Vol.2020-ICS-198 No.1
2020/3/7

雑談型会話システムにおける共通評価指標の提案

中村健太[1]　土屋誠司[2]　渡部広一[3]

概要：本研究では雑談型会話システムを評価する尺度として行動指標を提案する．これまで雑談型会話システムの研究では，各システムの開発目標に沿ってアンケートを用いた評価項目が作成されてきた．そのため，システム間での比較が困難であり，またユーザからの受けは軽視されてきた．そこで①ユーザの快適さ，②会話の自然さ，③会話の人間らしさ，を測定する行動指標を提案した．その結果，ユーザの快適さは会話時間の長さで測定できることを示した．しかし残りの 2 点に関しては今後の検討が必要であることが示唆された．

キーワード：雑談型会話システム，行動指標，ユーザ視点

Suggestion of the Common Evaluation Index Used Non-task-oriented Dialogue System

KENTA NAKAMURA[†1]　SEIZI TSUTIYA[†2]
HIROKAZU WATABE[†3]

Abstract: In this study, we propose a behavior index as a measure to evaluate non-task-oriented conversation system. Until now, research on non-task-oriented conversation systems has created evaluation items using questionnaires in accordance with the development goals of each system. For this reason, it is difficult to make a comparison between systems, and the reception from users has been neglected. Therefore, we proposed an behavior index that measures (1) user comfort, (2) naturalness of conversation, and (3) humanity of conversation. As a result, it was shown that the comfort of the user can be measured by the length of the conversation time. However, it was suggested that the remaining two items require further study.

Keywords: non-task-oriented conversation system, behavior index, evaluation from user

1. はじめに

近年，ロボットは様々な場面で人間のパートナーとして活躍することが望まれている．しかし日常的な場面で活躍させるためには依然として問題が多い．例えば教育場面では，ロボットを理科の授業のアシスタントとして参加させて，子どもたちの理解を促進させる試みが行われた[1]．しかし，授業への理解度に影響は見られなかった．この原因として，ロボットは限定的な応答しかできず，子どもからの呼びかけに対して自然なコミュニケーションが取れなかったことが挙げられている．また，家電製品との音声コミュニケーションはユーザのストレスを軽減し，ユーザをリラックスさせることが示唆されている[2].

これらから，ロボットを人間のパートナーとして活躍させるためにはコミュニケーション問題を解決する必要があるし，またその意義は大きいと考えられる．人間の基本的なコミュニケーション方法は, 自然言語を用いた意思疎通, すなわち会話である．そのため，ロボットが人間のパートナーとして活躍するためには会話を自然に行うシステムが必要である．

自然に会話を行うことができる会話システムが求められているが，会話システムはこれまで機能的な面で 2 種類に分類されてきた．1 つがタスク指向型の会話システムであり，ユーザの質問や要求に対して情報の提供を行うなどの特定のタスクを達成するものと定義される[3][4]. もう一方は非タスク指向型の会話システムであり，これは日常的な雑談を行うことでユーザを楽しませることが目的である会話システムと定義される[3][4].

タスク指向型の会話システムは古くから研究が行われている．例えば，吉田ら[5]は視覚障害者に向けて音声入力と音声フィードバックができるエアコン用のリモコンと通常のリモコンの比較を行った．それぞれのリモコンを用いてあらかじめ定められたエアコンを操作する課題を視覚障害者と目隠しされた健常者に行わせた．この時リモコンの操作時間と課題の達成度が測定された．その後，実験参加者には「将来的に使いたいか」を 5 件法で質問した．その結果，音声リモコンのほうがタスクを達成しやすく操作時間も短かった．また，それに対応して「将来的に使いたいか」という評価は音声リモコンのほうが通常のリモコンよりも有意に大きかった．ただし音声リモコンへの評価でさえ中

1 同志社大学大学院
Graduate School, Doshisha University
2 同志社大学
Doshisha University
3 同志社大学
Doshisha University

程度であった．また，音声フィードバックがかえってイライラさせた,音声入力が難しかったという感想が得られた．この結果から，音声機能を搭載したリモコンはある程度便利ではあるが，音声コミュニケーションという点では評価されにくいことが示された．近年では，このような問題点を改善した会話システムが開発されており Apple の Siri や NTT ドコモのしゃべってコンシェルや Yahoo!の音声アシストなどの製品が有名である．

