I'll」を「I」と「will」に回復する)
- 品詞 (表 4 に示す 35 個の品詞)
- 屈折情報 (表 5 に示す屈折情報)

3.4 単位間の関係と関係の属性

言語単位間の関係として、格・係り受け関係と日英対応関係を記述した。

3.4.1 格・係り受け関係

日本語の単語または単語連続の間の格・係り受け関係の情報を付与した^{15), 16)}。関係の属性としては、構文関係、深層意味関係および表層格関係を設定した。構文関係は格・係り受け関係の構文的情報を示す属性で、表 6 に示す 6 種類がある。深層意味関係は格・係り受けの深層的な意味情報を示す 63 種の属性で、詳細は文献 15) に記述されている。表層格関係は格関係を示す表層格助詞の情報で、語形そのもので表されている。格助詞相当の関係表現¹⁷⁾の一部も格助詞として扱った。表層上、格助詞が存在しないものは潜在格助詞¹⁸⁾を復元して付与した。また、時点性名詞などで格助詞の元来存在しないものは特殊記号で表層格関係を示した。たとえば、「今日 ε お送りします。」

の ε などがある。

3.4.2 日英対応関係

ADD の 1 つの特徴として、日本語の対話文とそれに対応する英語の対話文の両方をデータとしていることがある¹⁹⁾。日本語と英語で同一の意味を担う部分を日英対応関係として付与した。これには、対応する言語単位によって単語対応、文節対応、格構造対応、文対応、発話対応、ランダム対応の区別がある。すべての日英対応において単語対単語列の対応が取れば、単語対応のみを記述しておけばよいが、そのようなこ

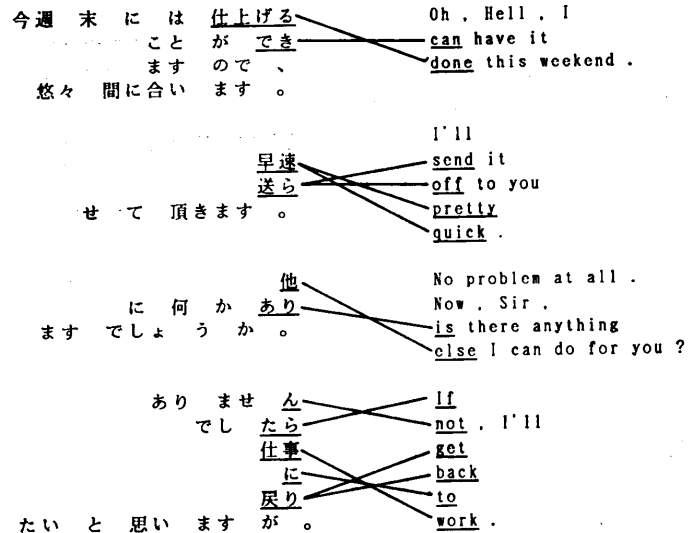


図 2 日英単語対応データ例
Fig. 2 Example of word correspondences between Japanese and English.



図 3 日英文節対応データ例
Fig. 3 Example of phrase (bunsetsu) correspondences between Japanese and English.

とは望めないで、単語より大きい単位の対応も考慮しなければならない。格構造対応の中には、各格要素ごとの対応も含まれる。また、格構造は埋め込み構造を持っているので、埋め込み深さの情報も付与した。イディオムなどで、言語単位間の対応が取りにくいものがある。これらをランダム対応として、対応させた。例えば、

- 早速送らせて頂きます。
- I'll send it off to you pretty quick.

の下線間の対応である。

対応の方向として、日本語から英語への方と、英語から日本語への方がある。日英対応の単位は対応元での単位である。対応元で1単位であるものが対応先では1単位にならない場合がある。例えば、日本語から英語への単語対応において

- 週末
- the end of week

では、「週末」は1単語であるが、対応する英語表現は4単語から成る。英語から日本語への対応には文節対応、格構造対応は存在しない。

図2～図5にこれらの対応のデータ例を示す。

4. ADD の利用

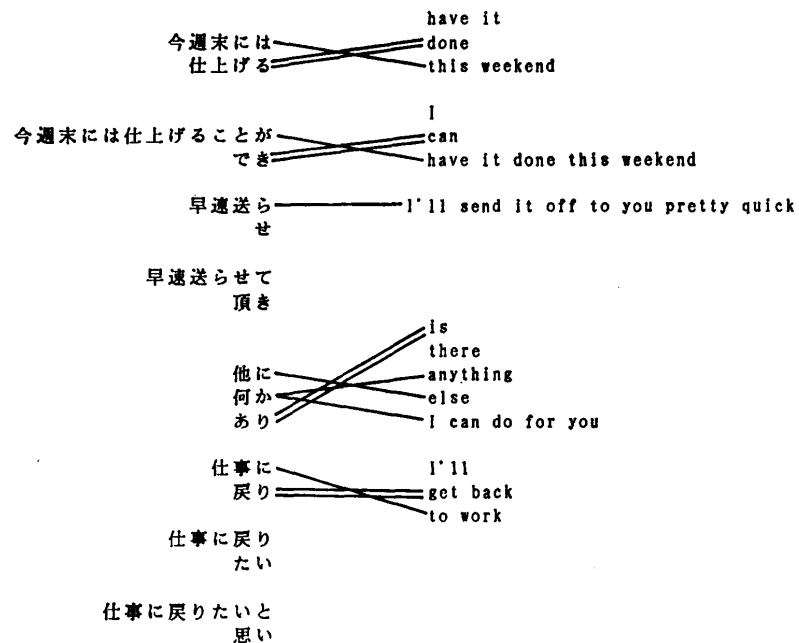
3章で述べた事前分析は ADD の利用に際して、有効である。本章では、この点を事前分析の内容別に述べる。詳細は、文献を参照されたい。

