言語情報に基づくジェスチャーの決定
－プレゼンテーションエージェントにおけるジェスチャー生成－

中野有紀子†　　岡本雅史‡　　李清‡

†科学技術振興機構社会技術研究システム　　‡東京大学大学院情報理工学系研究科

{nakano, okamoto, liqing}@kc.t.u-tokyo.ac.jp

1. はじめに

コンピュータグラフィックスの発展に伴い，表現力の高いアニメーションキャラクターが開発され，人対コンピュータのインタラクションを媒介するインタフェースエージェントが教育用コンテンツや各種 Web ページ等で頻繁に利用されるようになってきた．このようなエージェントのユーザインタフェースとしての魅力の１つは，その顔や体を使って表情やジェスチャーなどの非言語情報を音声言語に付与することができることである．

人間のコミュニケーション行動に関する従来研究では，様々な非言語情報の中でも特にジェスチャーは音声言語の理解を高める効果を持つと報告されている．音声がジェスチャーを伴って与えられた場合には，理解の正確性は，音声のみの場合と比べて約２倍に向上し[1]，さらに，口の動きや顔の表情よりもジェスチャーを伴った刺激のほうがよりよく理解されることが実験により実証されている[2]．従って，Reeves and Nass[3]が主張するように，人がエージェントに対して，人に対してと同様の反応をすると仮定すれば，音声言語と同期して適切なジェスチャーを自動的に生成することは，より効果的なインタフェースエージェントを実現するための重要な研究課題となる．

ジェスチャーの自動生成に取り組んだマルチモーダル生成に関する研究では，ジェスチャーは説明システムにおける教示内容[4, 5]，仮想学習環境におけるタスク状況[6]，あるいは会話エージェントにおける発話意図[7]に応じて決定される．しかし，これらのシステムでは，コンテンツを追加するために，コンテンツ作成者（例えば，学習教材を用意する教師など）が論理的な意味表現を記述しなければならず，大きな障害となる．一方，[8]は，テキストを入力すると，それに対するエージェントの動作を自動的に決定し，合成音声と同期したエージェントの動作スケジュールを出力するツールを提案している．しかし，従来研究では，入力されたテキスト中のどのような情報を使ってジェスチャーを決定すべきか，表層的な言語情報からジェスチャーをどの程度適切に決定することができるのかについての議論は少なく，また関連した言語学理論は計算モデルとして利用できるような形式になっていない．

以上の議論に基づき，本研究では以下の課題に取り組む．

- 言語表現を解析することによって得られる語彙的／統語的情報はテキストにジェスチャーを付与する上で，有効であるのか．
- もしそうであるならば，テキストからどのような情報を抽出し，それをエージェントの動作決定機構の中でどのように利用するのか．

以上の課題を解決するために，まず人によるプレゼンテーションのデータを収集し，ジェスチャーの出現とそこでの言語的特徴との関係について分析する．さらにこの分析結果に基づき，テキストからエージェントによるプレゼンテーションを自動生成するシステムを構築し，その機構について述べる．

2. 背景

ここでは，言語表現とジェスチャーの出現との関連について，言語学の理論を参考に考察する．

参照表現の言語的分量：McNeill[9]は，Firbas[10]により提案された「伝達のダイナミズム」（Communicative Dynamism (CD)）－所与のメッセージがコミュニケーションを前進させる度合い－という概念を変数として用い，CD が大きくなるほどジェスチャーが出現しやすいと述べている．さらに McNeill は，CD の尺度として「参照表現を構成する言語的情報の量」を取り上げている．例えば，代名詞は通常の名詞よりも CD が低く，名詞は修飾句／節を伴う名詞よりも CD が低い．この McNeill の議論は，文中の修飾関係を調べることにより，CD を，さらにはジェスチャーの出現位置を推定できることを示唆している．

