旅行会話基本表現に対する日本語パラフレーズデータの収集

金城 由美子* 青野邦生 安田圭志 竹澤寿幸 菊井玄一郎

ATR音声言語コミュニケーション研究所

1. はじめに

ホテル予約など基本旅行会話における音声翻訳システムの有効性は、対話実験により実証されている [1]。今後、音声翻訳システムがより広範な場面での会話を扱うためには、基本的な表現だけでなく、複雑ないいまわしなど多様な表現に対応していかなければならない。今後の音声翻訳システムおよび関連技術の研究には、言語表現の多様性を反映したコーパスが必須といえる。本稿では、多様な言語表現を含むコーパス作成の試みとして、現在構築中の旅行会話基本表現パラフレーズデータベースの概要について報告する。

多様な表現を、効率よく網羅的に収集する方法の一つとして「言い換え」が考えられる。本データベースは、セル形式言語データ収集法により収集された、日本語パラフレーズデータから構成される。パラフレーズデータは人手で作成されたものであるが、セル形式の利用により、1つの種文から多数のパラフレーズ文が得られ、表現の抜けも少なく、作業者の負担も少ないことが、テストデータ収集で確認されたのを受け、大規模なデータ収集を行った。

さらに、言語モデルへの適用や翻訳知識、パラフレーズ知識の自動獲得等のデータ利用に向けて、セル形式をXML形式に変換することにより、パラフレーズデータの管理に利便性を図っている。

2. データベースの構築

ATRにおいて作成した約20万の日英対訳文からなる旅行会話基本表現集 [2] に含まれる表現 約10,000ペアをオリジナルデータとして、日本語パラフレーズデータベースの構築を行った。データベース構築は、パラフレーズデータの収集、クリーニング、フォーマット変換の三段階を経て行われた。以下、それぞれについて述べる。

*E-mail: yumiko.kinjo@atr.co.jp

2.1 パラフレーズデータ収集

パラフレーズデータの収集は、人手によるパラフレーズ文の作成によって行った。それぞれ約500組の日英文を含む20ファイルをパラフレーズ元として使用し、6名の作業者が約3ヶ月間作業を行った。1つのオリジナルデータに対し、1名がパラフレーズ作業を行い、作業の重複はなかった。作業者は、いずれも英語の堪能な日本語話者であった。

パラフレーズ作業者には、オリジナルデータである日英文（以下、種文と呼ぶ）を参照してもらい、日本語文の厳密な言い換えや、英文の翻訳ではなく、状況を考えてそれにふさわしい自然な日本語表現を網羅的に記述してもらうよう指示を行った。テストデータ収集時には、パラフレーズ文が種文の影響を受けるのを避けるため、英文のみ参照としたが、文脈がとらえにくく英文解釈の誤りなどが生じたため、日英文を参照するようパラフレーズの条件を変更した。ただし、種文の場面、文脈等に関する情報は与えず、種文が使われるであろう一般的な場面を想定してもらった。

方言や、年齢や性別に特有な表現や、語用論的な省略は、パラフレーズの対象としないよう指示を行った。また、語順の転換によるパラフレーズも、規則による生成が可能なものと考え、対象としなかった。

セル形式による記述

パラフレーズの記述には、[3, 4] で提案されたセル形式を利用した[1]。テストデータ収集により、セル形式言語データ収集法は、表現の抜けが少なく、作業者の負担も少ないことが確認されている [3, 4]。基本的な作業方針はテストデータ収集の方針をひきつぎ、次のようなものとした。

- 1行は1文に相当するものとして扱う
- 言い換え可能な表現は、同一のセルに記述を行う

しかし、セルや行の具体的な分割基準は特に定めず、作業者の記述のしやすさを優先し、自由に記述しても

[1]EXCELを利用。

225	How is it going?			
225	調子どう。			
	調子 具合	は	どう いかが	X ですか
	調子 具合	は	どうだい	
	元気	かい ですか		
226	Tea with lemon, please			
226	レモンティーをください。			
	レモンティー	で を	GOBIC	
	レモンティー	を にして	GOBIA	
	レモンティー	を	いただき もらい 飲み 頼み	ます

表 1: セル形式データ

らった。表1にセル形式によるパラフレーズ記述の例を示す。

セル中の語または句をセルフレーズと呼ぶ。セルフレーズは語に相当することが多いが、言い換えが可能なら、統語的な語や句の単位にあてはまらない要素であっても構わないものとした。

「調子はいかがですか」と「元気かい」のように文の構成要素が大きく異る場合は、セルフレーズの追加ではなく、それぞれの文を異なる行に記述する。これらを種文に対し、苗文と呼ぶ。

表1の「調子はどう。」という種文は、3つの苗文にパラフレーズされている。左から右へと、各セルにつきセルフレーズを一つ一つ組み合わせて行くことでパラフレーズ文が表現できる。またセル中のXは空白記号として使用し、省略的な表現にも対応可能とした。最初の苗文は、セルフレーズが 2*2*1*2=8 で、「調子はどう」「調子はどうですか」「具合はどう」「具合はどうですか」など8通りのパラフレーズ文を表現できる。