非タスク指向型の会話システム（以後，会話システムとする）に関して，近年多くの研究開発が行われている．初期の研究では Eliza[6]などが有名である．吉田ら[7]は笑芸で用いられるテクニックを利用して，ユーモアを含む会話を行うシステムの提案を行った．その結果，ユーモア機能を搭載したシステムでは対話継続性や娯楽性などが比較的高く評価された．金子ら[8]は話題を理解して自然な応答を行う会話システムの提案を行った．その結果，一部の実験参加者からは人間のように自然な応答を行う場合があると評価された．また，石川ら [9]は感情表現を含む会話システムでユーザを説得するシステムの提案を行った．その結果，実験参加者の一部に感情表現による説得の効果が認められた．このように様々な目的に沿った会話システムの研究開発が行われている．

近年，自然に会話を行うことができる非タスク型会話システムが求められている．多種多様なシステムが開発されているため，その開発目的に沿ったアンケートによってユーザからの評価が測定されており，共通性と客観性に欠けると考えられる．そこでユーザの会話を，会話の快適さや会話の自然さを行動指標（行動を用いた測定尺度）によって数量的に測定することで，ユーザからの評価を客観的なものとすることができる．これによって会話システムのさらなる発展が狙えるものと考えられる．

1.1 既存の評価方法と問題点

多種多様な開発目標のために多くの会話システムが生まれてきたが，その中でもごく近年まで会話システムの 1 つの目標は明らかであったように思われる．それはチューリングテスト[10]に合格することであった．チューリングテストでは複数の審査員によって，コンピュータが思考しているか否かの判定が行われる．合格の 1 つの基準としては，その判定の誤答率が 30%より大きければ，そのシステムは思考しているとみなされ，チューリングテストに合格したものとされる．そしてついに 2014 年に Eugene Goostman という会話システムがチューリングテストに合格した．これで会話システムはその目標を達成したかのように思われたが，多くの問題点が指摘された．

目立つ問題点は次の 2 点である．1 つ目は，システムの設定は 13 歳のウクライナの少年であり,多少のおかしな発言や理解が及ばないことがあった場合に問題だと判断されにくいという点である．2 つ目は，そもそもの会話時間が 5 分間という短い時間であり，会話システムに問題があったにもかかわらず，それが時間内に検出されなかった可能性があるという点である．同様の問題点を指摘する研究者も多く[11]，会話を正しく評価する方法の開発，すなわちチューリングテスト自体の合格基準の改定が望まれている．しかしながら，チューリングテストの通過は会話システムの 1 つの目標ではあるが，そもそも万人が納得するような「思考をしている（人間らしい）と認められる」会話であると正しく評価することができるだろうか．むしろチューリングテストによる会話システムへの共通の評価とは別に,個々のユーザによる共通の評価が必要ではないだろうか．

1.2 評価尺度の提案

個々のユーザから測定すべき評価項目とはどのようなものだろうか．当然，開発目的によって測定すべき評価項目は異なるものである．しかしそのためにユーザからの共通の評価項目が存在せず，ユーザからの受けという点で会話システム間での比較が困難となっていることは事実である．我々が日常的にそうであるように，会話を楽しむ，という点において評価がなされるべきである．このような基準が定まれば，異なる会話システム間での比較が容易となり，会話システムのさらなる競争と発展につながると考えられる．そこで本研究では，ユーザが会話をどの程度楽しんでいるかという点で測定すべき評価とその尺度について提案する．

これまで生物学など他分野の考え方から着想を得て情報処理の分野は発展してきたが，これを会話システムの評価にも導入してはどうだろうか．会話を行動の 1 種類としてとらえれば，行動分析学の理論や手法を応用できるのではないだろうか．行動分析学の理論にのっとれば，ある行動に対して結果がフィードバックされ，フィードバックによってその行動の頻度や強度が増減し，またそのフィードバックを行う対象への好みまで変化する場合がある．この時，頻度や強度が増加すればその行動は行為の主体者にとって好ましい行動である．これを会話に置き換えると，まずユーザの発言に対してシステムの応答がフィードバックとして返される．そのフィードバックの好ましさ，快適さによって会話の内容や継続性が変化する．最終的にはフィードバックの繰り返しによってシステム自体への評価も変化する．またアンケートによる主観的評価による尺度とは異なり，行動は比較的客観的な尺度であるため，自然な評価を取りやすく偽ることが難しい．このような利点から，会話を行う際の尺度として行動による指標を導入すべきである．たとえば会話システムの使用時間やその際の発話回数などは簡単に測定することができる．これらは，実際のユーザのシステムへの好ましさ，会話の快適さを表す指標であるが，ほとんどの研究でこれらの測定や分析は行われていない．

