Internal Use Only（非公開）

TR－SLT－0036

旅行会話基本表現集に基づく日本語パラフレーズデータベースの構築

Constructing A Japanese Paraphrase Database based on Bilingual Basic Expression Corpus

金城 由美子　Yumiko Kinjo
津山 佳子　Yoshiko Tsuyama
竹澤 寿幸　Toshiyuki Takezawa
菊井 玄一郎　Genichiro Kikui

2003年3月25日

概要

本稿では、セル形式言語データ収集法により収集した日本語パラフレーズデータベース構築について報告する。二度行われたパラフレーズ収集およびデータクリーニング作業、その後行われたフォーマット作業とデータベースの概要を述べる。

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2 TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax :+81-774-95-1308

目 次

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構築した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。「言い換え/パラフレーズ」の応用は、音声翻訳システムのみでなく、文書校正支援、自動要約、質問応答など多岐に渡り、自動言い換え技術は自然言語処理技術の中核技術になるものとされている[2]。今後の自然処理研究において、パラフレーズの収集・分析は非常に重要な役割を果たすと思われる。

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ＡＴＲにおいて作成した旅行会話基本表現集の表現から、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。

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、各節でそれぞれにについて述べた後、パラフレーズデータの概要について説明する。

2　パラフレーズデータ収集

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[3]に含まれる表現をオリジナルデータとして、人手によるパラフレーズ文の作成を行った。パラフレーズデータの収集は、2度行われた。まず各回での収集の過程について延べ、それからセル形式によるデータ記述について説明を行う。

2.1　2001年4月パラフレーズデータ収集

それぞれ約500組の日英文を含む3ファイルset01,set02,set03(test set)をパラフレーズ元として使用し、1ファイル各3名、延べ9名の作業者が2001年4月から約3ヶ月間作業を行った。set01のみ、作業者が互いの作業内容の参照を行ったが、残る2ファイルについては独立に作業を行った。

パラフレーズ作業者には、オリジナルデータ（以下、種文と呼ぶ）の英文を参照してもらい、単なる翻訳ではなく、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。パラフレーズ文が種文の日本語表現の影響を受けるのを避けるため、英文のみ参照とした。英文は英語母語話者により校正済みのものを使用した。

作業者には種文の場面、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。また、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。語順の変更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。

その他、種文の日本語表現がわかりにくいなどの理由でパラフレーズ作業に適さないと判断されたものには「作業対象外」と記し作業対象から外した。また、パラフレーズ作業対象となった英文と全く同じ英文が後に出現する種文に含まれた場合[1]、重複する英文を“no sentence”に置換し、作業対象外であることを示した。

作業者への指示を付録Aに示す。

2.2　2001年9月パラフレーズデータ収集

それぞれ約500組の日英文を含む20ファイルset01(dev set),set001～set019(learning set)を種文として使用し、6名の作業者が2001年9月から約3ヶ月間作業を行った。うち5名が前回のパラフレーズ作業者であった。今回は、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。

2001年9月のパラフレーズ作業は、作業時期により前期・中期・後期に分けられる。パラフレーズ作業に関する指示の変更、英文修正の有無などで若干作業過程が異なるが、種文の場面、文脈等に関する情報、方言などのパラフレーズ対象外の表現については、全ての作業期間にわたって、2001年4月と同様の指示を行った。

前期の作業では、当初2001年4月と同様パラフレーズ文が種文の表現の影響を受けるのを避けるため、英文のみ参照とした。その後、全体の3割程度の英文が校正作業者による修正[2]を受けたため、日英文を参照してパラフレーズ文の見直しが行われた。

[1]日本語表現が異なっても、英文が同じ場合がある。
[2]修正は他の目的のために行われた。

英文のみを参照した場合、文脈がとらえにくく英文解釈の誤りなどが生じたため、中期・後期ではパラフレーズの条件を変更した。種文の日英文両方を参照してもらい、日本語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。

