

テキスト対話エージェントからの返信間隔の違いは ユーザの印象評価に影響を与えるのか

How do different switching pause durations between users and text-dialogue agent affect the users' impressions of this agent?

須田翔悟¹ 神保一馬¹ 小松孝徳¹ 山田誠二^{2,3}

Shogo Suda¹, Kazuma Jimbo¹, Takanori Komatsu¹ and Seiji Yamada^{2,3}

¹明治大学総合数理学部, ²国立情報学研究所, ³総合研究大学院大学

¹Meiji University, ²National Institute of Informatics, ³SOKENDAI

昨今の急速な AI 技術の発展により、ユーザとチャットを行うテキスト対話エージェントが普及しつつある。その一例であるチャットボットは、ユーザの入力に対して即時的な返信を行っているが、人間同士がチャットなどで雑談をする際には必ずしも即時的な返信を行っているわけではない。そこで本研究では、ユーザと雑談を行う目的のテキスト対話エージェントは即時的な返信よりも間隔を開けた返信の方がユーザに好印象を与えると考え、それを検証するための実験を行った。

1. はじめに

昨今の急速な AI 技術の発展により、ユーザとテキスト対話を行うテキスト対話エージェントが普及しつつあり、それらはチャットボットとして様々なシステムに実装されている。例えばヤマト運輸は、ユーザから送られるテキストメッセージによって配達時間や配達方法を指定することができる LINE bot を開発した[1]。また任天堂ではユーザからの簡単な問い合わせに対応するヘルプデスクとしてチャットボットを活用している[2]。チャットボットの多くは、上記のように問い合わせ対応などの具体的な業務を担当している一方、人間と自然な雑談を行うことを目的としたテキスト対話エージェントも開発されつつある。

その一例として挙げられるのが、Microsoft 社によって開発された AI コミュニケーション bot の「りんな」である[3]。「りんな」はディープラーニングを用いた自然言語処理技術により、ユーザと人間のような文脈を踏まえた自然な会話を行うことができる。しかしながら、「りんな」との会話においてユーザが「死ぬ」という単語を多用するなど、必ずしも「りんな」はユーザに受け入れられていなかったと Microsoft の研究チームは報告している。そこで著者らは、「りんな」がユーザから受けられなかった原因の一つとして、ユーザからのメッセージを受け取ると即座に返信を行うという即時応答性

にあると考えた。人間同士のテキストチャットでは、メッセージの送信後、その返信までの間隔は数秒間空くことは普通であり、返信が即座に返ってくることはほとんどない。そのため、「りんな」の即座の返信がユーザに対して違和感や圧迫感を与えていたと考えられる。

そこで本研究では、ユーザと雑談をする目的であるテキスト対話エージェントは、ユーザからのメッセージに対してどのくらいの間隔を開けて返信するべきなのかを実験的に調査することとした。具体的には、ユーザによるメッセージの送信からエージェントによる返信までの間隔（交替潜時）を「独立変数」、その際のユーザのエージェントへの印象を「従属変数」とした実験を実施した。

2. 関連研究

コンピュータシステムにおいて、ユーザからのアクションに対する応答時間は、満足度を向上させるための重要な要素であるため、コンピュータシステムにおける遅延に関する研究はこれまで盛んに行われてきた[4-8]。例えば、Hoxmeir ら[9]はシステムの応答時間の遅延とユーザの満足度の関係について調査し、システムの応答時間に遅延がない場合に最も満足度が高く、遅延が 3 秒~9 秒の場合は遅延がない場合に比べて満足度が約 3 分の 2 以下に減少し、遅延時間が 12 秒になるとそれが約 3 分の 1 に減少することを報告した。

しかし先の「りんな」の例のように、エージェントの「早すぎる応答」はユーザに機械的な印象を与えてしまうため、必ずしも応答が早ければ良いというわけではないことも報告されている [10-12]. Moon [13]は、参加者に返信間隔が操作されたテキスト対話エージェントと共に「砂漠の生存問題」に取り組んでもらう実験を実施した。「砂漠の生存問題」とは、砂漠で生き延びるために必要な道具 12 個の中から、どの道具を使用したいかのランク付けを行うというものである。その結果、短い返信間隔 (0~1 秒) や長い返信間隔 (13~18 秒) よりも、適度な返信間隔 (5~10 秒) のエージェントの方が、ユーザにとって説得力があると認識されていたことが明らかとなった。