「レモンティーをください。」の例では、GOBIAやGOBIBなどの記号が使用されている。日本語には、丁寧さなどに応じて「～ください」「～くださいますか」「～くださいませんか」など多数の文末表現が存在するので、言い換え可能な文末表現は、別表にまとめ、GOBIA等の記号で代用した。代用表現の使用により、数多く存在する文末表現の抜けをなくすと同時に、作業者の負担を減らし、効率的にパラフレーズ作業を行うことができる。

パラフレーズ作業の過程で、英文が重複するなどの理由で作業対象外とされたものが約15%程度あり、約8,500文に対するパラフレーズが得られた。

2.2 クリーニング

セル形式によるパラフレーズデータの一部[2]は、約1ヶ月間をかけ、10名の作業者によりクリーニング作業が行われた。クリーニング作業は言語学専攻の大学院生に依頼した。

クリーニングでは、パラフレーズ作業者の違いによるセル分割方法の差異などを解消するために、セル、苗文の統合・分割などが行われた。クリーニング時には、セルは基本的に語の単位で分割するよう指示をし、ある程度のデータの標準化を行った。同時に、パラフレーズ文の修正・追加なども行われた。この作業は、パラフレーズ作業者とは異る作業者が、クリーニング作業とクリーニング結果のチェック作業を2人1組で行った。

パラフレーズ作業は1名の作業者が行ったが、クリーニング時に2名の作業者が表現の修正・追加を行ったため、精度の高い、網羅的なデータが得られたと思われる。

また、数字や、アルファベットの表記の統一、種文が長い発話で分割された場合の枝番号の付与、などは全てのデータにクリーニングが施された。この作業は、パラフレーズ作業、クリーニング作業者とは異なる作業者1名が行った。ここでも、パラフレーズ文の明らかな間違いなどは修正が行われた。

2.3 フォーマット変換

クリーニング段階までの作業は、パラフレーズ文の見通しのよいセル形式で行われたが、セル形式は統計的な処理や、データの加工などを行うには適切とはいえないため、データフォーマットの変更を行った。セル形式をテキスト形式に変換し、全てのセルフレーズを組み合わせてパラフレーズ文への展開を行い、完成したパラフレーズ文は種文ごとにXML形式で保存することにした。

テキスト形式

セル形式は、パラフレーズ作業には都合が良かったが、データの加工や利用等に扱いやすい形式とはいえない。perlスクリプトを利用し、データ加工などに都合の良いテキスト形式に、フォーマットの変換を行った。

表2にテキスト形式データの例を示す。テキスト形式データは、基本的に1行が1つの苗文に対応している点はセル形式と同じである。セルの代わりにタブ区

[2]約60%。

```
225 How is it going?
225 調子どう。
    調子||具合　は　どう||いかが　X||ですか
    調子||具合　は　どうだい
    元気　かい||ですか
```

表 2: テキスト形式データ

切りを使用、セルフレーズはセパレータとして "||" を使用し、セル形式データと等価な情報を保持している。

XML 形式

全てのセルフレーズを組み合わせ、展開したパラフレーズ文は、データベースとして利用するため、種文等の情報を含む XML 形式とした。XML 形式データは、テキスト形式データを perl スクリプトにより展開して作成したものである。

表 3 に XML 形式データの例を示す。パラフレーズ文が膨大な数に上る場合もあるため、1 組の種文につき 1 つの XML ファイルを作成した。XML ファイルは、種文の含まれるファイルごとのディレクトリを作成し、データの階層化を図った。

```
<?xml version="1.0" encoding="euc-jp"?>
<UTTERANCE TID="" UID="225">
<JTEXT>調子どう。</JTEXT>
<ETEXT>How is it going?</ETEXT>
<PP_UNIT ID="0">
<JPP_GRP ID="1">
<JPP ID="1"><JT>具合はいかが</JT></JPP>
<JPP ID="2"><JT>具合はいかがですか</JT></JPP>
<JPP ID="3"><JT>具合はどう</JT></JPP>
<JPP ID="4"><JT>具合はどうですか</JT></JPP>
<JPP ID="5"><JT>調子はいかが</JT></JPP>
<JPP ID="6"><JT>調子はいかがですか</JT></JPP>
<JPP ID="7"><JT>調子はどう</JT></JPP>
<JPP ID="8"><JT>調子はどうですか</JT></JPP>
</JPP_GRP>
<JPP_GRP ID="2">
<JPP ID="1"><JT>具合はどうだい</JT></JPP>
<JPP ID="2"><JT>調子はどうだい</JT></JPP>
</JPP_GRP>
<JPP_GRP ID="3">
<JPP ID="1"><JT>元気かい</JT></JPP>
<JPP ID="2"><JT>元気ですか</JT></JPP>
</JPP_GRP>
</PP_UNIT>
</UTTERANCE>
```

表 3: XML 形式データ

XML 形式データでは、文番号や、種文 (JTEXT,ETEXT) だけでなく、同一の苗文から生まれたパラフレーズ文について、グループ化によって明示している (JPP_GRP)。

