

支援対話に現れた言語現象の分析

— 間投表現に着目して —

熊本忠彦 伊藤 昭

郵政省 通信総合研究所 関西先端研究センター

6H-3

1. はじめに

我々は、話し言葉対話によってユーザの計算機利用を支援するシステムを研究/開発している [1]。ユーザは、何らかの障害/問題が発生したとき、音声でこの支援システムに助けを求めることができる。支援システムは、ユーザの発話を理解し、そのときの計算機の状態に合わせて適切な応答を生成する。

以上のような役割を果たす対話システムを設計する際の基礎データとして、人間同士の支援対話（ユーザとコンサルタントの対話）を収集し、支援対話データベースを作成した [2]。本稿では、このデータベース中のユーザ発話文において観測された様々な言語現象のうち、特に間投表現に着目して、その出現頻度や出現位置を分析する。

2. 分析対象となるユーザ発話文

我々は、支援システムの代わりに人間コンサルタントが支援するという「対話による支援」実験を行い、ユーザの自由かつ自然な発話 (spontaneous speech) を収集した。そして、収集されたユーザ発話を書き起こすことによって、支援対話データベースを作成した [2]。本稿では、このデータベース中のユーザ発話文（ユーザ 42 人による発話文 855 文）を分析対象とする。

3. 間投表現の種類と出現頻度

ユーザ発話文において観測された間投表現 (35 種) を、その出現頻度とともに表 1 に示す。なお、表 1 に示されているクラスとは表層的または意味的に似ている表層表現を 1 つにまとめたものであり、その結果、音声データの書き起こし時に生じる、長音や促音などの有無に対する揺らぎも吸収される。

表 1 によれば、上位 10 種で全体の 81.6% を占めており、少数の語彙が多用されるという点では中川ら [3] や佐川ら [4]、村上ら [5] の分析とほぼ一致している。しかしながら、その内訳 (種類) という点ではバラツ

表 1 間投表現の種類と出現頻度

表層表現	頻度 (%)	クラス	頻度 (%)		
え	60 (18.7)	エ	110 (34.3)		
えっ	43 (13.4)				
えー	6 (1.9)				
えーっ	1 (0.3)				
えっと	27 (8.4)				
えーと	14 (4.4)				
えと	8 (2.5)				
えーっと	7 (2.2)				
えっとー	4 (1.2)				
えとー	3 (0.9)				
えとお	2 (0.6)	エット	67 (20.9)		
えっーと	1 (0.3)				
えっとお	1 (0.3)				
すいません	28 (8.7)				
すみません	15 (4.7)				
ごめんなさい	1 (0.3)				
すいませーん	1 (0.3)				
すいませんが	1 (0.3)				
たびたび すいません	1 (0.3)				
あ	21 (6.5)			スミマ セン	47 (14.6)
あっ	19 (5.9)				
あー	1 (0.3)				
ああ	1 (0.3)				
あの	26 (8.1)				
あのぉ	4 (1.2)				
あのー	1 (0.3)				
と	9 (2.8)				
っと	2 (0.6)				
とー	1 (0.3)				
あれ	6 (1.9)	ア	42 (13.1)		
その	2 (0.6)				
ん	1 (0.3)				
んー	1 (0.3)				
えのお	1 (0.3)				
や	1 (0.3)				
合計	321 (100)			アノ	31 (9.7)
		ト	12 (3.7)		
		アレ	6 (1.9)		
		ソノ	2 (0.6)		
		ン	2 (0.6)		
		その他	2 (0.6)		

An Analysis of Linguistic Phenomena in Advisory Dialogues — Focused on Interjectional Words
Tadahiko KUMAMOTO, and Akira ITO
Kansai Advanced Research Center, Communications Research Lab., MPT of Japan

