日本語における独話の特徴と文分割

丸山岳彦　　熊野正　　柏岡秀紀

ATR音声言語通信研究所
{maruyama,tkumano,kashioka}@slt.atr.co.jp

1　はじめに

ATRではこれまで「バイリンガル旅行会話コーパス[1]」を作成し，「対話」の音声翻訳システムの開発を進めてきた．これは「旅行会話」という限定されたタスクに関して対話の音声翻訳を行なうものであり，一定の成果を挙げている．これとは別に，現在，音声翻訳の適用範囲の拡張を目指して，「独話」の音声翻訳技術の開発に取り組んでいる．独話の音声翻訳技術の適用例には，講演や会議の通訳やニュースなど放送番組の通訳など，幅広い応用範囲が考えられる．

一般的に，講演やニュースなどの独話は対話に比べて1文が長く，文の構造も複雑になっているものと予想される．本稿では，現在ATRで収集している独話コーパスを対話コーパスと比較して特徴分析を行なった結果について述べる．コーパスを定量的に調査・分析し，独話の特徴を明らかにしておくことは，独話の音声翻訳の実現に知見を与えるものとして有効である．また，独話の音声翻訳を実現する際の問題点を指摘した上で，節境界による文分割について検討する．

2　独話コーパスの収集

現在ATRで独話コーパスとして収集しているのは，NHKで放送されている10分間の解説番組「あすを読む」である．これは，時事・経済・社会問題などのテーマについて，1人の解説委員が解説を行なうものである．我々はこの番組を「講演」と見なし，独話分析の対象としている[1]．

2001年1月現在，200番組の収録が完了している．これらをテキストに書き起こし，ATRの「変換主導型翻訳システム(TDMT)[4]」の体系に基づいて形態素付与作業を行っている．さらに，同時通訳者によって番組の同時通訳を行ない，日英の同時通訳コーパスとしての構築にも当たっている[2]．

[1] 番組の収録および使用については，NHKとの共同研究という形で許諾を得ている．

3　独話と対話の比較調査

本節では，収集した「あすを読む」のうち形態素付与の完了した50番組分について，特徴分析を行なった結果について述べる．「あすを読む」50番組分の形態素数[2]とほぼ同サイズに合わせた「バイリンガル旅行会話コーパス」(以下「旅行会話」)を用意し，独話と対話の比較調査を行なった．

3.1　文数と形態素数

両コーパス中に現れる文や形態素の数について調査を行なった．まず総文数[3]を求め，1ファイル[4]当たりの平均形態素数，1文当たりの平均形態素数，異なり語数を比較した．結果を表1に示す．

	あすを読む	旅行会話
総形態素数	101,268	101,264
ファイル数	50	251
総文数	3,010	9,412
平均形態素数 / ファイル	2,025	403
平均形態素数 / 文	33.6	10.7
異なり語数	7,428	2,807

表1: 総形態素数，総文数など

「あすを読む」1文当たりの平均形態素数は「旅行会話」の3.1倍となっており，1文が長くなる傾向にあるという独話の特徴を確かめることができる．

次に，1文に含まれる形態素数について調査を行なった．図1に示す．

1文に含まれる形態素数が最も多いケースは，「あすを読む」では178形態素，「旅行会話」では61形態素であった．また，「あすを読む」では59形態素までの文で全体の90%をカバーしたのに対し，「旅行会話」では21形態素までの文で90%をカバーした．これらの点からも，独話は対話に比べて1文が長いことが分かる．

[2] ここでは，間投詞や句点などの記号を含む．
[3] 文境界の認定は，書き起こし作業者の判断による．
[4] 1ファイルは，「あすを読む」の1番組あるいは「旅行会話」の1会話に相当する．