本研究における主な測定対象は①ユーザの快適さ，②シ

ステムとの会話の自然さ，③会話システムの人間らしさ，の 3 つとする．①の測定理由はこれまでに述べたため割愛する．②③は先行研究の多くで使用されている評価であり[12][13]，①との関連が非常に疑われるが，実際に①との関係を調査した研究は筆者の知る限り存在しない．また行動指標によってこれらの測定を試みた研究も同様に存在しない．そこで改めて①②③に関してアンケートと行動指標による測定を行い，それぞれの間で比較を行う．また，会話の相手をヒトとする条件を設け，会話の相手がシステムである場合よりも会話時間は長く，会話間時間は短く，単位時間当たりの発話回数は少なくなると予想する．具体的な測定内容に関してはのちに説明を行う．

2. 関連技術

本研究では北川ら[14]による自己開示を含む雑談型会話システム（以後，本システム）をもとに，上記の指標を用いた評価実験を行う．以下に関連技術と本システムの概要を記す．

2.1 MeCab

MeCab[15]とは，入力された文に対して形態素解析を行うシステムである．形態素解析は文を，言語として意味を持つ最小単位である形態素に分割し，またそれぞれの形態素の品詞を判別するものである．提案システムにおいてシステムからの質問に対するユーザの応答を形態素解析する際に用いる．

2.2 京大格フレーム

京大格フレーム[16]は，Web 上に存在するテキスト 16 億文から構築された大規模データベースである．用言は約 4 万語が登録されており，それぞれの用言に対して共起する名詞を格ごとに整理し，その頻度を取得することができる．本研究において格頻度知識ベースを作成する際に，京大格フレームのデータを用いる．以下の表 1 に，京大格フレームに動詞「泳ぐ」を入力した結果を示す．

本研究で使用するシステムは自己開示応答を特徴とする会話システムである．自己開示応答では，ある語に対して抱く感覚を発話に取り入れる．このような発話が会話中に含まれることによって，話者への質問とシステムからの自己開示の両方が行われる．これによりシステムはより自然な会話を行うことができると考えられる．

本システムは，自己開示応答を含む複数の応答手法によって会話を行う．図 1 は本システムの大まかな全体フローを表す．システムを起動すると入力待機状態となり，話者が文を入力すると，入力文が挨拶か否か確認する．挨拶であった場合は挨拶応答を用いて応答を出力し，最初の入力待機状態に戻る．挨拶でなければ入力文を会話履歴フレームに格納する．会話履歴フレームの状態から，話題転換応答以外の各応答手法で応答が生成できるかどうか確認する．

表 1　「泳ぐ」に対する京大格フレームの出力

名詞	格	頻度
魚	ガ格	1483
プール	デ格	1195
視線	ヲ格	1118

この時，応答可能な応答手法が 1 つもなければ話題転換応答が行われ，会話履歴フレームの中身をリセットした後に，開始時の状態に戻る．応答可能な応答手法があれば，あらかじめ定めた確率に応じて応答手法が選択され，その応答手法での処理が行われる．

2.3 会話履歴フレーム

入力文の情報を 7W1H と述語のフレームに分割して格納するシステムである．7W1H とは英語の疑問視に用いられる 6W1H（いつ，どこで，誰が，何を，誰に，何故，どのように）に「誰と」を表す Who+フレームを追加したものである．表 2 に意味理解システムの実行例として「私は友達と水族館でイルカを観た.」という文を入力した場合の結果を示す．本システムではこのような会話履歴フレームが作成され，その内容に対応した応答を行う．

入力文を各フレームに分類するため，格頻度知識ベースを作成した．格頻度知識ベースは京大格フレームを情報源としており，国語辞書に登録されている全用言 18046 語について共起する名詞と助詞を取得する．取得した名詞のシソーラス[17]における親ノードと助詞によって 7W1H に機械的に分類し，その頻度に応じてまとめたものである．分類時のルールを以下の表 3 に示す．