これらの先行研究においては、参加者は何かしら与えられた目的を達成するためにエージェントと会話を行うという実験環境が設定されている。一方で、本研究のようにテキスト対話エージェントと雑談を行うことを目的とした実験環境において、その返信間隔の影響を考察した研究は、管見の限り行われていないのが現状である。また返信間隔に関しても、先行研究では長くとも数十秒程度までしか設定されておらず、本研究で行うような 1 分以上の長い返信間隔の影響は考察されていない。

3. 予備調査

テキスト対話システムとユーザとの自然な返信間隔を考察するため、まずは最も自然なテキストコミュニケーションとみなすことができる「人間同士のテキスト対話」において、返信間隔がどのように認識されているかを把握する予備的な調査を実施した。本調査にはクラウドソーシングによって募集された 246 人 (18 歳~81 歳, 平均年齢 41.6 歳; 男性 165 人, 女性 80 人, 無回答 1 人) が参加した。調査における質問項目を表 1 に示す。

その結果、54.0%の回答者が SNS でのメッセージに気づいてから一時間以上の間隔を空けて返信すると回答していた。また、メッセージアプリの LINE でも同様に、一時間以上後に返信すると 40.7%が回答していた。特に 18 歳から 24 歳の若者層では、SNS のメッセージに気づいてから一日以上空けて返信すると 47%が回答し、LINE でも 44.4%が一日以上空けると回答するなど、返信間隔が極端に長い傾向にあることが明らかとなった。

また、メッセージには返事が急がれているものとしてでないものがある。返事が急がれている SNS でのメッセージに対しては 62.1%が 10 分以内に返信すると回答し、LINE でも 71.0%が 10 分以内に返

表 1 : 予備調査の質問項目

	質問
a-1	InstagramやTwitterなどのSNSのメッセージ機能において、普段どれくらい未読メッセージをためていますか
a-2	InstagramやTwitterなどのSNSのメッセージ機能において、メッセージに気づいてから、平均的にどれくらいの間隔を空けて返信しますか
a-3	InstagramやTwitterなどのSNSメッセージ機能において、返信が急がれているメッセージに対する返信時間はメッセージに気づいてからどのくらいですか
a-4	InstagramやTwitterなどのSNSメッセージ機能において、返信が急がれていないメッセージに対する返信時間はメッセージに気づいてからどのくらいですか
a-5	LINEで平均的にメッセージに気づいてからどのくらいの間隔を空けて返信しますか
a-6	LINEの未読メッセージの数はどのくらいですか
a-7	LINEを1日に何回開きますか
a-8	LINEで返事を急がれているメッセージに対して、メッセージに気づいてからどれくらいの間を空けて返信しますか
a-9	LINEで急がれていないメッセージに対して、メッセージに気づいてからどのくらい時間を空けて返信しますか

信すると回答していた。一方、返事を急がれていない SNS でのメッセージに対しては 45.1%が一時間以上の間隔を空けると回答し、LINE でも 39.0%が一時間以上の間隔を空けると回答していた。

このことからテキスト対話における会話の形態は、大きく二種類に分類できると考えられる。一つは、相手が早急な返信を求めているものである。こういった会話ではお互いが相手のメッセージを受け取ると即座に返信を行うので、会話のテンポは速くなる。また、目的が達成されれば会話は終了することが多い。このような会話を本稿では「同期的会話」と呼ぶ。一方、雑談のように会話の内容が早急な返信を求めない場合、返信間隔は長くなる傾向にあるといえる。こういった会話ではお互いが都合のいいタイミングで返信を行うため、会話のテンポは遅くなり、長期間にわたって会話が続くことが多い。このような会話を本稿では「非同期的会話」と呼ぶ。

本稿で着目するテキスト対話エージェントとユーザとの雑談という状況は非同期的会話とみなすことができるため、テキスト対話エージェントがユーザと雑談を行う場合、返信間隔を長くした方がユーザからの印象は向上すると著者らは考えた。特に、人間同士のテキスト対話においては一時間以上もしくは一日以上空けてから返信するケースが多くみられていたことから、これらの返信間隔を含んだ実験設定が必要だと考えられた。



図1: 「えだまめ」と参加者の会話

4. 実験

4.1. テキスト対話エージェント「えだまめ」

本実験にて参加者とテキスト対話を行うエージェントとして、LINE bot「えだまめ」を作成した。「えだまめ」はLINE Developer および Google App Scriptにて実装されたbotであり、LINE上での公式アカウントとして運用した。