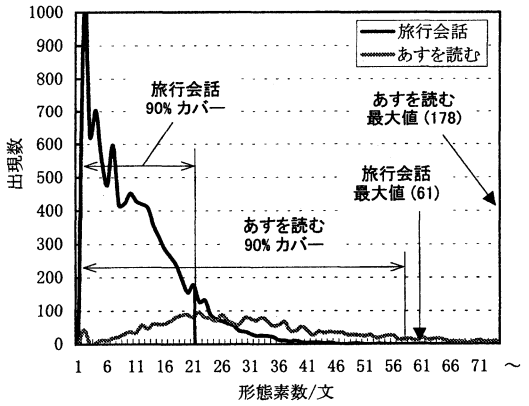

図 1: 1 文当たりの形態素数と出現数

3.2 従属節の種類と出現数

さらに，両コーパス中に現れる従属節 (接続助詞や述語の連用形などで表される節) の種類と出現数について調べるため，節境界を抽出する規則を作成した．これは「出現形_品詞_活用形_活用型」という 4 つの情報を用い，ある語の前後 0 ～ 3 語を見て節境界の判定を行なうものである．これを 3.1 節で見た同サイズの両コーパスに適用し，比較を行なった．表 2 に示す．

従属節の種類	あすを読む (計 6,444)		旅行会話 (計 3,482)	
並列節	733	(11.3%)	1,273	(36.5%)
理由節	178	(2.7%)	414	(11.8%)
条件節	396	(6.1%)	408	(11.7%)
譲歩節	109	(1.6%)	25	(0.7%)
テ節	914	(14.1%)	338	(9.7%)
連用節	653	(10.1%)	170	(4.8%)
連体節	2,273	(35.2%)	321	(9.2%)
補足節	379	(5.8%)	149	(4.2%)
引用節	569	(8.8%)	205	(5.8%)
間接疑問節	209	(3.2%)	179	(5.1%)
従属文	31	(0.4%)	0	(0%)

表 2: 従属節の種類と出現数

1 文中に含まれる平均従属節数は，「あすを読む」では 2.1，「旅行会話」では 0.3 となる．1 文に含まれる従属節の数が多いほど，文の構造は複雑になると予想される．この点から，独話は対話に比べて文構造が複雑であるという特徴が見て取れる．

通常，文末には述語部分の終止形や終助詞など言い切りの形が現れるが，「～ございますが。」や「～ですから。」のように従属節で文が終わる場合もある．従属節で文が終わる例の出現数を比べると，「あすを読む」が 3 例であるのに対して，「旅行会話」では 783 例あり，出現数に顕著な差が見られた．両コーパスにおいて，言い切りの形で文が終わる例，従属節の形で文が終わる例，それ以外の文中に現れる従属節，という三つの出現数は，表 3 のようになる．

	あすを読む (計 9,451)	旅行会話 (計 12,111)
従属節の形の文末	3	783
言い切りの形の文末	3,007	8,629
文中に現れる従属節	6,441	2,699

表 3: 従属節の生起位置と文末の数

従属節で文が終わる例は，文末を明確に言い切らない形で発話を終えることによって聞き手の反応を促すという，対話に特徴的な発話スタイルと考えられる．対話で観察されるこのような特徴が独話ではほとんど観察されないという点で，両者は異なる特徴を持つ話し言葉コーパスであると言える．

3.3 各従属節の分析

以下では，表 2 で出現数の割合の差が大きかった並列節，理由節，連体節，従属文について分析する．

3.3.1 並列節

ここでいう並列節とは，接続助詞「が，け (れ) ど (も)，し」や，判定詞連用形「で」，副助詞「とか，たり」などで表される従属節のことである．従属節全体に占める並列節の割合を比較すると，「旅行会話」が「あすを読む」の 3.2 倍であった．

「旅行会話」に出現した並列節 1,273 例のうち，接続助詞「が，け (れ) ど (も)」で表される並列節が 1,140 例あった．このうち，「～したいのですが。」のような形で文末に現れたものが 670 例あった．全出現数の半数以上が文末に現れているという点に，対話に特徴的な発話スタイルを見て取ることができる．一方，「あすを読む」に出現した並列節 733 例のうち，接続助詞「が，け (れ) ど (も)」で表されるものは 456 例あった．このうちで，文末に現れたのは 1 例のみであった．