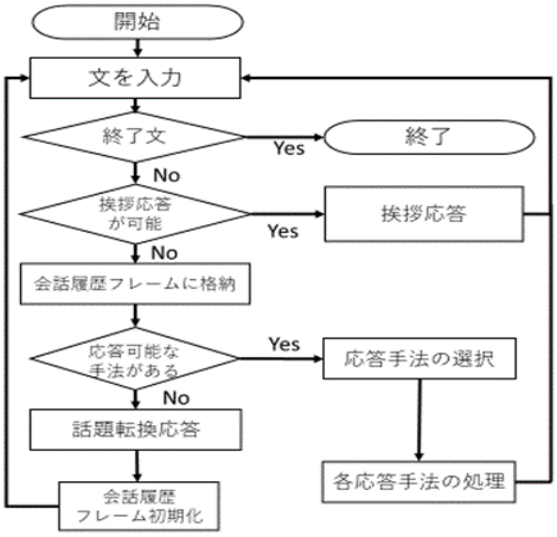

図 1　本会話システムのフロー

表 2　意味理解システムの一例

Who	Who+	What	When	Where	Whom	How	Why	述語
私	友達	イルカ		水族館				観た

表 3　7W1H の分類ルール

助詞	親ノード	分類格
で	乗り物，道具	How
	建造物，施設，場所	Where
	形容詞	Why
と	人物，人名	Who+
を	建造物，施設，場所	Where
	道具	What
に	時間	When
まで	時間	When
から	時間	When
	形容詞	Why
が	人物，人名	Who
	無生物	What
に	人物，人名	Whom
	時間	When
	建造物，施設，場所	Where
へ	人物，人名	Whom
	建造物，施設，場所	Where

表 4　本システムの各応答手法

応答手法	発話の種類	応答文の例
未登録語応答	質問	インスタグラムとは何ですか？
7W1H 応答	質問	どこで食べましたか？
掘り下げ応答	質問	友達の誰さんですか？
場所判断連想応答	質問	海に釣りに行ったのですか？
自己開示応答	自己開示	アイスクリームは甘いところが好きです.
話題転換応答	質問	ところで明日は何をしますか？

本システムではこのように会話履歴フレームを作成し，その内容に対応した応答を行う．表 3 は本システムにおける各応答手法の一覧である．

2.4 各応答手法

2.4.1 未登録語応答

未登録語応答とはシソーラスに登録されていない名詞と定義する．これが入力文に含まれる場合，それについて聞き返す.

2.4.2 7W1H 応答

7W1H 応答は，話題となっている格を特定した時，会話履歴フレームの該当する格に語が格納されていない場合，その格について質問する手法である.

2.4.3 掘り下げ応答

掘り下げ応答は，ある語について詳細を尋ねる応答である．例えば「動物園に行ってきた」に対して「どこの動物園なの？」,「料理を食べた」に対して「どんな料理なの？」等の質問をすることが，人間同士の対話においてみられ，それを再現するのが本応答である.

2.4.4 自己開示応答

自己開示応答は，話者の発話に含まれる語についてシステム側から主観を含んだ表現を行う応答である．感覚判断システムにおける 95 種の感覚語に対して好き嫌いを設定し，これを用いて応答を生成する．好き嫌いの設定の仕方は本研究においては，事前に手動で行うものとする.

自己開示応答で処理が可能な条件は 2 つある．1 つ目は会話履歴フレームに格納されたいずれかの語について感覚判断システムによって感覚語が取得できることである．例として，会話履歴フレームの What 格に「林檎」が格納されていた場合,「林檎」から感覚語｛甘い，赤い，丸い｝が取得できるため，自己開示応答による処理が可能となる．2 つ目は会話履歴フレームの述語が感覚語の場合である．例として話者が「猫は可愛い」という発話を行うと，会話履歴フレームの述語には「可愛い」という感覚語が格納されるため，自己開示発話による処理が可能となる.

3. 実験方法

3.1 実験参加者

理工学部の大学生および大学院生の男女 10 名（男性 8 名，女性 2 名）であり，平均年齢は 22.3 歳であった.

3.2 実験の流れ

実験参加者にキーボードを使用したチャット形式の会話を数分間行わせた．会話の相手をシステムとするシステム条件と会話の相手を実験者とするヒト条件の両方で会話を行った.

実験参加者に次のように教示した.『これからこちらの会話システムを数分間使用していただきます．会話が終了した後はこちらの会話システムに関する質問に答えていただきます』と教示した後，会話システムの操作方法について説明を行った．説明を終えた後『これからこちらのシステムと自由に会話を行ってください．会話における注意が 2 点ありますのでよく聞いてください. 1 点目ですが，普段あなたが公共の場で話すときのように会話を行ってください．もう 1 つは，これから 3 分間から 5 分間の会話を行っていただきます．3 分が経過した際にお伝えしますので，その後終わりたいなと思えばいつでも会話を終えてもらっても構いません．3 分が経過した後は私実験者からは時間の経過や実験の終了などお伝えしませんので，最大でも合計 5 分間，つまり 3 分経過と伝えた後 2 分が経過したなと思った場合は「さよなら」と入力して会話を終了してください．なお，3 分が経過するまでにどうしても会話をやめ

