「文節訳の出現順序」を考慮した翻訳単位の決定手法

丸山 岳彦　　柏岡 秀紀

ATR 音声言語コミュニケーション研究所
{takehiko.maruyama,hideki.kashioka}@atr.co.jp

1 目的

近年，音声翻訳技術の適用範囲の拡張を目指して，「独話」を対象とした音声翻訳の研究が始まっている [2, 5, 6]．独話の音声翻訳技術が用いられる場面には，講演・会議などの通訳（会議通訳）やニュース放送の通訳（放送通訳）などがあり，幅広い用途が考えられる．これらの場面では，通常，原発話に追従して訳出が行なわれる「同時通訳」としての翻訳が求められることから，そこで用いられる音声翻訳技術も同時通訳として実現されることが望ましい．

これまでの機械翻訳システムは，「文」を入力単位として翻訳処理を行なうものが大半である．しかし，独話は対話に比べて1文が長くなる傾向にあり，文構造も複雑になりやすいため，独話を対象とした翻訳処理では，正確な統語解析が困難になるという問題が生じる [5]．また，1文を入力単位として翻訳処理を行なう場合，入力文が長くなるにつれて訳出開始のタイミングが原発話から大きくずれていくという問題がある．同時通訳としての実用性を考えれば，翻訳文の出力は原発話になるべく追従していく方が望ましい．原発話への追従性を伴う翻訳処理を実現するためには，文全体の入力の終了を待ってから翻訳処理を開始するのではなく，入力と同時に翻訳可能な範囲をリアルタイムで検出し，部分ごとの翻訳結果を順次出力していくという，漸進的な処理が求められる．そこで問題となるのは，「翻訳可能な範囲」とはどのようなものかという点である．

本稿では，日本語の独話を英語に同時通訳することを想定し，「日英同時通訳としての処理に適切な日本語の翻訳単位とはどのようなものか」という問題について論じる．日本語の入力文の中で翻訳可能な単位はどの部分かを検討するための手法を示し，実験を行なった結果について述べる．

2 手法

2.1 チャンクの検出

人間による同時通訳の理論的研究では，聴取した原発話の意味が概念的に再構成され，意味的にまとまった部分ごとに翻訳結果が順次出力される過程がモデル化されている [1]．このような過程を機械的に再現することは困難であるが，意味のまとまりを考慮せず，訳出部の出現順序にのみ注目することによっても，漸進的な翻訳処理に有意義な単位を見つけることは可能である．以下では，「文節訳の並べ替え」という手法を用いて，日本語の原発話の中から訳出可能な範囲（以下，チャンクと呼ぶ）を見つけるための手順を示す．

2.2 手順: 文節訳の並べ替え

以下の手順によって，日本語の原発話からチャンクを検出する．一連の作業過程の例を，図1に示す．

1. 原発話の文節分割
   日本語を文節に分割する．
2. 文節訳の付与
   各文節に対応する英訳を人手で付与し，「文節訳」を作成する．ただし，対訳が付けられない文節には“＊＊＊”を付与し，次の行でまとめて訳出する．
3. 文節訳の並べ替え
   2. でできた文節訳を，なるべく英語の語順に沿うように，文節訳ごとに並べ替える．
4. チャンクの検出
   文節訳を並べ替える前の状態 (2.) と並べ替えた後の状態 (3.) を比較し，文節訳の出現順序が入れ替わった部分をチャンクと見なすことにより，対応する原発話の部分がチャンクとして検出される．