中期では、データの3割程度が、英文が不自然、または日本語文との組み合わせに問題があると判断され、いったんパラフレーズ作業を行わない「保留」とされた。「保留」分は、校正作業者によって修正が行われた後、パラフレーズ作業が行われた。

後期は、校正済みのデータを用いてパラフレーズ作業が行われた。

英文の修正が行われた場合もしくは修正の必要がないと判断された場合、英文の末尾に "#checed by R" など作業者を示すタグが付加される[3]。英文が修正される場合、元の英文には手が加えられず、修正された英文が元の英文の下に書き加えられる[4]。

付録B に前期・中期・後期の作業時期とパラフレーズ作業対象ファイル、作業者の対応表を付す。

2.3 セル形式による記述

パラフレーズの記述には、[4, 5] で提案されたセル形式を利用した[5]。2001 年4 月のパラフレーズデータ収集により、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 [4, 5]。基本的な作業方針は、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

ただし、パラフレーズのしやすさを最優先し、セルや行の具体的な分割基準は特に定めず、作業者に自由に記述してもらった。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

[3] 2 回修正が行われた場合 "#checked by RD" などとなり、後者が最終校正者のイニシャルとなる。校正作業者はR,D,Y の3名で、R,D は英語母語話者。

[4] 誤って修正前の英文が削除されたものもある

[5] 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

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た。使用した記号と文末表現の対応は付録 C を参照のこと。

3　クリーニング

収集したパラフレーズデータは、主にセル分割に関わるクリーニングと、表記に関わるクリーニングが施された。前者は一部データのみに行われたが、後者は全データに対し行われた。

3.1　セル分割に関わるクリーニング

収集したパラフレーズデータの約60%は、単語分割、パラフレーズ修正のクリーニング作業が行われた。この作業は、言語学専攻の大学院生に依頼した。パラフレーズ作業者との重複はなかった。

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、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。

2001年4月収集のパラフレーズは、set01の500文のみを2名の作業者に2001年8月からの約1ヶ月修正・確認を行ってもらった。

2001年9月収集のパラフレーズは、前期分、中期保留分を除く、中期、後期分について、set01のクリーニング作業者2名を含む10名に2002年4月からの約1ヶ月修正・確認を行ってもらった。作業方法・内容は2001年4月分と同等である。

クリーニング対象ファイルと作業者の対応など詳細については、付録Bを参照されたい。

2001年9月の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。

3.2　表記に関わるクリーニング

数字や、アルファベットなど表記の統一や、文番号の管理などについては、全てのデータにクリーニングが施された。

種文の英文が長すぎる場合、パラフレーズ作業者により分割が行われることがある。分割が行われる場合は'|'が区切りとして使用される。その場合、表2のように枝番号を付ける。分割なしにパラフレーズが行われた場合は—0 を、分割された部分に対しパラフレーズを行われた場合は—1,—2 など元の英文の出現順序に従って枝番号をつけた。

448	My name is Hamashima. | I have a reservation.
448	浜島ですが、予約をしています。
448—0	My name is Hamashima. | I have a reservation.
	予約をした浜島ですが
448—1	My name is Hamashima.
	浜島と申します
448—2	I have a reservation.
	予約をしています

表 2: 枝番号の例

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。表記に関わるクリーニングの際も、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。

4 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。今後のデータの加工、利用を想定して、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3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

```
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3: テキスト形式データ

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4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。データの階層化を図るため、各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に収められた。

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)。

```
<?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 4: XML 形式データ

5 パラフレーズデータの概要

本節では、2001 年 9 月に収集したパラフレーズデータの概要について述べる[6]。

作業対象外などを除く約 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 5 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 5: 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 [6]、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。