「えだまめ」は実験参加者からメッセージを受け取ると、実験条件に対応した返信間隔にてメッセージを返信する。なお、「えだまめ」は参加者のメッセージの内容に関わらず、決まった文章を返信することとした(図1)。これは、参加者の送信内容によって「えだまめ」と参加者との対話内容が変化することを防ぎ、対話の内容ではなく返信間隔の違いのみが参加者の心象に与える影響について考察するためである。よって、「えだまめ」の返信は参加者の送信内容に関わらず、できるだけ自然な会話になるように配慮して設計された。具体的には、参加者の返信に対して何らかの反応を示したのち、何らかの質問を返すという形式で会話を成立させるように設計した。

表2: 実験後アンケートの質問項目

	質問
b-1	「えだまめ」は親しみやすいと感じた
b-2	「えだまめ」と仲良くなれたと感じた
b-3	「えだまめ」には人間味があると感じた
b-4	「えだまめ」は好印象だった
b-5	「えだまめ」との会話に圧迫感を感じた
b-6	「えだまめ」は話しかけやすいと感じた
b-7	「えだまめ」との会話は返信しやすいと感じた
b-8	「えだまめ」との会話は盛り上がったと感じた
b-9	「えだまめ」と話すのは楽しいと感じた
b-10	「えだまめ」との会話は自然だと感じた
b-11	「えだまめ」に愛嬌を感じた

4.2. 実験設定

本実験では、ユーザのメッセージ送信から「えだまめ」の返信までの間隔を独立変数として設定した。独立変数の具体的な水準として、以下の六水準が設定された: 0秒, 5秒, 1分, 10分, 1時間, 24時間。1時間, 24時間という長い返信間隔は、本稿における予備調査にて、人間同士の非同期的会話において一般的な返答時間であったこと、0秒および5秒という短い返信間隔は、テキスト対話における有効な返信間隔として先行研究にて報告されていること、1分, 10分という中程度の間隔は、長い返信間隔と短い返信間隔の間を補完する条件となることから採用された。

また、実際に対話を行った「えだまめ」に対するアンケート調査の結果を従属変数として設定した(表2)。本アンケートは、b-1からb-10までの5段階の10問のリッカート尺度(最低評価が1, 最高評価が5: b-5は逆転項目)およびb-11「そう思った/そう思わなかった」の2択の質問から構成されている。本実験ではこの11問の質問項目を、「えだまめ」というエージェントの人格に対する評価である「人格的印象点」と、「えだまめ」との会話自体に対する評価である「機能的印象点」の二種類に分類して扱うこととした。具体的には、b-1からb-4までの四問の合計点を「人格的印象点」(最低点4点, 最高点20点, クロンバックの α 係数: 0.83), b-5からb-10までの六問の合計点を「機能的印象点」(最低点6点, 最高点30点, クロンバックの α 係数: 0.64)とした。また、表2の質問のほかにも、「えだまめ」に対する印象を自由に記入してもらう自由記述欄にも回答を求めた。

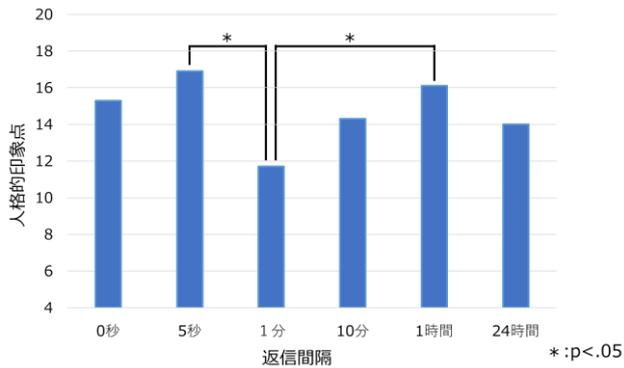


図2: 各水準における人格的印象点

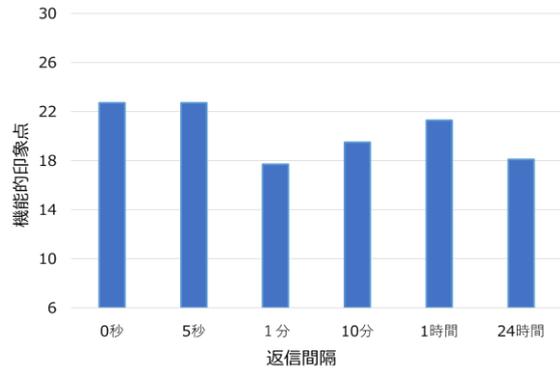


図3: 各水準における機能的印象点

4.3. 実験参加者

本実験には、60名の大学生および大学院生が参加した（18～23歳，平均年齢20.7歳；男性32名，女性28名）。これらの参加者は講義の課題の一環として本実験に参加したため，報酬は支払われなかった。また，これら60名の参加者は，無作為に独立変数の六水準に10人ずつ配置される，参加者間計画として実施された。なお，参加者にはテキスト対話エージェントの返信間隔については何も情報を与えなかった。