たくなった場合も「さよなら」と入力して終了してもらっても構いません．終わる際は会話の内容にかかわらず「さよなら」と入力し，実験者に終了の旨を伝えてください．練習として「さよなら」と入力してみてください』と教示を行った．実験参加者の入力後，不明な点などの質問を受け付ける．その後,『それでは準備がよろしければ「スタート」と入力して会話を始めてください』と教示を行った．会話が終了した後席を移動させてアンケートに回答させた．回答終了後，残りの条件で同様に会話を行わせ，また同様にアンケートに回答させた．その後，実験を終了した．

3.3 ユーザの快適さ

ユーザの快適さを測定する行動指標は，実際に会話を行った総時間と発話回数と発話の長さとした．2 分が経過すれば会話を終了してよいため，会話時間が必要以上に長ければ自発的に会話を行っていることとなり，その場合は会話を楽しんでいると考えられる．また，会話がはずんだ場合，会話の頻度や長さが増加すると考えられる．アンケートには「この会話システムを使用して楽しかった」「この会話システムを使用して満足している」「この会話システムをまた使用したい」の 3 項目を用いて,「1.全くそう思わない」から「5.非常にそう思う」までの 5 件法による評価を行わせる．質問項目は Grice の会話の公準[18]や，下岡ら[12]の質問項目を参考し作成した．

3.4 会話の自然さ

会話の自然さを測定する行動指標は，システムの発話からユーザが次に発話するまでの時間の長さとした．もしシステムの発話が自然ならば，ユーザは普段話すようにスムーズに会話を進めることができる．なお会話の内容自体が難解な場合にもユーザは発話に時間を要すると考えられるので,「会話の内容は難しかった」という質問項目を設けた．アンケートには「システムが提供する話題は適切だった」「システムとの会話はスムーズだった」「困ることなく普段話すように話せた」に関して上記と同様に答えさせた．これらの項目は[12]を参考に作成した．

3.5 会話の人間らしさ

人間らしさを測定する行動指標は，単位時間（30s）当たりのユーザの発話数とした．行動分析では消去バースト[19]と呼ばれる現象が確認されている．これは行動後に得られていたフィードバックが得られなくなるとその行動を急激に何度も繰り返し，その後次第に落ち着くという現象である．本研究では，普段と異なる会話になった際に消去バーストが生じるのではないかと考えた．つまり，普段から日常会話を行うものほどおかしな会話になった場合，短時間に発話を多く行い，しばらく会話を続けると発話頻度は元に戻るということになる．アンケートでは「このシステムには感情がある」「このシステムには意識がある」「このシステムには性格（個性）がある」に関して上記と同様に答えさせた．また，普段の会話量を測定するため「あなたは普段 line やメールのようなチャット式のコミュニケーションツールをどれぐらい使用しますか」「あなたは普段どれぐらい会話をしますか」に対して「1.ほとんど使用（会話）しない」から「5.頻繁に使用（会話）する」の 5 件法で測定した．

4. 実験結果

本研究では①ユーザの快適さ，②システムとの会話の自然さ，③会話システムの人間らしさに関して，会話の相手がシステムであるシステム条件と，会話の相手が人間であるヒト条件のそれぞれで測定を行った．

図 2 はシステム条件とヒト条件におけるシステム使用時間の平均を示す．両条件において，会話の終了条件とした会話時間（基準時間）の長さとした 300 秒よりも会話時間の平均は大きかった．そこでシステム条件で 1 群の t 検定を行ったが基準時間との間に有意差は見られなかった（両側検定，t （9） = 1.797, p = .106）．同様にヒト条件で 1 群の t 検定を行った結果，基準時間よりも有意にヒト条件での会話時間は大きかった（両側検定，t （9） = 3.412, p = .008）．両条件間で Wilcoxon の順位和検定を行った結果，有意差は見られなかった（T =18, r = .109, p = .631）．つまり，本システムとの会話よりはヒトとの会話のほうが比較的快適であることが示唆された．

