後置詞句における助詞の実現と非実現の比較分析

1M−6

保坂順子 竹沢寿幸 浦谷則好
ATR 音声翻訳通信研究所

1 はじめに

音声処理と言語処理を統合した話し言葉の音声翻訳に関する研究を進めている。日本語話し言葉の特徴として、後置詞句の助詞が発話されないことがあげられる[1]。音声認識では、助詞は音韻数が少ないため、誤認識の対象になりやすい[2]。また、言語処理では、格情報に依ることが多く、その欠落に対処することは難しい。本稿では、この問題に対処するため、助詞が発話されている後置詞句と発話されていないものを比較検討する。まず、後置詞句の音声認識を示す。次に、対話データベースを基に[3]、後置詞句を調査する。最後に、音声翻訳への応用の可能性を検討する。

2 後置詞句の音声認識

我々の音声翻訳システムの音声認識部では、制約として構文情報を使っている[4]。後置詞句での助詞の使用を自由にし、1a と 2a を音声入力したところ、1a-1c 及び 2a-2c が認識候補になった。

1a. こちらは 会議事務局です
1b. こちら 会議事務局です
1c. こちらが 会議事務局です

2a. 用紙を 送ってください
2b. 用紙 送ってください
2c. 用紙に 送ってください

1a-2c から、入力とは異なる助詞も認識されることが分かる。この様な候補から尤もらしい文を選択するには、実際の助詞の使用を調べる必要がある。

3 対話データベース検索

対話データベースを基に、普通名詞または代名詞を含む後置詞句で、助詞が使われる場合と使われない場合を調べた。対象とした国際会議に関するデータは、179 電話会話 (約 19 万単語)、175 キーボード会話 (約 7 万単語) である。話し言葉 (電話会話) と準書き言葉 (キーボード会話) も比較した。

表 1 に、助詞が使われている場合 (実現) と使われていない場合 (非実現) の出現頻度を示す。

表 1: 助詞の実現と非実現

	電話会話		キーボード会話	
	普通名詞	代名詞	普通名詞	代名詞
実現	12706	2451	6954	967
非実現	2504	539	1164	204
合計	15210	2990	8118	1171
非実現 %	16.5	18.0	14.3	17.4

表 1 から、準書き言葉より話し言葉で助詞の非実現が多いことが分かる。しかし、その差は顕著ではない。これは、我々のドメインでは、それほど親しい者同士の会話がないためであろう。

助詞の非実現に関して、その復元を試みた[5]。自然に復元できたものは、普通名詞では、話し言葉 31%、準書き言葉 19%、代名詞では、それぞれ 78% と 80% であった。代名詞の復元率が高いので、さらにその内訳を調べた。

3.1 話し言葉の代名詞と助詞

表 2 に代名詞と復元した助詞を示す[1]。「私たち」と「こちら様」もあり、それぞれ「は」(4 例) と「が」(1 例) を復元した。表 2 では、助詞が自然に復元できたもののうち 415 例 (98.8 %) を扱う。

表 3 に実際に発話されている代名詞と助詞の組合せを示す。ここで扱っている代名詞以外にも「あなた」「彼女」「誰」など多様だった。表 3 では、全組合せのうち 1656 例 (67.6%) を扱う。

「は」「が」と並んで、「を」も非実現が多いと言われる[1]。しかし、表 2 と表 3 から、「を」

Comparison of Postposition Occurrence and Non-Occurrence
Junko Hosaka, Toshiyuki Takezawa, Noriyoshi Uratani
ATR Interpreting Telecommunications Research Labs.