4.4. 実験手順

参加者はまず，配置された実験条件に応じた返信間隔を表出する「えだまめ」をLINEで友達に追加するように指示された。「えだまめ」を友達に追加すると，「えだまめ」から参加者に話しかけることで，「えだまめ」との対話が始まる。そして会話終了時に「えだまめ」から示されるURLにアクセスし，Google Formに実装された実験後アンケートに回答することで実験が終了した。

また，「えだまめ」との会話の途中で7日間以上返信をしなかった参加者は会話を放棄したとみなし，会話が全て終了していなくても実験後アンケートに回答してもらい，同時に会話を放棄した理由を記載するように指示された。

4.5. 実験結果

人格的印象点に関して，一要因参加者間分散分析（独立変数：返信間隔（六水準），従属変数：人格的印象点）を行ったところ，群の効果が有意であった [F(5,59) = 3.46, p < .05] (図2)。HSD法による多重比較の結果，1分水準よりも5秒水準および1時間水準の方が，人格的印象点が有意に高いことが明らかとなった (MSe = 9.7056, p < .05)。

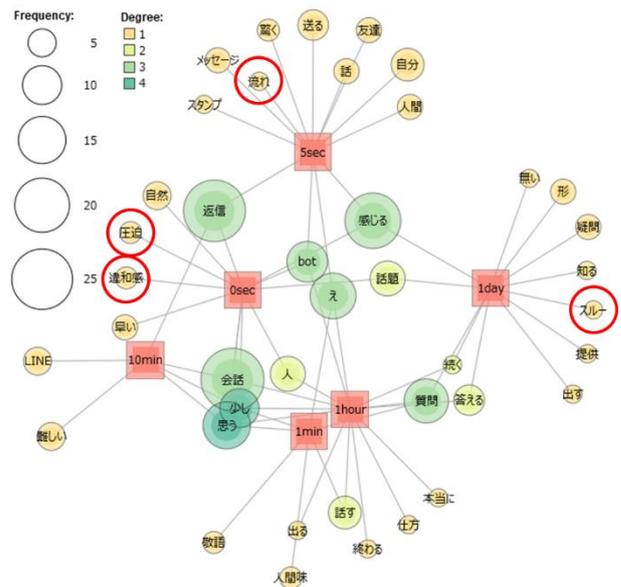


図4: 自由記述の分析結果

続いて，機能的印象点に関して，一要因参加者間分散分析（独立変数：返信間隔（六水準），従属変数：機能的印象点）を行ったところ，群の効果が有意傾向が観察されたものの [F(5.59) = 2.27, p < .10] (図3)，HSD法による多重比較の結果，どの水準間においても有意差は観察されなかった (MSe = 21.7741, p < .05)。

実験後アンケートにおける自由記述を分析した結果，0秒水準において「圧迫」や「違和感」，5秒水準において「流れ」，24時間水準において「スルー」という単語が特徴的に観察された。これらの関係性は，KHcoder[14]による共起ネットワーク図においても確認することができた (図4)。

なお，「えだまめ」との会話を放棄した参加者数は0秒水準で1人，10分水準で2人，1時間水準で2人，24時間水準で1人であったが，5秒水準および1分水準では放棄者は確認されなかった。

5. 考察

5.1. 各水準における人格的印象点

本実験の結果より、「えだまめ」というエージェントの人格に対する評価である人格的印象点は、5秒水準において最も高い得点を示していることが分かった。Moon [13]は、同期的会話においては短い返信間隔(0~1秒)や長い返信間隔(13~18秒)よりも、適度な応答時間(5~10秒)の方が、ユーザからの評価が高いことを報告していた。このMoon [13]の研究と本実験とでは会話の形態が同期的か非同期的かという違いがあるが、その結果はほぼ同一であったといえる。よって、会話の形態によらず、その返信間隔が5秒の場合に、テキスト対話エージェントの評価が高くなることが明らかとなった。