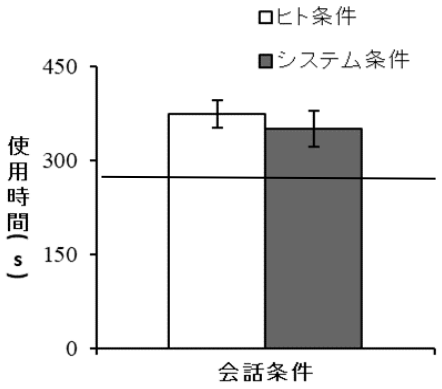

図 2　条件ごとの会話の長さ（エラーバーは標準誤差）

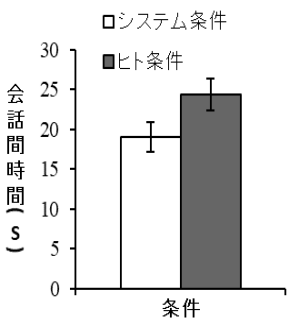

図 3　条件ごとの会話間時間（エラーバーは標準誤差）

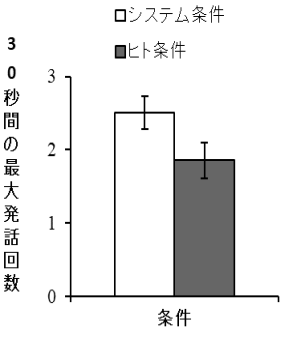

図 4　条件ごとの単位時間当たりの発話数（エラーバーは標準誤差）

図 3 は両条件における会話間時間の平均を示す．会話間時間はシステムの発話後にユーザが発話するまでの時間とした．システム条件の平均会話間時間はヒト条件のそれよりも大きかった. そこで両条件間で Wilcoxon の順位和検定を行った．その結果システム条件の平均会話間時間はヒト条件よりも有意に小さかった（$T = 5$, $r = -.501$, $p = .025$）.

図 4 は両条件における任意の単位時間におけるユーザの発話数を示す. 発話回数がもっとも多かった 30 秒間を任意の単位時間とした．システム条件における発話数は，両条件の発話間時間の長さに応じて調整を行った．システム条件の単位時間あたりの発話数平均はヒト条件のそれよりも大きかった. そこで両条件間で Wilcoxon の順位和検定を行った．その結果，システム条件の発話数平均はヒト条件よりも有意に大きかった（$T = 3$, $r = .547$, $p = .014$）.

図 5 では a .平均会話時間と平均会話間時間の散布を，図 6 では b.平均会話間時間と単位時間当たりの発話数の散布を，図 7 では c.平均会話時間と単位時間当たりの発話数の散布を，それぞれヒト条件で示した.

図 5 に関して平均会話時間の相関係数は $r = .339$ であり弱い正の相関がみられたが，有意な値ではなかった（$p = .338$）. 図 6 に関して平均会話間時間の相関係数は $r = .-297$ であり弱い負の相関がみられたが，有意な値ではなかった（$p = 405$）. 図 7 に関して平均会話時間の相関係数は $r = -.034$ で相関は見られず, 有意な値ではなかった（ $p = .926$）.