主題・題述：さらに，McNeill は，文の主題は通常最も CD が低く，一般的にはジェスチャーを伴わない，ジェスチャーは通常題述（主題について述べている部分）に付随すると論じている．ここで注目すべき点は，英語の場合は，主題・題述の認定が語順だけからでは困難であるが，日本語の場合には，提題助詞（eg. 「は」等）と呼ばれる主題を有標化する言語的手段が備わっているということである．従って，日本語に特化すると，提題助詞で有標化されていれば，統語解析によって文の主題部分を同定し，ジェスチャーが出現しにくい場所として識別することができるはずである．

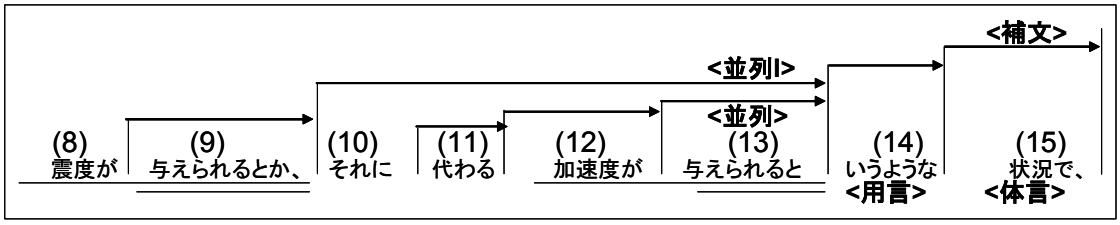

図 1：係り受け構造とジェスチャー出現位置との関係
下線部はジェスチャーの出現部分を，二重下線部はストローク部分を示す

さらに，WH 疑問文においては，文頭の疑問詞は文内容の方向を示す標識であり，疑問詞で尋ねているものが焦点であると考えられるが，疑問詞を主題とするのか題述とするのかは議論が分かれている[11, 18]．従って，文の情報構造において特殊な位置づけにある疑問詞というものがジェスチャーを誘発する要因になるのかどうか，データ分析により明らかにする必要がある．

新・旧情報：主題・題述による文の情報構造と深くかかわる概念として新・旧情報がある．通常，旧情報は題述性が低く，新情報は高い．このことを利用すると，初めて言及された表現であるか否かを調べることによって，表層表現から題述性が推定でき，さらには，題述性の低い旧情報よりも，高い新情報の方がジェスチャーを伴いやすいと予想される．

対比的関係：談話中の重要な概念間の対比関係を明確化するためにアクセントが用いられるが[12], [13]は，この研究[12]と Kendon の韻律とジェスチャーについての理論[14]に基づき，対比的ジェスチャーの意味表現からの生成を行っている．このことを統語的に捉えてみると，同様の関係を統語的な並列構造を見つけることにより同定できる可能性がある．

図 1は統語分析として係り受け分析を用い，文中の各文節間の係り受け関係とジェスチャーの出現位置とを示したものである．文節(8)～(9)と(10)～(13)は並列関係にあり，両者ともジェスチャーを伴っている．また，文節(14)は補文であり，文節(15)の体言に係り，修飾している．つまり，文節(15)は節によって修飾された言語的分量の大きい体言である．

3. データ分析

どのような言語的特性がジェスチャーの出現に関係しているのかを調べるために，7 人分のプレゼンテーションと質疑応答をビデオに録画し，各発表者につき約３分分を書き起こした．書き起こされたデータは合計 2124 文節，その中に 343 のジェスチャーが含まれていた．

ジェスチャーの分析：約半数のデータに関しては，３人の分析者が議論してジェスチャーの出現位置を認定した．ジェスチャーの認定方法について一致した基準が得られた後，一人の分析者が残りのデータを分析した．ジェスチャーは準備期，ストローク，消失期とから構成され，ストロークはもっとも目立った音節と共起することが多い[14]．そこで，ジェスチャーの開始と終了の時間に加え，ストロークを打った時間もアノテーションに加えた．

言語情報の分析：日本語係り受け解析器[15]を用いて，以下の項目に関して各文節の特徴を計算した[1]．

(a) 所与の体言が節や補文により修飾されているか（参照表現の言語的分量）
(b) 体言に後続している助詞の種類（「は」，「が」，「を」，その他）
(c) 所与の文節が疑問詞を含むか
(d) 所与の文節中の全ての内容語が１つ前の文中で既に述べられているか（新／旧情報）
(e) 所与の文節が並列句を構成しているか