これまでの先行研究では本実験で着目したような1分以上の返信間隔は実験条件として使用されていなかったため、例えばMoonの結果を踏まえると、返信間隔が長くなるほどエージェントの評価は減少し続けると考えられていたといえる。しかし本実験の結果からは、1分水準においてエージェントの印象点は最も低くなるものの、1時間水準になると1分水準に比べて有意に印象点が高くなっていったことが明らかとなった。さらにその得点は印象点が最も高かった5秒水準と同じような値であったといえる。この結果は、著者らの予備調査において、人間同士のテキスト対話における一般的な返信間隔が1時間や24時間であったことと一致しており、非同期的会話においては長い返信間隔の方がエージェントへの印象が良くなるという著者らの仮説がある程度支持されたといえる。しかしながら、予備調査で一般的な返信間隔であった24時間水準では、1時間水準と比べて人格的印象点が低くなっていったことも確認された。これは、本実験における参加者とエージェントが親密でなかったためだと考えられるため、この要因に関する追加実験の必要性があるといえる(例、対話相手が親密なエージェントの場合/初対面のエージェントの場合)。

今回の実験から雑談のような非同期的会話では返信間隔が1時間水準であっても、エージェントに対する評価は5秒水準の場合と同程度であることが明らかになった。返信間隔が短いと会話はすぐに終わってしまうが、返信間隔が長いと会話のテンポは遅くなり、会話が長期間にわたって続くことが多い。つまり、雑談を行うテキスト対話エージェントは、返信間隔を1時間程度に設定することでユーザから高い評価を得ながら、長期的な関係を維持すること

ができるといえる。この知見は、HAI分野の目標の1つである「ユーザと長く付き合っていけるパートナーとしてのエージェントの開発」[15]の達成に有用であると考えられる。

5.2. 各水準における機能的印象点

会話のシナリオによって生じたエージェントへの評価である機能的評価点は、5秒水準における評価が0秒水準と同程度であるという点を除いて、人格的印象点と似たような傾向を示していたといえる。しかし分散分析において有意差は観察されず、さらには機能的印象点におけるクロンバックの α 係数も低い値を示していた(0.64)。特にクロンバックの α 係数が低い値を示していることは、どの水準でも参加者が同じような回答をしていたことを示唆しており、各水準間における機能的印象点に有意な差がないことと合致している。よって、「えだまめ」の返信間隔の違いは、「えだまめ」との会話自体に対する印象には影響を与えていなかったことが明らかとなった。

5.3. 各水準における自由記述欄の傾向

次に、参加者の自由記述欄から各水準についての特徴的な記述に着目する。

0秒水準では、「圧迫」や「違和感」といった単語が特徴的に観察された。これらは、「返信が早く圧迫感があった。」「即座に返信が来るので違和感があった。」というようなネガティブな文脈で使われていた。このように0秒水準ではネガティブな単語が自由記述欄で使われていたにも関わらず、人格的印象点は比較的高い値を示していた。これは、返信間隔が0秒の場合、参加者とエージェントが即座に返信しうするため、会話のテンポが速く、違和感や圧迫感があってもそれが顕在化する前に会話自体が終了していたことが原因だと考えられた。

5秒水準では、「流れ」という単語が特徴的に使用されていた。具体的には、「会話の流れが自然だった。」というような使われ方をしていた。エージェントからの返信内容は共通であるのも関わらず、5秒においてのみ「流れ」という単語が観察されたため、5秒程度の適度に短い返信間隔においては、ユーザはエージェントとの会話を自然なものと感じやすくなり、その結果として人格的印象点が高くなったと考えられた。

24時間水準においては、「スルー」という単語が特

微的に使用されていた。よって、返信間隔が極端に長くなると、ユーザは会話のかみ合わなさや不自然さに気がつきやすくなり、その結果としてエージェントに対する評価が低下したと考えられた。これは、0 秒水準においてエージェントに対して違和感があったにも関わらず評価が高かったことと関連していると推測された。つまり、短い返信間隔ではユーザも即座に返信をするため、会話の些細な違和感や不自然さを気にする間もなく会話が終了してしまうが、逆に長い返信間隔では会話が持続し続けているために、それらの欠点に意識が向いてしまい、結果として低評価に結びついていたと考えられた。

