係り受けに基づく話し言葉コーパスの統計的分析

大野 誠寛　　村瀬 隆久　　松原 茂樹　　河口 信夫　　稲垣 康善　　(名古屋大学)

1　はじめに

音声対話処理システムを実現する上で自然発話の解析が不可欠であるが、話し言葉にはフィラーや言い淀み、言い直しなどが頻出するため、それを従来の言語解析手法を用いて処理することは難しい。ロバストな解析処理の実現に統計情報を用いることは効果的であるが、より精緻な確率モデルを構築するために、話し言葉の特徴を明らかにする必要がある。本稿では、大規模音声対話コーパスを用いた対話音声の係り受けに関する分析について述べる。分析の結果、対話音声には、係り受けの交差、係り文節から受け文節への前方修飾、係り先のない文節の存在などが特徴的な現象を伴って頻出することを確認した。

2　係り受け分析済み音声対話コーパスの構築

自然な対話音声に係り受け分析を与えるために、名古屋大学 CIAIR 車内音声対話コーパス [1] を用いた。コーパスには、車両運転中のドライバとナビゲータとの対話が収録されている。収集した音声データは書き起こされ、ポーズで分割された各対話音声を発話単位として、その開始時間及び終了時間を記録している。車内音声対話コーパスの 81 対話からドライバの 7,781 発話を取り出し、分析の対象とした。係り受けの付与は人手で実施し、品詞体系や係り受け文法は京大コーパス [2] の作成基準に、また、文節切りについては日本語話し言葉コーパス [3] の基準に準拠している。ただし、対話音声特有の現象については以下の基準を新たに設けた。

- 受け文節が省略された文節の係り先は存在しない。
- 話し言葉特有の言い回し表現 (「こっから」、「食べてえ」など) については、新たな辞書項目を設けて、形態素ごとに品詞を定めた。
- 対話ターンを 1 つの修正対象とする。発話単位を越える係り受けも認めるが、対話ターン間に跨った係り受けは認めない。
- 「じゃ」、「はい」などの感動詞にも受け文節を与える。

3　係り受けコーパスの統計的分析

日本語係り受け解析では、一般に、係り先の唯一性、後方修飾性、係り受けの非交差性といった制約を仮定するとともに、係り受けが文を跨らないことを前提とする。しかし、話し言葉の場合，係り受けに関する上述の性質に従わないことも多いと予想され，話し言葉解析においてこれらを絶対的な制約として用いることは賢明ではない．それに対して著者らは，統計情報を用いた文節間係り受け確率モデルの定式化について検討し、上述の性質に反する係り受けを特定するためのロバストな解析手法の実現を試みている [4]．そこで本研究では，対話音声の解析に適した係り受け確率モデルを構築することを目的として，(1) 係り先がない文節、(2) 前方修飾、(3) 係り受けの交差、(4) 発話単位を跨ぐ係り受け、についてそれぞれの出現頻度及び特徴的現象を分析した。分析には，前節で述べた係り受けコーパスを用いた。分析の対象とした 81 対話には、 24,256 個の文節が存在し、 11,859 個

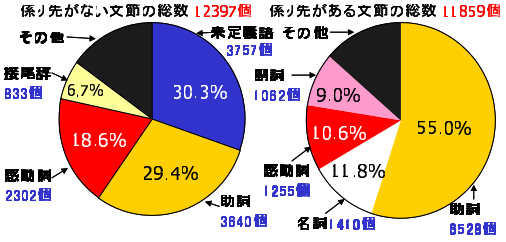

図 1: 係り先のない文節の品詞による分析

図 2: 前方修飾の品詞による分析

の係り受けが存在した。 1 発話単位あたりの平均係り受け数は 1.52 個であり、 1 ターンあたりでは 1.95 個であった。分析結果は以下の通りである。

(1) 係り先がない文節　総文節の約 51.1% が係り先がない文節であった。その特徴を調べるため、係り先がない文節の最終形態素の品詞の分布を調査し、係り先がある文節の場合と比較した。結果を図 1に示す。フィラーや言い淀みなどの未定義語の出現は係り先がない文節に特徴的である。

(2) 前方修飾　250 個存在した。全ターン 6,078 個の 4.1% に出現しており、無視できる頻度ではない。その中の 85.6% に相当する 214 個がターンの最後の文節に出現しており、文節の位置情報が前方修飾の特定に有用であることがわかる。一方、前方修飾の起こった係り受けの文法的な特徴を明らかにするために、係り文節の最後の形態素の品詞と受け文節の最初の形態素の品詞の組の分布を調べた。前方修飾が起こってない場合との比較を図 2に示す。 2 つのグラフに大きな違いはなく、前方修飾との関係は見当たらない。

(3) 交差　交差する係り受けは 9 回あるが、全係り受けの 0.08% に過ぎない。

(4) 発話単位を跨ぐ係り受け　発話単位に跨る係り受けは 92 個存在した。これは、複数の発話単位から構成されている 1,362 ターンの 6.8% に出現することを意味する。従って、発話単位を話し言葉解析の処理単位とすることは、必ずしも適当でないと判断される。

今後は、以上の分析結果をもとに、話し言葉の統計的係り受け解析手法を定め、それを用いた対話音声の解析実験を実施する予定である。

参考文献

[1] Kawaguchi, N., et al.: Multimedia Data Collection of In-Car Speech Communication, *Proc. of 7th Eurospeech*, 2027–2030 (2001).

[2] 黒橋ら: 京都大学テキストコーパスプロジェクト，第 3 回言処年大， 115-118 (1997).

[3] 前川ら: “日本語話し言葉コーパスの設計”, 音声研究, 4(2), 51-61(2000).

[4] Matsubara, S., et al.: Stochastic Dependency Parsing of Spontaneous Japanese Spoken Language, *Proc. of 19th COLING* (2002).