言語単位の認定と、形態素解析結果を利用して、日本語の文節パターンを抽出した。このパターンは文節発声の音声認識のための文節内確率文法の各規則に付随している確率の推定に利用された²⁰⁾。また、文末に出現する接続助詞のパターンを利用して、精度の良い文節間文法の設計が行えた²¹⁾。

係り受け関係を利用して単語間の共起に関する知識ベースを構築した。これを利用して音声認識の候補削減を行った²²⁾。

日英対応関係を利用して、用例主導型翻訳のための知識ベースを構築した²³⁾。これは「名詞+格助詞〈の〉+名詞」の〈の〉の部分英語に訳す場合の曖昧性を用例を用いて解消しようとするものである。

日英対応と係り受け関係を利用して、単語を意味的にクラスタリングする場合の精度の向上が図れた²⁴⁾。通常、単語の係り受け関係のみを用いて、クラスタリングを行うが、多義性による精度の劣化が生ずる。これを日英対応データによって減少させることができた。



2 重線：述語要素の対応
実線：その他の格要素の対応

図4 日英格構造対応データ例

Fig. 4 Example of clause correspondences between Japanese and English.

今週末には仕上げる事ができますので、悠々間に合います。
Oh, Hell, I can have it done this weekend.

早速送らせて頂きます。
I'll send it off to you pretty quick.

No problem at all.

他に何かありますか。
Now, Sir, is there anything else I can do for you?

ありませんでしたら、仕事に戻りたいとおもいますが。
If not, I'll get back to work.

図5 日英文対応データ例

Fig. 5 Example of sentence correspondences between Japanese and English.

各種事前分析にまたがる利用例として以下がある。
2文間の継続関係に関する知識を抽出し、文内の省略要素の補完を行った²⁵⁾。これは、省略要素のない継続する2文のデータに基づいて、省略要素のある入力に対して、その省略部分を推定する手法であり、省略の多い話し言葉の解析に有効である。また、キー対話と電話対話の事前分析結果を統計的に比較し、両者の各種言語現象に関する類似点と相違点を明らかにした²⁶⁾。

5. おわりに

対話に基づいて構築された言語データベース ADD の構築方法とその利用例について述べた。ADD は日本語と英語の対応がある大規模な対話データベースとしては唯一のものであり、また、各種事前分析を施すことによって効率的な利用が可能であるという特徴を持つ。

ADD は関係データベースを拡張した形式で利用されている^{27), 28)}。また、関係データベースの各表をテキスト形式に出力したもので利用可能である。

ADD は、順次整備できたものから、公開することを予定している²⁹⁾。公開に当たっての利用条件は現在検討中である。

謝辞 本研究および発表の機会を与えられ、また有益な助言を与えられた、本研究所 葉原耕平会長に感謝する。本対話データベースの構築は、元研究員橋本一男、井ノ上直己、幸山秀雄、工藤育男の各氏をはじめ、多くの研究員や外部企業の関係者との共同作業による。これらの各氏に深く感謝する。