以上の統語的な要因に加えて，以下の語彙的な特徴に関しても分析を行った．

(f) 所与の文節が強調の副詞（eg. 大変，非常に）を含むか，あるいは強調の副詞を含む文節の直後の文節であるか
(g) 所与の文節が cue word（eg., それでは，従って）を含むか，あるいは cue word を含む文節の直後の文節であるか
(h) 所与の文節が数詞（eg., 数千人，99 回）を含むか

Cassell[16]は，ジェスチャーは話の区切りを示すと述べており，これは，談話の区切りを示す cue word がジェスチャーと共起しやすいことを示唆している．

以上の分析方法を用い，ジェスチャーストロークの出現位置と各文節の言語的特性との相関関係を調べた．

結果：分析結果を表 1に示す．ジェスチャー出現のベースラインは 10.1%である（つまり，約 10 文節に 1 回の頻度でジェスチャーが出現する）．体言が節によって修飾されている場合，ジェスチャーの出現確率は 38.2%である．代名詞やその他の体言が格助詞の「を」を伴って文節を構成し，かつその文節に新情報（前の文で言及されていない内容語）が含まれている場合には，28.1%の場合においてジェスチャーが共起している．

さらに，並列構造の一部となる文節であれば，ジェスチャーストロークが起こる確率は 47.7%であり，疑問詞を含む文節である場合には 41.4%，cue word を含む文節である場合には 41.5% となる．また，強調の副詞に続く文節である場合や数詞を含む文節に

[1] 解析器の係り受け解析エラーによる影響を取り除くため，全データの約 13%において係り受けに関する誤りを人手で修正した．

おいてもジェスチャーの出現確率が比較的高い（出現確率はそれぞれ，35%，39.3%）．

以上に示すように，表 1に列挙された言語的特徴を持つ文節では，それ以外の文節に比べて，ジェスチャーの出現確率が約 3〜5 倍となる．また，このモデルは観測されたジェスチャーの約 75%を説明することができる．これらの分析結果から，ジェスチャーを使うべき場所と使うべきでない場所とを識別する上で，語彙的・統語的情報が有用であることが明らかになった．

表 1：分析結果

条件				ジェスチャー出現率
[C1]	参照表現の言語的分量	(a) 節により修飾された体言		0.382
[C2]		代名詞，その他の体言	(b) ヲ格 & (d)新情報	0.281
[C3]	(c) WH疑問詞			0.414
[C4]	(e) 並列句			0.477
[C5]	強調の副詞	(f) 強調の副詞		0.244
[C6]		(f') 強調の副詞に後続		0.350
[C7]	(g) Cue word			0.415
[C8]	(h) 数詞			0.393
[C9]	その他（ベースライン）			0.101

4. システムの実装

4.1. システムの概要

本節では，画像／映像，音声，エージェントアニメーションを統合した放送型メディアを自動生成する Web アプリケーション，SPOC(Stream-oriented Public Opinion Channel) [17]上で動作する会話エージェントシステム CAST(The Conversational Agent System for neTwork applications)を提案する．SPOC の番組視聴画面と CAST のシステム構成を図 2に示す．

CAST はテキストを入力とし，エージェントのアニメーションスケジュールとエージェントの発話となる合成音声を自動的に計算し，出力する．図 2に示すように，CAST は (1)エージェント動作決定機構(Agent Behavior Selection Module (ABS))，(2)言語タグ付与機構(Language Tagging Module (LTM))，(3) エージェントアニメーションシステム，(4)音声合成装置，の 4 つの主要モジュールからなる．CAST に入力されたテキストは，まず ABS に送られる．ABS がそのテキストを LTM に送ると，言語情報がタグ付けされ，ABS に返される．ABS は，この言語情報を用いて，どの文節でどんな動作を用いるかを決定する．そして，顔表情やジェスチャー等のエージェントの動作タグをテキストに付与する．最後に，これがタイムスケジュールに変換されてエージェントアニメーションシステムで実行される．

