幼児の日常生活における対話データの分析と名詞概念獲得システム SINCA への適用

内田 ゆず　　荒木 健治
Yuzu Uchida　　Kenji Araki
{yuzu,araki}@media.eng.hokudai.ac.jp

北海道大学大学院 情報科学研究科
Graduate School of Information Science and Technology, Hokkaido University

1. はじめに

人間の言語活動は，対話の中から言語を獲得する，記憶に基づいた対話を行う，質問に対して論理的な応答を行う，ユーモアや皮肉を理解する，環境や相手に適した言葉を選択する，感情を表現する，などの様々な能力から成り立っている．我々は，人間と同等の言語能力を持つロボットの実現を目指している．その研究の第一歩として，ユーザの画像に対する発話から名詞概念（画像に対するラベル) を獲得していくシステムである SINCA（System for Noun Concepts Acquisition from utterances about Image）の構築・提案を行った[1].

我々は，幼児が実際に日常的に聞いている対話データを分析することで，SINCA をさらに改善するための手掛かりが得られるのではないかと考えている．そこで我々は，幼児の日常生活を撮影したビデオデータを書き起こすことによって独自の対話データの作成を行った．撮影対象となった家族は，二人の幼児とその両親の 4 人で構成されている．撮影期間は約 4 ヶ月間であり，約 82 時間分のビデオデータが得られた．これまでに，名詞概念獲得システム SINCA にこのデータを入力して行った実験についても報告している[2].

本稿では，長男に向けられた発話と長女に向けられた発話の違いなど，対話データをさらに詳細に分析した結果について報告する． また，データの分析結果と SINCA への適用結果について考察する．

2. 対話データの概要

幼児がいる家庭における日常生活をビデオカメラ（Everio GZ-MG255，Victor）を用いて撮影を行った．撮影対象となった家族は， 日本語を母語としており，2 歳 7 ヵ月の男児，12 ヵ月の女児（ビデオ撮影開始時の年齢）とその両親の 4 人で構成されている．

多くの場合，撮影場所は上記の家族が住む家の中であった．ただし，まれに自家用車内や祖父母宅，

図 1　ビデオに収録された日常風景

駅などの外出先で撮影された場合もある．撮影は母親か父親がビデオカメラを手で持った状態か，箪笥の上などのやや高い位置に固定した状態で行われた．図 1 に収録された場面の例を示す．家庭における自然な発話が行われる場面の撮影を目的としたので，課題などは一切設けていない．

撮影は 2007 年 7 月 21 日から同年 11 月 23 日にかけて行われ，約 82 時間分のビデオデータが得られた．

第一著者が得られた動画データを視聴し，その発話内容の書き起こし作業を行った．書き起こし対象は大人の発話（主に両親）であり，なるべくできる限り発話の通りにひらがな表記で記録を行った．音量が小さく内容が聞き取れない場合など，書き起こしが不可能な音声は”***”の記号で記録した．

実際の発話の一部を以下に示す．文中の D は長女の名前を表す．

- もうあいすかわないよおとうさん（父親→長男）
- ほらかまれるわありんこに（父親→長男）
- ねむいのかな D ちゃんは（父親→長女）
- うえひこうきとんでるひこうき（母親→長男）
- こおりとって（母親→長男）
- りんごいまおうちにないもん（母親→長男）
- なにたべてんの D（母親→長女）

3. 対話データの分析

2.で述べた方法で書き起こしを行った対話データの一部を分析し，発話者と発話対象の違いが発話内容などにどのような影響を及ぼすのかの調査を行った．分析対象とした発話は 3,954 文である．

3.1 発話回数と発話の長さ

図 2 に発話者と発話対象ごとの発話回数を示す．母親から長男への発話が最も多く，全体の 41.6%を占めている．それに対し，母親から長女への発話は全体の 4.8%となっており，最も少ない．父親も長男への発話が多く，長女への発話は少ない傾向にある．長男は不完全ながらも 1 語～2 語の発話を行うことができ，両親の発話に応答を返すことができる．自ら両親に向けて話しかけることもある．したがって，長男は両親と対話を行うことができる．一方で，長女はまだ話すことはできない．この違いが両親の発話回数に影響していると考えられる．

図 3 に発話者と発話対象ごとの発話の長さを示す．両親同士の発話は長く，子どもに向けた発話は短いという特徴が見られる．長男への発話と長女への発話を比較すると，母親から長女への発話が平均して 0.5～0.9 文字ほど短くなっているが，顕著な差は見られなかった．