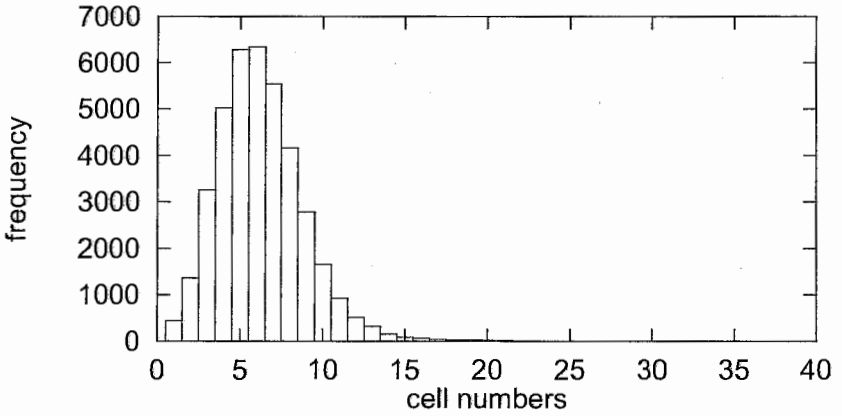

図 1: 1 苗文あたりのセル数

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 2001 年 4 月分収集分 [4] 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 6 参照)。苗

[6]2001 年 4 月収集パラフレーズについては、同一のファイルに対し、作業者ごとに異なるパラフレーズ作業ファイルが存在し、set01 を除き統合作業が行なわれていないのでここでは取り上げない。

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 6: 拡大率

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、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。

6　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。

言語表現の多様性の反映を目指すという点では、パラフレーズデータベースは、自動翻訳システムを用いた対話実験による日英対話データ [7] と同じ目標を持つと思われる。今後両者の比較・分析を行っていく必要がある。

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も今後の課題である。

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] 佐藤理史. なぜ言い換え/パラフレーズを研究するのか. 言語処理学会第7回年次大会ワークショップ論文集. 言語処理学会, 2001.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 菊井玄一郎, 竹沢寿幸, 鈴木弥生, 西野 敦士. 自動翻訳システムを用いた日英対話データの収集. 言語処理学会第9回年次大会発表論文集, pp. 529–532. 言語処理学会, 2003.

付録Ａ　作業者への指示

日本語表現収集について
目的と方法

目的

ＡＴＲでは、日本語音声を認識するシステムを開発しています。話した日本語を認識してくれるようなシステムを作りたいわけですが、そのためには、あらかじめ認識すべき日本語表現をコンピュータに覚えさせる必要があります。その日本語表現を収集するのが今回の目的です。

方法

収集する日本語表現は、旅行で使う表現に限っています。そこで、市販の旅行英会話の書籍を利用することにしました。ひとつの方法として、書籍に掲載されている和訳をコンピュータに入力することが考えられます。

ただ、書籍には訳が大抵ひとつしかありません。ところが、実際には、ひとつのことを言うのにいろいろな表現が可能です。その可能な表現をすべて集めるのが今回の主旨です。

掲載されている和訳に影響を受けないために、敢えて英語の方を見てもらいます。そうすると、英語をそのまま訳したのでは必ずしもこと足りず、このような内容を日本語なら普通どう言うかということを考えてもらわないといけないということになります。

① 与えられた英語表現を読み、状況を考えつつ、内容を理解してもらいます。
② 同じ状況において、日本語で言うなら普通どのように言うかという視点で、自然な普通の日本語表現をすべて網羅するつもりで出していってもらいます。