4.2. エージェント動作の決定

言語情報タグの付与：まず ABS では，LTM が入力テキストを解析し，3節で議論された言語情報を自動的にテキストに付与する．例えば，図 1の文節(9)に以下のような言語情報が付与されたとする．

｛テキスト ID:1，文 ID:1，文節 ID:9，係り受け_from:8,係り受け_to:13，文節タイプ:用言，言語的分量：NA，格：NA，WH 疑問：false，新／旧情報：新，並列関係：13，強調副詞：false，Cue-Word：false，数詞：false｝

例えば上の例では，この文節が含まれるテキストの ID が 1，文の ID が 1 であり，この文節の ID は 9 である．文節 8 がこの文節に係っており，この文節は文節 13 に係る．この文節は新情報を伝達し，文節 13 と並列関係にある．

ジェスチャーの決定：次に ABS は，各文節に対し，ジェスチャーを付与すべきかどうかを表 1に示すデータの分析結果に基づいて決定する．例えば，先に示した例では，文節(9)は表 1の[C4]（文節が並列構造の構成素である）の場合に当てはまるが，この場合には，システムは 47.7%の確率でジェスチャーを該当文節に付与する．現在のシステムでは，ジェスチャーの形態のデフォルトとしてビートジェスチャーを採用している．ビートジェスチャーとは，手を上下に振るような身振りであり，発話の意味内容とは直接的に関連せず，発話の中で強調される部分に出現しやすい．一方，強調される文節中の概念に対して，特定のジェスチャーがエージェントアニメーションシステムのライブラリに定義されている場合（例えば，「大きい」という概念を表現するジェスチャーがライブラリに既に登録されている場合）には，それが優先して用いられる．ジェスチャーが決定されると，エージェント動作タグが XML 形式で付与される．ABS ではジェスチャーに加えて，表情や視線などの動作も決定されるが，詳細は[17]に譲る．図 3にエージェント動作の XML の例を示す．この XML では，2 番目と 6 番目の文節で対比のジェスチャーを行い，8 番目の文節でビートジェスチャーを行うことを指示している．

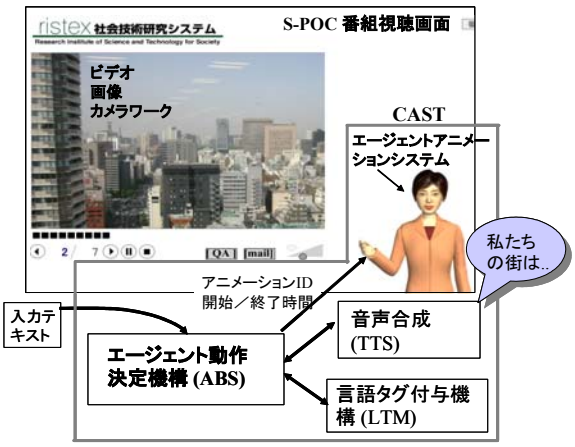

図 2：SPOC-CAST の概要

```
<Gaze type="towards">
   震度が
        <Gesture_right type="contrast" handshape_right="stroke1@2">
        与えられるとか
        </Gesture_right>
   それに
   代わる
   加速度が
        <Gesture_right type="contrast" handshape_right="stroke2@2">
        与えられると
        </Gesture_right>
   いうような
        <Gesture_right type="best" handshape_right="stroke1">
        状況で
        </Gesture_right>
     ...
```

図 3：エージェント動作スクリプト例

4.3. スケジューリング

最後に動作スクリプトの XML はタイムスケジュールに変換される．XML のテキスト部分を音声合成装置に入力することにより，ABS は音素や文節区切りの時間情報を合成エンジンから取得し，これをもとにエージェント動作のタイムスケジュールを計算すると同時に，読み上げの合成音声をファイルに保存する．ABS から出力されたタイムスケジュールは，アニメーションのコマンド列としてエージェントアニメーションシステムで解釈され，スケジュール通りに実行される．