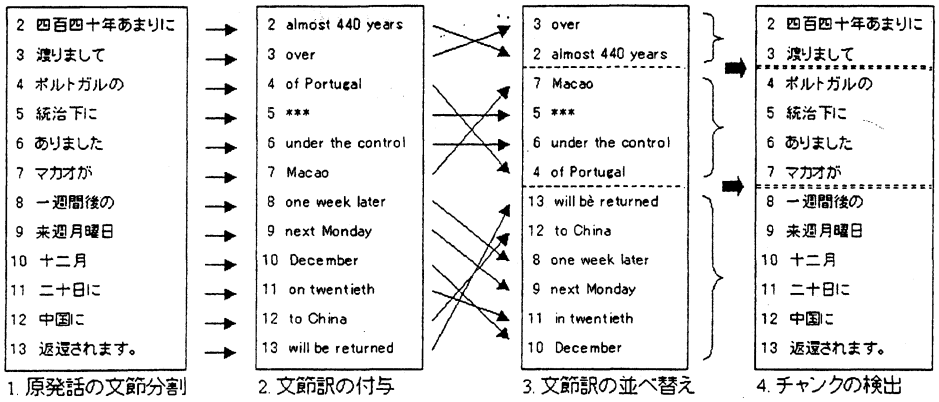


図 1: 「文節訳の並べ替え」の手順

2.3 「文節訳の出現順序」を考慮した翻訳単位

図1に示した手順は，意味的なまとまりは考慮せず，単に文節訳を並べ替えることによってチャンクの境界を得ようとするものである．日英翻訳に有用なチャンクを見つけるためには，日本語と英語の語順の違いを考慮し，日本語が英語に翻訳されるときにどの部分がまとめられるかを観察すればよい．

手順 2.「文節訳の付与」によって，日本語の語順に即した英語の対訳を得ることができる（なお，単語ごとの対訳ではなく，文節ごとの対訳を採用したのは，人手で対訳を付与する際の容易性という事情による．また，文節の意味の曖昧性をできるだけ排除するために，文節訳を付与する際には 3 文節先まで参照してよいというルールを設けた）．また，手順 3.「文節訳の並べ替え」によって，英語の語順に即した文節訳を得ることができる．ここで，両者の間で並び替えられた範囲，つまり「文節訳の出現順序」の差分を取ることによって（手順 4.），原発話のどの部分をまとめれば英語の部分に訳出することができるかを見つけることができるわけである．

図1の例では，原発話が結果として 3 つのチャンクに分割されている．チャンクの境界となった部分の統語的特徴を観察すると，3 行目の「テ節」，7 行目の「格助詞ガ」，そして 13 行目の「文末（句点）」というように，構造的に切れ目の大きい箇所であることが分かる [7]．これは，原発話を部分ごとに翻訳していくときに，統語的な境界がチャンクの境界として利用されていることを示している（4.1節参照）．

3 実験

2節で示した手順に従って，日本語の独話データからチャンクを検出する実験を行なった．

3.1 使用データ

まず，使用したデータについて述べる．ATR では現在，NHK の解説番組「あすを読む」を収録し，独話コーパスとして書き起こし作業を行なっている．2002 年 1 月現在，300 番組分の収録と書き起こしが完了している．今回は，このうち 30 番組分の書き起こしデータを同時通訳の原発話と想定して使用した．データの詳細を表 1に示す[1]．

番組数	総文数	総形態素数	総文節数
30	1,749	52,407	22,994

表 1: 「あすを読む」データ

3.2 結果

「あすを読む」30 番組分のデータに対して，2節で示した手順を適用し，チャンクを検出した．検出されたチャンク数などを，表 2に示す．総文数 1,749 文が 6,819 チャンクに分割された．1 文当たり平均 3.89 のチャンクが含まれることになる．1 チャンク当りの平均形態素数は 7.68，平均文節数は 3.37 であった．