ビデオに収録された親子の対話の中で特徴的な現象として，幼児の発話に対して大人が行うオウム返しがあった．オウム返しは幼児の言語獲得に重要な役割を果たすことが報告されている[3]．今回分析を行ったデータの中には，91 回のオウム返しが存在する．長女はまだ話すことができないこと，父親は母親に比べると長男の発話を正しく理解できないことから，必然的に母親が長男の発話に対して行うオウム返しとなっている．長男が正確ではない発音で発話を行い，母親はそれを正しく解釈し，正確な発音でオウム返しを行う様子が多く観察された．オウム返しが幼児の発音などを修正するためのフィードバックとして機能している点を参考にして，我々が開発した名詞概念獲得システム SINCA のフィードバック部分にもオウム返しを用いることを検討している．

ビデオデータの中には，呼びかけも多く含まれていた．両親の発話の中で，長男の名前が含まれているものは 331 文（全体の 13.1%），長女の名前が含まれているものは 178 文（全体の 35.6%）存在した．大人同士の発話には呼びかけはほとんど見られない（全体の 0.8%）ため，幼児に対する発話の特徴であると考えられる．

3.2 品詞の分布

発話者や発話対象によって発話に含まれる品詞に特定の傾向が現れるかの調査を行った．収集した発話文を話者，発話対象ごとに分類し，形態素解析を行った．分析対象となる発話文はオノマトペ[1]を含んでいる，ひらがなで表記されている，などの特徴がある．形態素解析器は，連濁，反復形オノマトペを辞書に登録するのではなく，動的に認識を行う JUMAN[2]（ver.6.0）を使用した．図 3 に両親から長男への発話，両親から長女への発話，両親の間で行われた発話にそれぞれ形態素解析を行った結果を示す．品詞の分類は，JUMAN の仕様によるものである．この結果から，発話対象の違いには関係なく，名詞，動詞，助詞の割合が高いことが明らかになった．また，幼児を対象にした発話と大人を対象にした発話では，助詞の数に約 6 ポイントの差があることが明らかになった．幼児に対する発話は，1 語発話が多く含まれることに加え，助詞によって格関係を明示する必要のない単純な内容であるためだと考えられる．

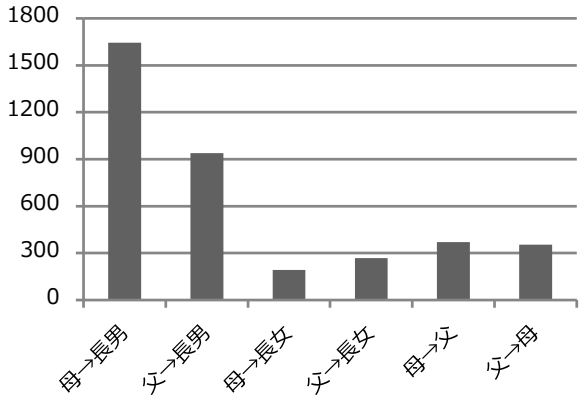


図 2　発話回数

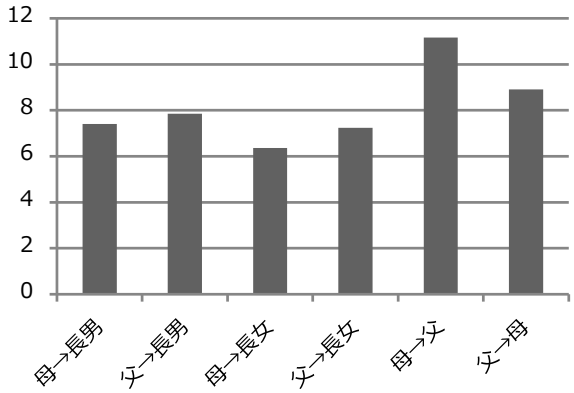
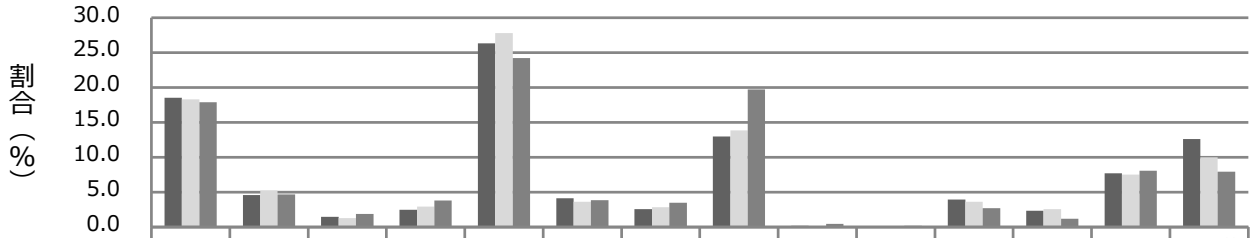