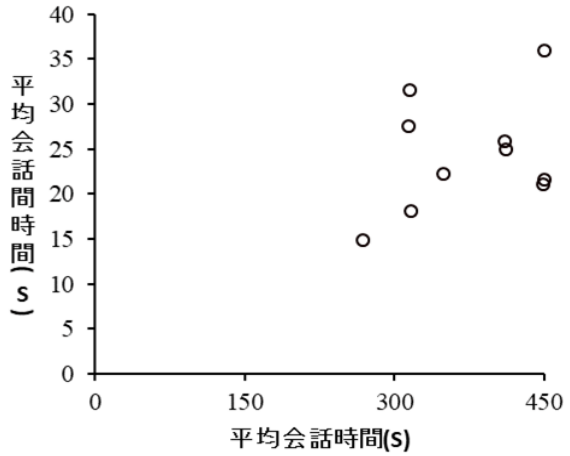

図 5　a. 個人別の平均会話時間と平均会話間時間の散布

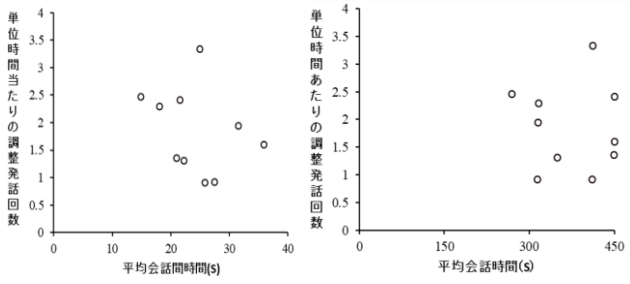

図 6　b. 個人別の平均会話間時間と発話数の散布（左図）
図 7　c. 個人別の平均会話時間と発話数の散布（右図）

表 5　質問項目および測定概念ごとの平均

		システム条件	ヒト条件	測定概念
1	システムが提供する話題は適切だった	2.90	4.50	自然さ
2	システムとの会話はスムーズだった	2.10	4.70	自然さ
3	この会話システムをまた使用したい	2.50	3.90	快適さ
4	このシステムには感情がある	1.50	3.00	人間らしさ
5	この会話システムを使用して楽しかった	3.10	4.10	快適さ
6	会話の内容は難しかった	2.30	1.30	内容確認
7	このシステムには意識がある	2.10	3.50	人間らしさ
8	この会話システムを使用して満足している	2.50	4.00	快適さ
9	困ることなく普段話すように話せた	1.80	4.70	自然さ
10	このシステムには性格(個性)がある	2.00	2.60	人間らしさ
①	ユーザの快適さ	2.70	4.00	快適さ
②	会話の自然さ	2.27	4.63	自然さ
③	会話の人間らしさ	1.87	3.03	人間らしさ

表 5 ではアンケートにおける質問項目ごとの平均（表中の 1-10）と測定概念ごとの平均（表中の①②③）を条件ごとに示した．測定概念はそれぞれの項目がどの概念と対応するかを示した．①ユーザの快適さに関して両条件間で Wilcoxon の順位和検定を行った．その結果，ヒト条件のユーザの快適さはシステム条件よりも有意に大きかった（$T = 7.5$, $r = .444$, $p = .047$）. ②会話の自然さに関して両条件間で Wilcoxon の順位和検定を行った．その結果，ヒト条件の会話の自然さはシステム条件よりも有意に大きかった（$T = 0$, $r = .615$, $p = .006$）. ③会話の人間らしさに関して両条件間で Wilcoxon の順位和検定を行った．その結果，ヒト条件の会話の人間らしさはシステム条件よりも有意に大きかった（$T = 0$, $r = .583$, $p = .009$）.

5. 考察

本研究では①ユーザの快適さ，②システムとの会話の自然さ，③会話システムの人間らしさに関して，会話の相手がヒトもしくはシステムという条件でそれぞれ測定を行った.

①の行動指標には会話の長さを使用した．実験参加者は基準時間（300s）が経過した場合に会話をやめるよう指示されていたが，ヒト条件でのみ基準時間を超過して会話を継続した．これは会話が快適であったため自発的に会話を継続した，もしくは快適な時間を過ごしたため体感時間よりも実際の時間は経過していた，と考えられる．また，アンケートでもヒト条件のほうがシステム条件よりも快適であることが示された．これらから会話時間の長さは「ユーザの快適さ」の行動指標として妥当であることが示唆された.

②の行動指標には会話間時間を使用した．本研究ではヒト条件のほうで会話間時間が小さい，つまりスムーズに会話を行うと予想した．しかし，システム条件のほうで会話間時間が小さかったため，この結果からはスムーズな会話

を行うことができたと示唆される．ただしアンケートではヒト条件のほうで会話が自然だったと評価された．また会話間時間が大きいほど会話時間も大きい傾向が図 5 から認められる．これらから，自然な会話ほど考える時間が増加するため会話間時間がある程度長くなる，ということが考えられる．

③の行動指標には単位時間当たりの発話回数を使用した．会話の相手がシステムである場合のほうが単位時間当たりの発話回数が大きかった．普段とは異なる会話，つまり人間らしくない会話に対して消去バーストが発生し，局所的に発話頻度が増加したと考えられる．アンケートでもヒト条件のほうで会話が人間らしいと評価されており行動指標の結果と一貫している．

しかしながら，アンケートではヒト条件での人間らしさは 3.03 であり，システム条件と比べると比較的大きいが，会話の相手がヒトであるにもかかわらず高い評価とは認められない．これに加え，消去バーストは普段の会話と異なる場合に生じるため，会話が人間らしくない場合に限らず会話が不自然である場合にも生じるものである．ヒト条件で会話の自然さがアンケートで高く評価されていたことが正しいとすれば，単位時間当たりの発話回数は会話の自然さを測定していたと考えられる．

6. 結論

6.1 本研究の限界と展望

本研究の問題として，会話の人間らしさを測定する行動指標をほとんど提供できなかったことが挙げられる．これは示唆に富むものでもあるが，アンケートにおいて会話の相手がヒトであるにも関わらず人間らしさが評価されなかった．質問項目が不十分であったとも考えられるが，そもそも万人の考える人間らしさがあるのか，またそれを妥当な尺度で測定できるのか検討すべきである．