次に，1 チャンクに含まれる文節数を図 2に示す．

[1] 形態素解析，文節への分割は，JUMAN[3] および KNP[4] を用いて行なった．

	「あすを読む」
総チャンク数	6,819
平均チャンク数 / 文	3.89
平均形態素数 / チャンク	7.68
平均文節数 / チャンク	3.37

表 2: チャンク検出結果

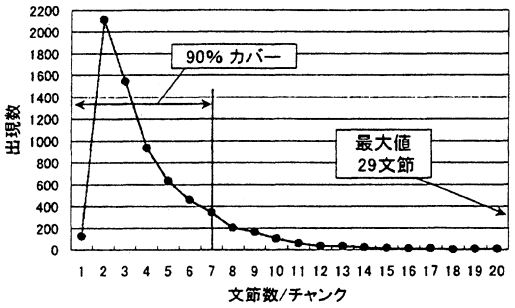


図 2: 1チャンクに含まれる文節数の分布

1チャンク中の文節数の分布で最も多いのは，2文節の2,113例であった．また，1チャンクが7文節で構成されるものまでで，全体の90%をカバーした．1チャンク中に最も多く文節が含まれていたケースは，29文節であった．

4 考察

4.1 チャンク境界の分析

本実験で検出されたチャンク境界を観察し，品詞や統語的特徴について分析を行なった．出現数が多かったチャンク境界の種類について，表3に示す[2]．

1,716	文末	153	時間表現
856	係助詞ハ	145	接続助詞ト
792	接続表現	128	体言止
610	格助詞ガ	114	格助詞ヲ（主語）
327	接続助詞テ	112	係助詞モ
207	定型表現	111	引用節
199	連用節	94	接続助詞ケレドモ
154	接続助詞ガ	...	

表 3: チャンク境界の品詞，統語的特徴

[2] 「文末」の数値 (1,716) が総文数 (1,749) より少ないのは，「今晩は。」などを「定型表現」として分類したためである．

これらの統語的特徴は，大きく3つに分類することができる．以下，それぞれの特性について分析を示す．

a. 主語に翻訳されるもの

「係助詞ハ，モ」や「格助詞ガ」[3] など，英語の主語として訳出される要素がチャンクを形成しているケースが上位を占めた．また，日本語のヲ格（目的語）が英語の主語として並べ替えられるケースもあった（「格助詞ヲ（主語）」）．英語の場合，主語は文頭に置かれることから，主語に該当するこれらの要素がチャンクとして先に訳出され，後から述語句全体がまとめて訳出されるというパタンが多く見られた．

b. 定型的な表現

接続詞や慣用的な接続表現などの「接続表現」や，「ご覧のように」「ある意味で」「これを言い換えますと」など定型的な言い回しの「定型表現」，「来週の月曜日」「七十四年に」「この夏以降」のような「時間表現」などがチャンクとして多く検出された．特にこれらの表現が文頭に出現する場合に，チャンクとして利用されるケースが多く見られた．

c. 節境界

「接続助詞テ，ガ，ト，ケレドモ」で導かれる従属節や，「連用節」，「引用節」など，主節に対する従属度の低い従属節がチャンク境界として検出された．従来，従属度が低い従属節は文の切れ目になりやすいということが指摘されているが [5, 7]，本実験の結果はこの点に合致する．

以上，文節訳の出現順序に注目して検出したチャンクの分析を行なった．本稿での手順は意味のまとまりを考慮したものではないが，結果としては，意味のまとまりを持った部分が検出されていると考えられる．

4.2 チャンクとなるための条件

4.1節では，チャンク境界として検出された表現について分析を行なった．しかし，これらの表現は，全ての場合においてチャンクとして利用されるわけではない．そこで，チャンク境界として検出された各表現について，その表現がチャンク境界になっていない場合を調べた．出現数の多かったものを，表4に示す．

「係助詞ハ，モ」や「格助詞ガ」など，先に「主語に翻訳されるもの」として分類したものが上位を占めた．データを確認したところ，これらの要素が埋め込み構造の中にある場合，チャンクとして利用されてい

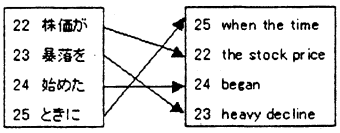
[3] 「～では」のように格助詞を前接させるものや，「～のも」「～ことが」のように補足節の形を取っているものも含む．

757	係助詞ハ	26	定型表現
712	格助詞ガ	21	接続表現
387	係助詞モ	19	ヨウニ節
350	接続助詞テ	17	副詞
285	時間表現	15	タメニ節
47	連用節	13	従属文
44	場所表現	13	引用節
38	間接疑問節	...	