図 3　発話の平均文字数

[1] 擬声語．擬音語と擬態語の総称．
[2] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html



	動詞	形容詞	判定詞	助動詞	名詞	指示詞	副詞	助詞	接続詞	連体詞	感動詞	接頭辞	接尾辞	未定義語
■親→長男	18.5	4.6	1.5	2.5	26.3	4.1	2.6	13.0	0.2	0.1	3.9	2.3	7.7	12.6
■親→長女	18.3	5.2	1.3	2.9	27.8	3.6	2.8	13.9	0.1	0.1	3.6	2.6	7.5	10.0
■親→親	17.9	4.7	1.9	3.8	24.2	3.9	3.5	19.7	0.5	0.2	2.7	1.2	8.1	7.9

図 4　発話者と発話対象ごとの品詞分布

4. 名詞概念獲得システム SINCA の概要

4.1 入力

入力は画像と文の対である．入力画像は Web カメラ（USB-CAMCHAT2／アイ・オー・データ機器. 有効画素数：30 万画素）からキャプチャされた画像（以降画像 P と呼ぶ），入力文は画像 P を見せながらユーザが幼児に話かける発話 1 文（以降文 S と呼ぶ）である．入力画像は，ユーザが自由に被写体を選び撮影するものである．

入力文は全てひらがなで表記され，入力文に形態素解析などの前処理は一切施されない．ひらがな表記を用いているのは，ユーザによって表記に揺れが生じることと，入力された文字列自体に意味が含まれてしまうことを避けるためである．また，形態素解析などの前処理を行わないのは，未知語に対応するためである．

4.2 画像認識

過去に同じ被写体が写った画像が入力されたかどうかを判断する．ここでは，エボリューション・ロボティクス社 の ERSP3.1 (Evolution Robotics Software Platform) に含まれる”ERSP ビジョン”を用いた．ERSP 3.1 は，ロボット製品の作成を目的とした総合開発プラットフォームで，照明や物体の位置が管理されていない現実的な環境の中でも，ロボットやシステムが２次元と３次元の物体を認識することができる画像認識ツールである．

4.3 共通部分抽出

システムは入力を得ると，過去に画像 P とともに入力された文と文 S を比較して，字面が一致する文字列を切り出す．この切り出された文字列を共通部分と呼ぶ．これ以降の処理で共通部分は，画像 P に対応するラベルの候補として扱われる．

4.4 基本スコア付与と出力

抽出された共通部分には基本スコアが付与される．基本スコアとは，その共通部分のラベルとしての確からしさを表した値であり，出現頻度が高く，文字数が多く，他の画像と共に出現することのない共通部分ほど高いスコアが与えられる．

基本スコアが閾値 S を超えた共通部分は，画像 P に含まれる事物のラベルに適している可能性が高いと判断され，テキストとして出力される．

4.5 出力に対する評価

システムの出力に対してユーザは次の 3 つのキーワードのうち，最も相応しいものを選び，入力する．

・じょうず：ラベルとして適切である
・おしい　：ラベルとしては適切でないが
　　意味はわかる
・ちがうよ：意味がわからない

幼児がこれらのキーワードを完全に理解するとは考えられないが，実際には，大人の表情や声の調子で感じ取ることのできる情報は多い．本手法ではそれらの情報の代わりにキーワードを用いることとする．ユーザの反応によってその共通部分のスコアの再計算を行う．「じょうず」の場合はスコアを増加させ，「おしい」の場合は維持，「ちがうよ」の場合は減少させる．

4.6 名詞概念獲得

「入力」から「ユーザの評価」の処理を繰り返した結果，再計算されたスコアが閾値 T を超え，さらに「じょうず」という評価を得たことがある共通部分は画像 P のラベルとして獲得される．

