

電話対話データベースの構築

6F-1

江原暉将 小倉健太郎 森元 逞

ATR自動翻訳電話研究所

1. はじめに 従来、書き言葉を主な対象としてきた機械翻訳技術を自動翻訳電話に拡張するためには、話し言葉独特の言語現象の把握が重要である。そのために、これら言語現象を豊富に含んだ大規模な言語データベースが必要となる。そこで、我々は対話データベースADD (ATR Dialogue Database) の構築を進めてきた。本文では、ADDの概要、データの収集状況について述べる。ADDを利用した各種言語現象の抽出については、他の講演で述べる[小倉5][井ノ上3][橋本3]。

2. 収集対象 一口に話し言葉と言っても、いろいろなものがある。ここでは、自動翻訳電話の研究目的から、「対話」を対象としている。また、「世間ばなし」的なおしゃべりではなく、情報の伝達や行為の要求などの明確な目的を持った「目的指向型」の対話を対象にした[森元]。伝達メディアとしては、電話会話が主体であるが、書き言葉を用いた対話として、キーボード会話および手紙による会話データも収集している。実際の会話を収集することは、通信の秘密の問題があるので、現在は模擬会話実験によってデータを収集している[小倉][篠崎][篠崎2]。模擬実験を行う時の会話内容として、現在は、・国際会議の申込に関する参加者と事務局の対話・旅行に関する旅行会社と客の対話の2つのタスクを選択している。会話データの他に、比較のための書き言葉データとして、新聞のデータも若干収集している。図1に最近の収集状況を示す。

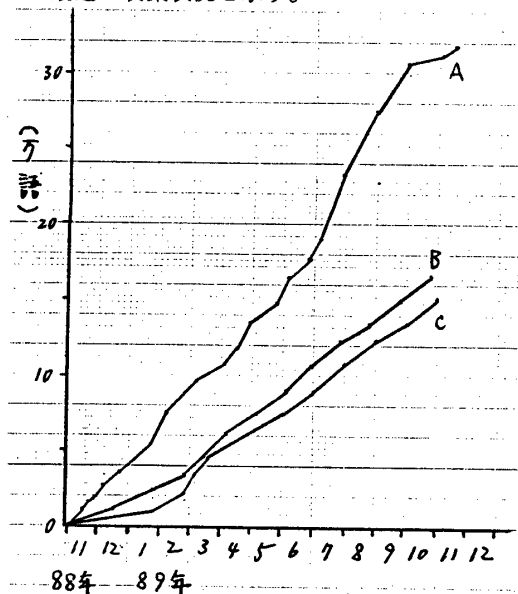


図1 最近のデータ収集状況

図中、A、B、Cは文字化済み、係り受け付与済み、日英対応付与済みの各データを示す。

3. 付与情報 模擬実験によって収集された電話会話データは、まず、テープの聞き取りによって文字化される。キーボードデータなどは、収録当初から文字化されている。次に、各種の情報が付与される。情報を付与するに当たって、付与する言語単位を決めなくてはならない。それらは表1の第1列に示すものを用いた。下の方ほど大きな単位である。以下これらの単位を説明する。

表1 付与情報

言語単位	分割	属性	関係	日英対応
文字	○			
形態素	○	○	○	○
句	○			○
節	○			○
文	○			○
発話	○			○
会話	○	○		○

①文字単位：話し言葉を対象としているので「文字」よりも「音韻」単位の方が望ましいが、録音テープを文字化したものを原データとしているので、文字を最小単位とした。日本語の場合は表記と読みの文字であり、英語の場合は、表記の文字のみである。韻律情報のうち、イントネーションについては、基準が明確でないので、現状では付与していない。ポーズについては、長さに応じて、読点(コンマ)、句点(ピリオド)またはリーダー(．．．)で表現している。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に特殊記号を使って示した。・間投詞は「」で囲む。・言い直し、言い淀みは()で囲む。・英語の場合、数字の後に<>で囲んで読みを添える。・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」などは{ }で囲む。図2に文字化された例を示す。②形態素単位：短単位[国研]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[国研]に近い。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形態素とした。③句単位：日本語では文節であり、

(接頭語) + 自立語 + (接尾語) + (付属語*)