表 4: チャンク境界の候補がチャンクにならなかった場合

ないことが多かった．例えば(1)の「格助詞ガ」は連体節の中にあるため，チャンクとしては利用されず，when 節の内部で並べ替えが行なわれている．

(1) 株価 が | 暴落を | 始めた | ときに ||



また，「係助詞ハ，モ」や「格助詞ガ」には，「〜ことにはなりません」「地理的にも極めて近い」など，主語としては訳されない例も多く存在する．このような場合にもチャンクとしては利用されていなかった．

さらに，「時間表現」や「場所表現」，「定型表現」なども，チャンクにならない場合があった．4.1節でも述べたように，これらの要素は，文頭に出現した場合には単独で訳出可能なものとしてチャンクとして利用される．しかし，図1の8,9行目や10,11行目のように，主述関係に挟まれた位置に出現した場合，英語の語順としては述語の後に訳出される要素であるため，独立したチャンクとしては利用されていない．これらの要素については，その出現位置によってチャンクになるかどうかが決まると言える．

また，「テ節」や「連用節」などの従属節もチャンクにならない場合があった．これらは，主節に対する従属度が高い場合，つまり文構造の大きな切れ目になっていない場合には，チャンクとして利用されていないと考えられる．

以上のように，本稿の手法でチャンクとして検出された要素が，実際にチャンクとして用いられるかどうかは，その出現順序や統語構造を考慮する必要がある．ある要素をチャンクとして認定するための条件については，今後より詳しく分析を行いたい．

5 課題

日本語の独話を英語に同時通訳することを想定し，「文節訳の並び替え」という手法によって日本語の翻訳単位について検討した．本手法で検出されたようなチャンクを翻訳単位としておくと，同時通訳としての追従性を実現し，かつ英語の自然な語順に即した翻訳結果を得ることができる．

今後の課題として，チャンクを自動的に検出するための知識の獲得が挙げられる．4.2節で述べたように，ある要素をチャンクとして認定できるかどうかはその出現位置や統語構造に依存して決まるため，チャンクを自動的に検出するには，その認定基準について詳しく検討しなければならない．また，対訳コーパス内の日本語と英語をチャンク単位で対応づけるなど，本稿で示したチャンクの概念をアラインメントの単位として活用することも考えられる．

さらに，「翻訳処理の単位」と「翻訳出力の単位」を分けて考えることによって，チャンク単位で翻訳を行なう処理と，その翻訳結果を出力するタイミングを制御する処理とに分離することができる．翻訳と出力の各処理をモジュール化することによって，漸進的な翻訳処理にかかる負荷を下げ，より効率的にチャンクを活用することができるものと考えられる．

謝辞: 本研究は通信・放送機構の研究委託により実施したものである．

参考文献

[1] 船山仲他 1996. 同時通訳における処理単位について. 通訳理論研究 10, pp.4–13. 通訳理論研究会.

[2] 柏岡秀紀・田中英輝 2001. 講演の同時通訳データの分析. 言語処理学会第 7 回年次大会発表論文集, pp.433–436.

[3] 黒橋禎夫・長尾真 1999. 日本語形態素解析システム JUMAN version 3.61. 京都大学.

[4] 黒橋禎夫・長尾真 1998. 日本語構文解析システム KNP version 2.0 b6. 京都大学.

[5] 丸山岳彦・熊野正・柏岡秀紀 2001. 日本語における独話の特徴と文分割. 言語処理学会第 7 回年次大会発表論文集, pp.429–432.

[6] 松原茂樹・相澤靖之・河口信夫・外山勝彦・稲垣康善 2001. 同時通訳コーパスの設計と構築. 通訳研究 1, pp.85–102. 日本通訳学会.

[7] 南不二男 1974. 現代日本語の構造. 大修館書店.