日本語入力方法

留意事項

＜１＞ セルによる文の分断

文のどこで区切るかについて、特にルールはありません。

＜２＞ セルに語句がなくても可能

「セルに語句がなくても可能」であるときは、半角大文字エックス（ Ｘ ）を。

＜３＞備考

何か説明を入れたいときは、一番右のセルに入力。

最初に半角シャープ（ ＃ ）を。

＜４＞語順の入れ替えが可能（セルの交換可能）

一番右のセルに入力。「#」及びアルファベットは、半角英数入力で。

例： #BC 交換　　セルＢとセルＣとの交換が可能。

　　#B(CD)交換　　セルＣとセルＤを離さずセットにして、セルＢと交換が可能。

＜５＞ 複数の文がある場合

◇自然な日本語を考えて、英語と一対一対応する場合（語順も日本語と英語が同一。）

　　第一文の英語日本語、第二文の英語日本語・・・というふうに文ごとに。

◇日本語が英語と一対一対応しない場合

　　長くなければ、文ごとに切らずに続けて入力してください。ただ、長くなって見づらくなると、切るなど適当にしてください。

付録 B ２００１年９月パラフレーズおよびクリーニング作業詳細

	パラフレーズ作業開始時期	文番号	作業担当者名	
			パラフレーズ	クリーニング（チェック）
前期	2001年	set01 1～508	P1	＜作業なし＞
	9月10日～	set001 1～220	P1	
	10月7日	set005 1～240	P2	
		set007 1～480	P3	
		set010 1～360	P4	
		set012 1～240	P5	
		set013 1～508	P6	
		set014 1～140	P6	
中期	10月8日～	set001 221～508	P1	C1 (C5)
	11月11日	set002 1～508	P1	C2 (C3)
		set003 1～187	P1	C3 (C2)
		set005 241～508	P2	C4 (C8)
		set006 1～199	P2	C5 (C1)
		set007 481～508	P3	C5 (C1)
		set008 1～508	P3	C6 (C9)
		set009 1～132	P3	C7 (C9)
		set010 361～508	P4	C7 (C9)
		set011 1～359	P4	C6 (C10)
		set012 241～508	P5	C8 (C4)
		set014 141～508	P6	C9 (C7)
		set015 1～508	P6	C2 (C3)
		set016 1～48	P6	C4 (C8)
		set019 161～227	P5	C1 (C5)

後期	11月12日～ 12月23日	set003 188～508	P1	C3 (C2)
		set004 1～508	P1	C10 (C6)
		set006 200～508	P2	C5 (C1)
		set009 133～508	P3	C7 (C9)
		set011 360～508	P4	C6 (C10)
		set016 49～508	P6	C4 (C8)
		set017 1～508	P6	C8 (C4)
		set018 1～508	1～20 P1, 21～130 P2, 131～310 P3, 311～508 P4	C9 (C7)
		set019 1～161, 228～508	1～160 P4, 228～508 P5	C1 (C5)

・P4を除く５名が２００１年４月分のパラフレーズ作業者
・C3, C６の２名が２００１年４月分のクリーニング作業者

付録 C　GOBI ファイル

依頼表現 置き換え記号

GOBIA	GOBIB	GOBIC	
ください	ください	お願い	します
くださいますか	くださいますか		いたします
くださいますでしょうか	くださいますでしょうか		できますか
くださいませんか	くださいませんか		できますでしょうか
くださいませんでしょうか	くださいませんでしょうか		できませんか
くださらないですか	くださらないでしょうか		できませんでしょうか
くださらないでしょうか	くださるでしょうか		できるでしょうか
くださるでしょうか	いただけるでしょうか		できないでしょうか
もらえますか	いただけますか		したいのです
もらえますでしょうか	いただけますでしょうか		したいのですが
もらえませんか	いただけませんか		したいのですけど
もらえませんでしょうか	いただけませんでしょうか		
もらえないですか	いただけないでしょうか		
もらえないでしょうか	いただきたいのですが		
もらえるでしょうか	願えますか		
いただけるでしょうか	願えませんか		
いただけますか	願えませんか		
いただけますでしょうか	願えませんでしょうか		
いただけませんか	願えないですか		
いただけませんでしょうか	願えないでしょうか		
いただけないですか	願います		
いただけないでしょうか			
くれますか			
くれますでしょうか			
くれませんか			
くれませんでしょうか			
くれないですか			
くれないでしょうか			
ほしいのです			
ほしいのですが			
ほしいのですけど			
もらいたいのです			
もらいたいのですが			
もらいたいのですけど			
いただきたいのです			
いただきたいのですが			
いただきたいのですけど			