参考文献

- 1) 大須賀節雄 (監訳): 人工知能大辞典, p. 921, 丸善 (1991).
- 2) Brown University: Brown Corpus, Tech. Report, Brown University (1967).
- 3) Beale, A. D.: Lexicon and Grammar in Probabilistic Tagging of Written English, *Proc. of 26th Annual Meeting of the ACL*, pp. 211-216 (1988).
- 4) 国立国語研究所: 電子計算機による新聞の語彙調査, 秀英出版 (1970).
- 5) 国立国語研究所: 高校教科書の語彙調査, 秀英出版 (1983).
- 6) Svartvik, J. and Quirk, R. (eds.): *A Corpus of English Conversation*, Liber Laromedel Lund (1980).
- 7) 国立国語研究所: 談話語の実態, 秀英出版 (1955).
- 8) 工藤育男, 森元 逞: 対訳対話コーパス作成のためのキーボード会話収録システム, 電子情報通信学会論文誌, Vol. J-75, D-II, No. 4, pp. 749-761 (1992).
- 9) 森元 逞ほか: 自動翻訳電話研究用言語データベースの収集について, 第 36 回情報処理学会全国大会論文集, 4 U-5 (1988).
- 10) 小倉健太郎ほか: 言語データベース収集支援システム, 第 36 回情報処理学会全国大会論文集, 4 U-4 (1988).
- 11) 篠崎直子ほか: 言語データベースの品質管理, 第 36 回情報処理学会全国大会論文集, 4 U-3 (1988).
- 12) 篠崎直子ほか: 言語データベース作成のためのシミュレーション会話, 第 37 回情報処理学会全国大会論文集, 5 B-8 (1988).
- 13) 松村 明 (編): 日本文法大辞典, p. 216, 明治書院 (1971).
- 14) 国立国語研究所: 話しことばの文型(1)(2), 秀英出版 (1960, 1963).
- 15) 井ノ上直己ほか: 係り受け意味関係の問題点とその考察, 電子情報通信学会研究会資料, NLC 88-3 (1988).
- 16) 井ノ上直己ほか: 言語データベース用単語間の関係データ, 第 37 回情報処理学会全国大会論文集, 5 B-7 (1988).
- 17) 首藤公昭, 檜原斗志子: 日本語の文構造のわく組を与える表現, 福岡大学総合研究所報, 第 63 号, pp. 1-43 (1983).
- 18) 平井 誠ほか: 格の強度と述語の構文および意味属性を用いた格構造の変換生成について, 情報処理学会論文誌, Vol. 28, No. 3, pp. 240-249 (1987).
- 19) 小倉健太郎: 言語対比データの構築について, 電子情報通信学会創立 70 周年記念全国大会, No. 1642 (1987).
- 20) 北 研二, 森元 逞: 音声認識システムにおける確率文法の有効性, 第 41 回情報処理学会全国大会論文集, 7 M-2 (1990).
- 21) 保坂順子ほか: 対話データベースを利用した音声認識のための構文規則, 情報処理学会研究会資料, NL-83-13 (1991).
- 22) 柿ヶ原康二, 森元 逞: SL-TRANS における文節候補の削減, 第 39 回情報処理学会全国大会論文集, 4 G-6 (1989).
- 23) 隅田英一郎, 飯田 仁: 用例主導型機械翻訳, 情報処理学会研究会資料, NL-82-5 (1991).
- 24) Inoue, N.: Automatic Noun Classification by Using Japanese-English Word Pairs, *Proc. of the 29th Annual Meeting of the ACL*, pp. 201-208 (1991).
- 25) Kudo, I.: Local Cohesive Knowledge for a Dialogue-machine Translation System, *Proc. of the COLING '90*, Vol. 3, pp. 391-393 (1990).

- 26) Ehara, T. et al.: ATR Dialogue Database, *Proc. of ICSLP '90*, Vol. 2, pp. 1093-1096 (1990).
- 27) 小倉健太郎ほか: 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, NL-69-4 (1988).
- 28) 橋本一男ほか: 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, pp. 245-248 (1989).
- 29) 江原暉将ほか: ATR 対話データベースの内容, ATR 自動翻訳電話研究所テクニカルレポート, TR-I-0186 (1990).

(平成3年6月3日受付)

(平成4年1月17日採録)



江原 暉将 (正会員)

昭和42年早稲田大学第一理工学部電気通信学科卒業。同年, 日本放送協会入社。昭和45年より放送技術研究所勤務。かな漢字変換, 機械翻訳の研究に従事。平成元年(株)ATR自動翻訳電話研究所に出向。対話データベース, 音声言語処理の研究に従事。平成3年日本放送協会に復帰。現在, 放送技術研究所画像研究部主任研究員。電子情報通信学会, Association for Computational Linguistics, 機械翻訳協会各会員。



小倉健太郎 (正会員)

昭和29年生。昭和53年慶應義塾大学工学部管理工学科卒業。昭和55年同大学院修士課程修了。同年日本電信電話公社入社。昭和61年(株)ATR自動翻訳電話研究所に出向。平成2年日本電信電話(株)に復帰。現在, NTT 情報通信網研究所知識処理研究部主任研究員。機械翻訳システム, 自動翻訳電話の研究に従事。電子情報通信学会, 人工知能学会, 計量国語学会各会員。



篠崎 直子 (正会員)

1962年生。1985年神戸大学理学部数学科卒業。同年(株)東洋情報システム入社。1986年9月より1989年3月まで(株)ATR自動翻訳電話研究所に出向。言語データベースの研究開発に従事。1989年4月よりシャープ(株)に勤務。現在, 技術本部情報技術開発センターに所属。



森元 暹 (正会員)

昭和43年九州大学電子工学卒業。昭和45年同大学院修士課程修了。同年電電公社電気通信研究所入所。オペレーティングシステム, データベース検索の研究, 実用化に従事。昭和62年より(株)ATR自動翻訳電話研究所へ出向。音声言語翻訳システム, 特に, 音声言語統合方式, 音声言語翻訳方式の研究を行っている。現在, データ処理研究室長。電子情報通信学会会員。



樽松 明 (正会員)

昭和36年早稲田大学理工学部電気通信学科卒業。同年国際電信電話株式会社入社。以来同社研究所にて, パターン認識, 音声情報処理, 端末システムなどの研究に従事。工学博士。昭和61年より(株)ATR自動翻訳電話研究所に出向。同社代表取締役社長。自動翻訳電話の研究に従事。電子情報通信学会, 日本音響学会, IEEE各会員。