表2 間投表現の直前にある表現

間投表現の直前	頻度	(%)
文頭	211	(65.7)
間投表現	47	(14.6)
提題助詞	11	(3.4)
格助詞	11	(3.4)
名詞接続助詞	7	(2.2)
名詞相当	7	(2.2)
述語接続助詞	6	(1.9)
単語の一部	6	(1.9)
述語活用形	4	(1.2)
接続詞	4	(1.2)
述語基本形	3	(0.9)
終助詞	2	(0.6)
副詞的名詞「とき」	1	(0.3)
連体詞相当	1	(0.3)
合計	321	(100)

キが見られ、対話ドメインなどの違いによって音声理解システムに登録されるべき間投表現の種類が異なってくる事がわかる。

4. 間投表現の出現位置

間投表現がどのような表現の後に続くのかを調べた。その結果を表2に示す。表2によれば、間投表現は文頭に現れる場合が最も多く、全体の65.7%に達している。この値は中川らの分析結果(39.4%)に比べてかなり大きい。その原因として、支援対話が質問応答型の対話であり、ユーザが連続して発話しないこと、発話と発話の間に比較的に間があること、1文あたりの発話文字数が19.7(間投表現を除くと18.8)と少ないことなどが考えられる。なお、文頭に現れた間投表現の内訳は、「エ(91)」、「スミマセン(35)」、「ア(33)」、「エツト(27)」、「アノ(12)」、「ト(8)」、「アレ(4)」、「その他(1)」であり、中川らの分析結果とは若干異なっている。対話ドメインの違いが原因と考えられる。

また、表2によれば、次に出現頻度が多いのは間投表現であり、支援対話では間投表現が連続的に現れやすいことがわかる。そこで、間投表現の連鎖回数を調べてみると、単独発生が241回、2個連鎖が21回、3個連鎖が10回、4個連鎖が2回、計274回であった。単独発生および2個連鎖の場合が全体の95.6%を占めており、中川らの分析結果と同じ結果となった。ここで2個以上の連鎖(33回)に対し、クラスレベル

表3 間投表現(クラス)のbigram

1つめ=2つめ	頻度	(%)
スミマセン=アノ	10	(21.3)
アツ=スミマセン	8	(17.0)
スミマセン=エツト	7	(14.9)
エツ=エツト	3	(6.4)
ト=ト	2	(4.3)
エツ=ト	2	(4.3)
アノ=スミマセン	2	(4.3)
その他(13パターン)	各1	(各2.1)

でbigramを調べてみた。その結果を表3に示すが、全bigram(47個)のうち、29個(61.7%)が「スミマセン」を含んでおり、その分布に偏りがあることがわかる。

5. まとめ

本稿では、支援対話中のユーザの発話文において観測される様々な言語現象のうち、間投表現に焦点を当て、その出現頻度や出現位置を分析した。その結果、間投表現の種類は少数ではあるが、種類そのものは対話ドメインなどによって異なってくる事がわかった。また、間投表現の出現位置は、文頭、間投表現の後、助詞の後の3カ所で全体の約92%をカバーしていることがわかった。支援対話の特徴と言える。

参考文献

- [1] 熊本, 伊藤: “支援対話の解析に基づく対話処理方式の提案”, 信学論, J77-D-II, 8, pp. 1492-1501 (1994).
- [2] Kumamoto, T., Ito, A., and Ebina, T.: “Design and Construction of an Advisory Dialogue Database,” IEICE Trans. Inf. & Syst., E78-D, 4, pp. 420-427 (1995).
- [3] 中川, 小林: “自然な音声対話における間投詞・ポーズ・言い直しの出現パターンと音響的性質,” 日本音響学会誌, 51, 3, pp. 202-210 (1995).
- [4] 佐川, 大西, 杉江: “自己修復を含む日本語不適格文の分析とその計算機による理解手法に関する考察,” 情処学論, 35, 1, pp. 46-52 (1994).
- [5] 村上, 嵯峨山: “自由発話音声認識における音響的および言語的な問題点の検討,” 信学技報, NLC91-57/SP91-100, pp. 71-78 (1991).