3.3.2 理由節

ここでいう理由節とは，接続助詞「から，ので」で表される従属節のことである．従属節全体に占める理由節の割合は，「旅行会話」が「あすを読む」の 4.3 倍であった．

「旅行会話」に出現した理由節 414 例のうち，「～ですから。」のような形で文末に現れているものが

85例あった．先の並列節の場合と同様，対話に特徴的な発話スタイルと言える．また，残りの329例のうち221例は文末の依頼表現を修飾するものであった．先に理由を述べてから依頼を行なうという形もまた，対話に特徴的なものと思われる．一方，「あすを読む」に出現した理由節178例のうち，文末に用いられるケースは観察されなかった．また依頼表現を修飾するものは全体の中で3例見つかったが，いずれも引用節の内部に収まるものであり，発話者による依頼の理由を表す例は1例もなかった．

さらに，理由節を表す接続助詞「から」と「ので」の出現数に顕著な違いが観察された．「あすを読む」では「から」が92例，「ので」が86例だったのに対し，「旅行会話」では「から」が18例，「ので」が394例であった．

3.3.3 連体節

連体節には，節が被修飾名詞を直接修飾する場合(ゼロ形式による接続)と，「という，ための，ような」などの形式によって節と被修飾名詞が接続される場合がある．両者を合計した結果(表2)，従属節全体に占める連体節の割合は，「あすを読む」が「旅行会話」の3.8倍であった．また，ゼロ形式で接続される連体節のみに注目すると，「あすを読む」で1,472例，「旅行会話」で257例であり，「あすを読む」の方が「旅行会話」の5.7倍の出現数となっている．

連体節は埋め込み構造を発生させ，文の構造を複雑化する．特に「あすを読む」においては1文中に複数の連体節が現れるケースも多く観察されたことから(表4)，構造が一層複雑化しているものと思われる．

出現数	あすを読む	旅行会話
1	1,009	301
2	376	10
3	115	0
4	33	0
5	7	0

表 4: 1文中における連体節の出現数

3.3.4 従属文

「真性モダリティを持たない文」または「従属文」と呼ばれる，独話に特徴的な表現が「あすを読む」で観察された．これは文相当の形式が他の文の部分に従属するもので，従属文全体を受ける指示語が文の他の部分に存在するという特徴がある [3]．

(1) 制裁発動によって目的が達成される、 /
そう考えるのは楽観的に過ぎるでしょう。

このような表現は，講演などの独話に特徴的な表現と考えられる．

以上，独話の特徴について，対話と比較して調査した結果について述べた．

4 独話の文分割

前節で見た独話の特徴を踏まえて，本節では独話の音声翻訳を実現する際の問題点とその対処策について論じる．

4.1 問題の所在

独話の音声翻訳で問題となるのは，1文が長いことによる構文解析精度の低下である．ここでは，翻訳処理の単位を短くすることを目的として，文を分割する処理について検討する．

機械的に分割を行なうためには，分割点を決めるための何らかの指標を文の中から見つける必要がある．本稿では，従属節に関して分析を行なった結果をもとに，文分割の手がかりとして節境界に注目する．

4.2 節境界の抽出と分割

3.2 節で述べた節境界を抽出する規則を用いて，「あすを読む」に現れた全ての節境界で分割を行なった．分割結果の例を図2に示す．

原文: 「十月の失業率は四点六パーセントと依然高い水準で三百万人を越す人が仕事を求めてます。」

十_数詞__ 月_普通名詞__ の_連体助詞__ 失業率_普通名詞__ は_係助詞__ 四_数詞__ 点_普通名詞__ 六_数詞__ パーセント_普通名詞__ と_格助詞__ 依然_副詞__ 高い_形容詞_形容詞_基本 水準_普通名詞__ で_判定詞_形容動詞_連用 / 並列節 /
三_数詞__ 百_数詞__ 万_数詞__ 人_普通名詞__ を_格助詞__ 超す_本動詞_五段サ_基本 / 連体節 /
人_普通名詞__ が_格助詞__ 仕事_サ変名詞__ を_格助詞__ 求め_本動詞_一段_た て_助動詞_一段_連用 ます_助動詞_特殊サ_基本 。_記号__ / 文末 /

図 2: 節境界での文分割の例

以下では，分割された1つの単位を「セグメント」と呼ぶ．節境界の全てを分割点として「あすを読む」を分割すると，分割後の総セグメント数は9,451となり，もとの総文数3,010文から3.1倍に増加した．