5. 議論

本稿では，テキストにジェスチャーを自動的に付与し，ナレーション音声と同期したエージェントアニメーションを出力するメディア変換技術を提案した．そのために，まず，実際のプレゼンテーションのデータを分析し，テキストにジェスチャーを付与するために有効な語彙的／統語的情報を統計的に明らかにした．特に，文節が並列構造を構成素する場合には，約半数の場合においてジェスチャーが付随し，また，節によって修飾されている場合にもジェスチャーの出現頻度が高くなることがわかった．これらの結果は，助詞のタイプや新／旧情報といった局所的な情報を用いて推定された主題・題述性よりも，係り受け関係から得られる文全体の構造的情報のほうがジェスチャー決定にはより有用であることを示唆している．

本稿で提案したモデルでデータの全てを説明できるわけではない．残りはさらにモデルを精緻化していくことにより，説明されなければならない．そのためには，意味や語用論の情報を統合したモデルへと改良していくことが今後の課題である．

参考文献

1.Berger, K.W. and G.R. Popelka, *Extra-facial Gestures in Relation to Speech-reading.* Journal of Communication Disorders, 1971. **3**: p. 302-308.

2.Rogers, W., *The Contribution of Kinesic Illustrators towards the Comprehension of Verbal Behavior within Utterances.* Human Communication Research, 1978. **5**: p. 54-62.

3.Reeves, B. and C. Nass, *The Media Equation: how people treat computers, televisions and new media like real people and places*. 1996, Cambridge: Cambridge University Press.

4.Andre, E., T. Rist, and J. Muller, *Employing AI methods to control the behavior of animated interface agents.* Applied Artificial Intelligence, 1999. **13**: p. 415-448.

5.Rickel, J. and W.L. Johnson, *Animated Agents for Procedural Training in Virtual Reality: Perception, Cognition and Motor Control.* Applied Artificial Intelligence, 1999. **13**(4-5): p. 343-382.

6.Lester, J.C., B. Stone, and G. Stelling, *Lifelike Pedagogical agents for Mixed-Initiative Problem Solving in Constructivist Learning Environments.* User Modeling and User-Adapted Interaction, 1999. **9**(1-2): p. 1-44.

7.Cassell, J., M. Stone, and H. Yan. *Coordination and Context-Dependence in the Generation of Embodied Conversation*. in *INLG 2000*. 2000.

8.Cassell, J., H. Vilhjalmsson, and T. Bickmore. *BEAT: The Behavior Expression Animation Toolkit*. in *SIGGRAPH 01*. 2001.

9.McNeill, D., *Hand and Mind: What Gestures Reveal about Thought*. 1992, Chicago, IL/London, UK: The University of Chicago Press.

10.Firbas, J., *On the Concept of Communicative Dynamism in the Theory of Functional Sentence Perspective.* Philologica Pragensia, 1971. **8**: p. 135-144.

11.Halliday, M.A.K., *The linguistic study of literary texts*, in *Essays on the Language of Literature*, S. Chatman & S. R. Levin, Editors. 1967, Houghton Mifflin: Boston.

12.Prevost, S.A. *An Informational Structural Approach to Spoken Language Generation*. in *34th Annual Meeting of the Association for Computational Linguistics*. 1996. .

13.Cassell, J. and S. Prevost. *Distribution of Semantic Features Across Speech and Gesture by Humans and Computers*. in *Workshop on the Integration of Gesture in Language and Speech*. 1996. Newark, DE: WIGLS.

14.Kendon, A., *Some Relationships between Body Motion and Speech*, in *Studies in Dyadic Communication*, A.W. Siegman and B. Pope, Editors. 1972, Pergamon Press: Elmsford, NY. p. 177-210.

15.Kurohashi, S. and Nagao, M., *A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures.* Computational Linguistics, 1994. **20**(4): p. 507-534.

16.Cassell, J., *The Development of the Expression of Time and Event in Narrative*, in *Department of Psychology/Linguistics*. 1991, The University of Chicago: Chicago, IL. p. 233.

17.Nakano, Y.I., Murayama, T., and Nishida, T., *Multimodal Story-based Communication: Integrating a Movie and a Conversational Agent.* IEICE Transactions, (to appear).

18. 野田尚史. 1996. 『「は」と「が」』，新日本語文法選書 1，東京：くろしお出版.