本研究での会話時間は 5 分間であり各条件 1 回のみであった．そのため，実際のユーザは会話システムを繰り返し使うことになるが長期的な快適さや会話の自然さは別の議論が必要であろう．またその意味では，我々が普段長い時間を通じて人間関係を構築しお互いを知ることを考慮すると，そもそも人間らしさは長期的な尺度によってのみ測定できる可能性には留意したい．

6.2 まとめ

本研究では会話システムの共通評価尺度の作成のため，①ユーザの快適さ，②会話の自然さ，③会話の人間らしさに関して行動指標を提案した．その結果，①ユーザの快適さに関しては「会話時間の長さ」をおおむね妥当な行動指標として提案できた．②会話の自然さに関しては当初の予想とは異なるが，「会話間時間の程度」もしくは「単位時間当たりの発話回数の大きさ」が行動指標となりうる可能性を提供できた．③会話の人間らしさに関しては行動指標とアンケートの両方で問題が生じたため，長期的な視点も考慮しつつ，今後の検討が必要である．

謝辞 本研究に協力頂いた皆様に，謹んで感謝の意を表する．

参考文献

[1] 小松原剛志，塩見昌裕，神田崇行，石黒浩，萩田紀博．理科室で授業の理解を支援するロボットシステム．日本ロボット学会誌, 2015, Vol. 33, no.10, p. 789-799.
[2] 徳永礼．ロボット家電との音声会話がユーザに与える効果．エンタテインメントコンピューティングシンポジウム論文集, 2015, p. 179-185.
[3] 小林峻也，萩原将文．ユーザの嗜好や人間関係を考慮する非タスク指向型対話システム．人工知能学会論文誌, 2016, DSF-502.
[4] 宅和晃志，吉川大弘，古橋武．非タスク指向型対話システムにおけるあるあるツイートからの共感誘発型発話生成手法に関する検討．知能と情報, 2018, vol. 30, no. 5, p. 744-752.
[5] 吉田諒，安村通晃．音声とテンキーを統合した視覚障害者向け携帯電話型家電リモコンの試作と評価．情報処理学会研究報告音声言語情報処理, 2007,no.7, p. 17-22.
[6] Weizenbaum, J.. ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 1966, vol. 9, no. 1, p. 36-45.
[7] 吉田裕介，萩原将文．複数の言語資源を用いたユーモアを含む対話システム．知能と情報, 2014, vol. 26, no. 2, p. 627-636.
[8] 金子稜，吉村枝里子，土屋誠司，渡部広一．話題を考慮した自然な会話システムの構築．研究報告知能システム , 2016, no. 2, p. 1-8.
[9] 石川葉子，水上雅博，吉野幸一郎，鈴木優，中村哲．感情表現を用いた説得対話システム．人工知能学会論文誌, vol.33,no. 1, DSH-B_1.
[10] Turing, I. B. A. . Computing machinery and intelligence-AM Turing. Mind, 1950, vol. 59, no. 236, 433p.
[11] 東中竜一郎．チューリングテスト「合格」のシステム．情報処理, 2014, vol. 55, no.9, p. 904-907.
[12] 下岡和也，徳久良子，吉村貴克，星野博之，渡部生聖．音声対話ロボットのための傾聴システムの開発．自然言語処理, 2017, vol.24, no.1, p. 3-47.
[13] 畑健治，小倉卓也，萩原将文．言語資源を用いた非タスク指向型対話システム．日本感性工学会論文誌, 2011, vol. 10, no. 4, p. 515-522.
[14] 北川智裕，土屋誠司，渡部広一. 2018. 自己開示発話を取り入れた雑談型対話システムの提案. SIG-KBS, vol.5, no. 03, p. 13-18.
[15] https://taku910.github.io/mecab/
[16] 河原大輔，黒橋禎夫．高性能計算環境を用いた Web からの大規模格フレーム構築，情報処理学会 自然言語処理研究会, 2006, vol. 171, no.12，pp.67-73.
[17] NTT コミュニケーション科学研究所，日本語語彙体系，岩波書店，1997.
[18] Grice, H. P. 1989. Studies in the Way of Words. Harvard University Press.
[19] Lerman, Dorothea C.; Iwata, Brian A. Prevalence of the extinction burst and its attenuation during treatment. Journal of applied behavior analysis, 1995, vol. 28, no.1, p. 93-94.