1セグメントの長さが短くなった一方，極端に少ない形態素から成るセグメントの数が増加した．短すぎ

るセグメントを翻訳しようとしても，格要素などの情報が不足していたり，談話全体としての結束性が損なわれたりして，翻訳品質が劣化する恐れがある．そこで，全ての節境界の中から，実際の分割点とするものの選択を行なうことにする．

4.3 分割点の選択

南不二男によると，従属節はその形式によって主節への従属度がある程度決まっており，従属度が低くなるほど文の切れ目になり易い [5] ．この性質を利用して，従属度の低い従属節を以下のように選択し，その節境界を分割点とした．

> 並列節，理由節，条件節，テ節の一部 (丁寧体＋テ)，
> 連用節の一部 (連用形接続)，間接疑問節，従属文

この分割の結果，分割後の総セグメント数は 5,004 となり，分割前の 1.6 倍となった．また，分割後の 1 セグメントに含まれる平均従属節数は 1.4 となった．

分割前，全ての節境界による分割，分割点の選択を行なった上での分割，という各条件について，1 セグメント中の形態素数を比較した結果を表 5 に示す[5] ．

形態素数	分割前	全ての節	選択後
1–5	63	2,826	377
6–10	173	3,777	1,052
11–15	370	1,833	1,227
16–20	480	683	932
21–25	480	229	595
26–30	409	65	386
31–35	308	19	202
36–40	218	12	109
41–45	155	5	56
46–50	115	0	30
51–60	124	1	23
61–70	57	1	10
71–	58	0	5
計	3,010	9,451	5,004

表 5: 1 セグメント中の形態素数と出現頻度

分割点を選択することによって，全ての節で分割した場合に問題となった，極端に少ない形態素から成るセグメントの数は減少した．が，逆に 1 セグメントが長いままのケースが多く残った．例えば 31 以上の形態素を含むセグメントは 435 例あり，全ての節で分割した場合の 38 例に比べ 11.4 倍となっている．

[5] ここでは，間投詞や記号類は除いてある．

5 まとめ

本稿では，ATR で独話コーパスとして収集している「あすを読む」を題材として，独話の特徴分析を行なった．形態素や文，従属節の出現数や種類などを対話コーパスと比較・分析することによって，いくつかの知見を得た．独話では，対話に比べて 1 文内に含まれる形態素数や従属節の数が多く，対話よりも 1 文が長いこと，そして文の構造が複雑になっていることが分かった．また，対話と独話それぞれに特徴的な表現が存在することから，同じ話し言葉のコーパスでも性格の異質なものであることが分かった．これまで，独話の特徴についての十分な調査・分析はなされてこなかったが，今回の調査によって実際的・定量的な分析を行なうことができた．こうした分析は，今後独話の音声翻訳システムを構築していく上で，有効な指針を与えるものとして期待することができる．

また，独話の特徴分析を行なった結果を利用して，節境界に注目した文分割について検討した．ここでは全ての節境界で分割した場合と分割点を選択した上で分割した場合を挙げ，選択点を変更することによってセグメント長が調節できることを示した．ただし，分割を行なった後，そのセグメント内に含まれるべき格要素が欠落していたり，他のセグメントに係る成分が含まれていたりする可能性がある．分割結果を翻訳処理に活用するためには，このような未決定要素・余剰要素を処理するための対策が必要となる．また，分割結果が翻訳処理に適しているかどうかは，分割後のセグメント長の閾値なども考慮した上で，評価基準を設定し，評価を行なう必要がある．今後は，独話コーパスの特徴分析を進めるとともに，独話の翻訳処理を行なう際の問題点を具体化に検討し，文分割処理，分割結果の妥当性などについて検討したい．

参考文献

[1] T. Morimoto, et. al., 1994, "A speech and language database for speech translation research," *Proc. International Conference on Spoken Language Processing*, pp.1791–1794.

[2] 柏岡秀紀 (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.

[3] 野田尚史 (1989). "真性モダリティをもたない文." 日本語のモダリティ, pp.131–157. くろしお出版.

[4] 古瀬蔵・山本和英・山田節夫 (1999). "構成素境界解析を用いた多言語話し言葉翻訳." 自然言語処理, **6**(5), pp.63–91.

[5] 南不二男 (1974). 現代日本語の構造. 大修館書店.