4.7 ラベル獲得ルール生成

ラベル獲得ルールとは，再帰的な名詞獲得を行うためのルールである．人間は過去に得た知識を活用し，より効率的に学習を進めていく．本手法ではそのような再帰的な学習を以下で述べるように実現している．

システムが文字列 S1 をある事物に関する正しいラベルとして獲得すると，その事物に関する過去の入力文のうち，文字列 S1 を含む文から，ラベル獲得ルールを生成する．ラベル獲得ルールとは，図 3 のようにラベルの部分を変数とすることで，入力文を抽象化したものである．次に，生成したラベル獲得ルールに合致する入力文があった場合，変数部@1 に相当する部分を切り出し，スコアを上昇させる．

5. SINCA を用いた実験

5.1 実験方法

2.で述べた対話データを SINCA に入力した場合，名詞概念を獲得可能であるかを検証する実験を行う．ビデオ撮影によって収集した発話文のうち，親子の間で共同注意が成立している場面の発話を入力文の候補とする．さらにその中で出現頻度が高い 10 種類の名詞（あいす，あり，ばす，こおり，らーめん，おせんべい，とんねる，りんご，じゅーす）に関する 353 文からランダムに選択したものを入力文とする．入力文はキーボードで入力し，入力画像は上記の名詞に対応した静止画像を Web カメラで撮影したものを用いた．

5.2 実験結果

実験の結果，10 種類の画像に対して適切なラベルを対応付けることに成功した．一つの名詞概念を獲得するまでに必要な入力回数は，5.3 回であり，分散は 0.21 であった．獲得されたラベル獲得ルール数は，44 個であった．ラベル獲得ルールが適用された回数は，1 回であった．

5.3 考察

ビデオ撮影によって収集した入力文を用いた実験から，SINCA は実際に幼児が聞いている発話を入力として用いた場合，名詞概念を獲得可能であることが明らかになった．

3.1 で述べたように，幼児への発話は一語発話が多く，短い傾向がある．実験で入力文の候補とした 353 文のうち 66 文（18.6%）が一語発話であり，そのうち 7 文が実際の入力文として用いられた．また，入力として用いられた 53 文のうち，23 文（42.6%）は本来必要であるはずの助詞が欠落していた．

SINCA は字面の一致のみを手がかりに名詞の候補の抽出を行う． 助詞は種類が豊富ではないため，SINCA が入力文から共通部分を抽出する際に，ラベルに助詞を伴った文字列を抽出することがある．助詞が含まれない入力文を用いると，ラベル部分に隣接する文字の種類が大幅に増えるため，このような誤りを回避することが可能になり，正確なラベルの獲得が促進される．同様に，一語発話も SINCA の名詞概念獲得手法には有効である．

6. まとめ

幼児の日常生活を収録したビデオデータを書き起こし，大人の発話の分析を行った．分析の結果，大人は短い発話で幼児に話しかけることが明らかになった．また，大人の発話に形態素解析を行った結果，幼児に対する発話には助詞が少ないことが明らかになった．

上記のデータを入力文として名詞概念獲得システム SINCA の名詞概念獲得実験を行った．SINCA は形態素解析などを用いず，字面の一致のみを手がかりに名詞の候補を抽出する．したがって，名詞に隣接する文字が多様な（あるいは存在しない）文の集合が理想的な入力となる．これは大人が幼児に話しかける発話文の特徴に合致する．SINCA への入力文には，日本語文法に即した文よりも，大人が幼児に対して日常的に発話しているような文が適していることが示唆された．

一方で，日本語において助詞は動詞の格関係を示すなど，文の成立に重要な機能を果たしている．幼児が動詞を獲得する過程では助詞から得られる情報も必要だと考えられる．助詞と名詞獲得，動詞獲得との関係については今後の研究課題である．

参考文献

[1] 内田ゆず, 荒木健治, "画像に対する発話を対象とした名詞概念獲得システム SINCA", 知能と情報（日本知能情報ファジィ学会誌）, Vol.20, No.5, pp.685-695, Oct. 2008.

[2] Yuzu Uchida and Kenji Araki, "Evaluation of a System for Noun Concepts Acquisition from Utterances about Images (SINCA) Using Daily Conversation Data", Proceeding of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT) 2009 conference, pp.65-68, Jun. 2009.

[3] 正高信男, “0 歳児がことばを獲得するとき―行動学からのアプローチ”, 中公新書, 1993.