3. パラフレーズデータの概要

約 8,500 文の種文から、セル形式言語データ収集法により集められたパラフレーズデータを展開した結果、2,300 万文あまりのパラフレーズ文が得られた。苗文は、構成要素としてセルおよびセルフレーズを含むので、それらの総数と共に表 4 に示す。

種文	8,505
苗文	39,227
セル	245,660
セルフレーズ	460,580
1 苗文の平均セル数	6.26
1 苗文の平均セルフレーズ	11.74
パラフレーズ文	23,366,634

表 4: パラフレーズ結果

1 つの苗文に含まれるセルの数は最小が 1、最大が 38、平均 6.26 であった。種文となった旅行会話基本表現集の日本語 1 文の平均の語数は、6.87 であり [5]、セルフレーズは述語部分に複合的な表現が含まれることが多いことを考慮すると、妥当な数値であるといえる。1 苗文あたりのセル数の分布を図 1 に示す。

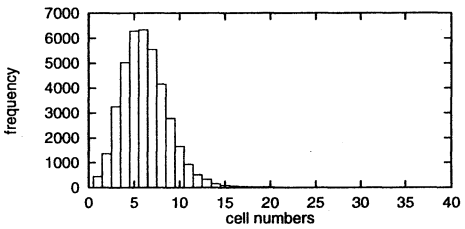

図 1: 1 苗文あたりのセル数

種文に対するパラフレーズ文の数を拡大率と呼ぶ。全データの平均拡大率は、約 2,700 であった。最小拡大率は 1、最大拡大率は 610,304 であった。1 つのセルに含まれるセルフレーズの数をセル拡大率と呼ぶ。セル拡大率は 1.87 であった。2,700 という拡大率はテストデータ収集時 [3] の拡大率 436.9 と比べても非常に高いが、セル拡大率のみでこれを説明するのは難しい。

種文から苗文への拡大率は 4.61、苗文からパラフレーズ文への拡大率は 595.68 という点から考えると、文末表現が高い拡大率をもたらしている可能性が考えられる (表 5 参照)。苗文の約 20%に文末表現記号が使用されており、今回使用した 3 種類の GOBI ファイルには、

拡大率	2747.40
セル拡大率	1.87
種 → 苗	4.61
苗 → パラフレーズ	595.68

表 5: 拡大率

計 68 の文末表現が収録されている。やはり、高い拡大率は文末表現の働きによるところが大きいといえる。

また、セル形式を利用してのパラフレーズ作業、クリーニング作業を通じて、各データに対して少なくとも 2 名、多い場合は 4 名の作業者が表現の抜けなどのチェックを行っており、この点も高い拡大率に寄与していると思われる。

4. おわりに

本稿では、旅行会話基本表現に対する日本語パラフレーズデータベースの構築過程と、収集したパラフレーズデータの概要について述べた。セル形式言語データ収集法により、パラフレーズデータを効率的に集めることができ、非常に高い拡大率が得られた。拡大率の高さは、豊富な文末表現によるところが大きく、文末表現の拡張や、文末以外の表現にも同様の言い換え表を作ることで、拡大率の増加を図れるものと思われる。このようなセルフレーズの拡張には、シソーラスなどの利用も考慮していく必要がある。

また、セルフレーズの拡張だけでなく、苗文の活用について考えていく必要がある。苗文は、種文とパラフレーズ文の橋渡しをする存在となっているが、苗文間の関係など、その特徴については不明な点が多い。今後は、苗文の分析を含めた、パラフレーズデータの詳細な分析によりパラフレーズ知識の獲得を目指す。

また、パラフレーズデータを言語モデルや、言語翻訳に適用した場合の効果を明らかにすることが今後の課題である。

謝辞 データ作成に貢献いただいた津山佳子氏をはじめとする作業者の方々に心から感謝いたします。

本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである。

参考文献

[1] 菅谷史昭, 竹沢寿幸, 隅田英一郎, 匂坂 芳典, 山本誠一. 音声翻訳システム: ＡＴＲ-ＭＡＴＲＩＸの開発と評価. 情報処理学会論文誌, Vol. 43, No. 6, pp. 2230–2241, 7 2002.

[2] Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. Third International Conferencse On Language Resources and Evaluation (LREC 2002)*, Vol. I, pp. 147–152, 2002.

[3] 菅谷史昭, 金城由美子, 竹沢寿幸, 菊井 玄一郎, 山本誠一. 大規模言語データベース収集法の一提案. 情報処理学会第 64 回（平成 14 年）全国大会講演論文集 (2), pp. 2–81–82. 情報処理学会, 2002.

[4] Fumiaki Sugaya, Takezawa Toshiyuki, , Genichiro Kikui, and Seiichi Yamamoto. Proposal of a very-large-corpus aqcuisition method by cell-formed registration. In *Proc. Third International Conference On Language Resources and Evaluation (LREC 2002)*, Vol. I, pp. 326–328, 2002.

[5] 竹沢寿幸, 菊井玄一郎, 鈴木弥生, 西野 敦士. コーパスベース音声翻訳研究のための対話データ収集. 音声言語情報処理, Vol. 45, No. 12, pp. 71–76, 2003.