[1]助詞の復元の際に、「が」か「は」、「を」か「は」または「の」か「は」に迷ったものは、「が」「を」「の」とした。

表 2: 話し言葉の代名詞と復元助詞 (%)

	は	が	に	を	で	の
こちら	12.4	11.4	1.2	0.2	0.7	0.2
そちら	2.4	4.0	0.2			
これ	1.7	1.2	0.2	0.2		0.2
それ	0.5	0.5	0.2	0.5		
皆さん		1.0	0.5			
私	45.5	6.7	0.2			0.7
私ども	3.6	2.1	0.2			0.2

表 3: 話し言葉の代名詞と助詞 (%)

	は	が	に	を	で	の
こちら	0.4	0.5	1.1	0.1	1.9	14.9
そちら	0.2	0.1	1.5	0.1	1.3	10.1
これ	4.8	0.5	0.4	0.4	0.4	0.1
それ	4.7	0.6	1.4	2.2	2.4	0.1
皆さん		0.2	0.7			0.2
私	1.7	2.4	0.4		0.0	5.8
私ども	0.1	0.4	0.1	0.1	0.5	4.4

はあまり使われないことが分かる。「私」は単独で発話されることが多く、ほとんど「は」が復元される。一方、助詞が発話される場合は、「の」が多い。「こちら」は「の」を伴うことが多いが、復元される場合は「は」と「が」が多い。

3.2 準書き言葉の代名詞と助詞

表4 に、代名詞と復元した助詞を示す。表4 では、助詞が自然に復元できた全 163 例を扱う。

表 4: 準書き言葉の代名詞と復元助詞 (%)

	は	が	に
こちら	14.1	31.9	
そちら		1.2	0.6
これ			
それ			
皆さん		1.8	
私	42.3	5.5	
私ども	2.5		

表5 に実際に使われている代名詞と助詞の組合せを示す。全組合せのうち 586 例 (60.6%) を扱う。

話し言葉同様、「を」の使用率は低い。代名詞の後で「を」が復元されたものは 1 例もなかった。「私」の後には「は」の復元が多く、「は」が発話されることも多い。

表 5: 準書き言葉の代名詞と助詞 (%)

	は	が	に	を	で	の
こちら	2.7	0.4	0.5		2.8	6.1
そちら	0.6	0.2	1.9	0.2	1.1	5.1
これ	1.1	0.3	0.4	0.1	0.9	
それ	4.2	0.6	2.2	1.9	2.9	
皆さん	0.1					0.2
私	9.1	4.1	0.5		0.1	6.6
私ども	0.6	0.4	0.1	0.1	0.1	2.2

4 おわりに

3 節から、代名詞により、助詞の実現、非実現の傾向が異なることが分かる。また、「に」の復元など、話し言葉の方が多様である。しかし、全体的には、両者とも似た分布を示している。

この調査を音声翻訳に応用する場合、音声認識で「こちら」や「そちら」の後の「の」の尤度をあげることが考えられる。解析では、「こちら」を単独で処理する場合、「は」や「が」を補うことが考えられる。また、生成では、「私」の後には助詞を使わず、自然な文にすることが考えられる。

話し言葉の特徴として、さらに、文末に終助詞が使われることがあげられる [6]。終助詞は、音声認識でも誤認識されやすい。今後、対話データベースを使い、終助詞の調査も進めていく予定である。

参考文献

[1] Shibatani, M. (1990, 1992[3]): The Languages of Japan, Cambridge University Press, Cambridge.

[2] 保坂, 竹沢, 江原 (1991): "対話データベースを利用した音声認識のための構文規則", 情報処理学会 自然言語処理研究会 83-13.

[3] 江原 (1990): "ATR 対話データベースの内容", ATR テクニカルレポート TR-I-0186.

[4] 竹沢, 森元, 谷戸, 鈴木, 嵯峨山, 樽松 (1993): "ATR 音声言語翻訳実験システム ASURA", 第 46 回情報処理学会全国大会 6B-5.

[5] 保坂, 竹沢, 浦谷 (1992): "対話データベースを使った無助詞名詞句の分析", 人工知能学会 第 3 回言語・音声理解と対話処理研究会 SIG-SLUD-9203-1.

[6] 水谷 (1985): 日英比較話しことばの文法, くろしお出版.