である。()は任意選択、*は0個以上の任意選択を表す。英語では句単位で分割していない。④節単位：1つ

の述語にいくつかの格要素名詞句が係った構造であり、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、日本語では、被修飾格要素が述語の後方に存在する。⑤文単位：発話意図の1まとまりを文単位とした。文字化に当たって、韻律情報・文型も利用して、文の認定をし、句点（ピリオド）で文の切れ目を表示した。判断しがたいときは、文の途中であるとした。⑥発話単位：話者の交替によって発話単位を分割した。ただし相づちは話者の交替とは見なさなかった。⑦会話単位：電話会話の開始から終了までとした。

申込者：[あ] もしもし。

担当者：はい。

申込者：[あの一、えーと] 団体旅行で {はい}

[あの一] テニスのサークルの合宿で {はい}

[えーと] 男女25人ぐらいの {はい}

グループで {はい} [えーと] 合宿で軽井沢の方に行きたいんですけど。

担当者：はい

申込者：そちらの申込みの方でお電話したんですけど。

担当者：[あ] はい、ありがとうございます。

図2 文字化の例

言語単位には、各種属性が付与し得る。形態素単位の属性として、次のものを付与した。・表記 ・読み ・標準表現（活用を終止形にしたり、異表記を統一した表現） ・品詞 ・活用型 ・活用形 ・音便形。品詞体系としては、31個の品詞カテゴリーを用いた[吉本]。発話者の性別や年齢などを、会話単位の属性と見なして付与した。

次に、言語単位間の関係として、格・係り受け関係を記述した。ここでは、日本語の形態素の間の構文関係と意味関係を付与した[井ノ上] [井ノ上2]。

本言語データベースの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[小倉2]。日英対応関係として、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 ・ランダム対応に分けた。全ての日英対応において単語対単語の対応が取れば、単語対応のみを記述しておけばよいが、そのようなことは望めないもので、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、各格要素毎の対応も含まれる。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。

図3に文節対応のデータ例を示す。

4. 統合管理システム 収集した対話データベースを統合管理するために、関係データベースを拡張した形にデータを収容した[小倉3]。また、ユーザーインターフェースとして、フレーム表現のデータ構造を持つインタ

ーフェースを実現した[橋本] [小倉4] [橋本2]。

5. おわりに ATRで構築している電話対話に基づく言語データベースについて述べた。今後、このデータベースが広く利用されることを期待する。

文献

[井ノ上] 井ノ上直己ほか：係り受け意味関係の問題点とその考察、信学研、NLC88-3、1988。

[井ノ上2] 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、情全大、37回、5B-7、1988。

[井ノ上3] 井ノ上直己ほか：係り受け関係データから見たキーボード会話と電話会話の比較、情全大、40回、1990。

[小倉] 小倉健太郎ほか：言語データベース収集支援システム、情全大、36回、4U-4、1988。

[小倉2] 小倉健太郎：言語対比データの構築について、信学界創立70周年記念全大、1642、1987。

[小倉3] 小倉健太郎ほか：言語データベース統合管理システム、情全大、37回、5B-6、1988。

[小倉4] 小倉健太郎ほか：言語データベース統合管理システム、情研資、NL69-4、1988。

[小倉5] 小倉健太郎ほか：慣用表現を利用した形態素情報収集法、情全大、40回、1990。

[国研] 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。

[篠崎] 篠崎直子ほか：言語データベースの品質管理、情全大、36回、4U-3、1988。

[篠崎2] 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、情全大、37回、5B-8、1988。

[橋本] 橋本一男ほか：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ーフェース、情全大、37回、2C-4、1988。

[橋本2] 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、1989。

[橋本3] 橋本一男ほか：対話データベースを用いた各種言語現象の検索、情全大、40回、1990。

[森元] 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、情全大、36回、4U-5、1988。

[吉本] 吉本 啓：日本語品詞の分類、ATRテクニカルレポート、TR-I-0008、1987。

今週末には	Oh, Hell,
仕上げる	I can
ことが	have it done
できますので、	this weekend.
悠々	
間に合います。	
早速	I'll send it off to you
送らせて頂きます。	pretty quick.
	No problem at all.
	Now, Sir,
他に	is there
何か	anything
ありますでしょうか。	else
	I can do for you ?
ありませんでしたら、	If not,
仕事に	I'll get back
戻りたいと	to work.
思いますが。	

図3 日英文節対応データ例