5. 4. 本実験の限界と今後の方向性

本実験では「えだまめ」からの返信内容はあらかじめ決められており、「えだまめ」は参加者からの質問に返答する機能は有していなかった。そのため、一部の参加者は、自由記述にて「えだまめに質問を無視された。」と回答していた。今後はユーザからのメッセージに対して、そのメッセージに応じた返答を表出できるテキスト対話エージェントにおいても、本研究と同様の結果が得られるかを調査していきたいと考えている。

なお、本研究の結果を精査したところ、参加者の性別の違いが人格的印象点に異なる影響を与えていたことが示唆された。具体的には、男性に比べ女性の方がエージェントに対する評価が低い傾向にあったことが確認された。ただし、実験参加者の男女の数に偏りがある状態での比較のため、この段階ではこれ以上この結果について言及することはできない。将来的には実験参加者数を増やすと同時に、実験参加者の性別などを統制した環境にて実験を行っていきたいと考えている。

6. おわりに

本研究では、テキスト対話エージェントがユーザと雑談を行う場合、返信間隔を長くした方がユーザからの印象は向上するという仮説を立て、それを検証するための実験を行った。その結果、返信間隔が5秒もしくは1時間の場合に、エージェントへの印象が有意に高くなることが明らかとなった。従来の研究では1分以上の返信間隔での実験は行われておらず、エージェントの評価は返信間隔が長くなるにつれて悪化すると考えられていた。しかし本研究の結果より、雑談においては返信間隔が1時間の場合であっても、エージェントの評価が高くなるという

ことが明らかとなり、ユーザにとって対話しやすいテキスト対話エージェントの設計に貢献できる知見を得ることができたといえる。

参考文献

- [1] ヤマト運輸, “LINE で受け取るをもっと便利に”, <https://www.kuronekoyamato.co.jp/yt/campaign/renkei/LINE/>, 参照 2022 年 1 月 23 日
- [2] 任天堂, “ソーシャルメディアアカウント”, <https://www.nintendo.co.jp/social/>, 参照 2022 年 1 月 23 日
- [3] Xianchao Wu, Kazushige Ito, Katsuya Iida, Kazuna Tsuboi, Momo Klyen: “りんな：女子高生人工知能”, 言語処理学会 2016 年次大会 発表論文集, (2016)
- [4] Appel J, von der Pütten A, Krämer NC, Gratch J.: “Does Humanity Matter? Analyzing the Importance of Social Cues and Perceived Agency of a Computer System for the Emergence of Social Reactions during Human-Computer Interaction”, *Advances in Human-Computer Interaction*, Vol. 2012, pp. 1–10, (2012)
- [5] Hoxmeier JA, DiCesare C.: “System Response Time and User Satisfaction: An Experimental Study of Browser-based Applications”, *Proceedings of the 6th Americas Conference on Information Systems (AMCIS)*, Proceedings. 2000 Jan 1:347. (2000)
- [6] 志和敏之, 神田崇行, 今井倫太, 石黒浩, 萩田紀博, 安西祐一郎.: “ロボットの反応時間と反応遅延時における問投詞の効果”, *日本ロボット学会誌*, Vol. 27, No. 1, pp.87-95, (2009)
- [7] Mentzer G, Cryan J, Teclehaimanot B.: “Two Peas in a Pod? A Comparison of Face-to-Face and Web Based Classrooms”, *Journal of Technology and Teacher Education*, Vol. 15 No. 2, pp. 233–246, (2007)
- [8] 石井俊太郎, 下田怜奈, 藤江真也.: “チャットボットの発言タイミング制御とその印象評価”, *In IEICE Conferences Archives 2018 Sep 12*. The Institute of Electronics, Information and Communication Engineers, (2018)
- [9] Hoxmeier JA, DiCesare C.: “System Response Time and User Satisfaction : An Experimental Study of Browser-based Applications”, *Proceedings of the 6th Americas Conference on Information Systems (AMCIS)*, pp. 140-145, (2000)
- [10] Kang L, Tan CH, Zhao JL, “The impact of intra-transaction communication on customer purchase behaviour in e-commerce context”,

[11] Gnewuch U, Morana S, Adam MT, Maedche A, “Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction.”, Proceedings of the European Conference on Information Systems (ECIS), (2018)

[12] Holtgraves TM, Ross SJ, Weywadt CR, Han TL, “Perceiving artificial social agents”, Computers in Human Behavior 23, pp.2163-2174, (2007)

[13] Moon. Y, “The effects of physical distance and response latency on persuasion in computer- mediated communication and human-computer communication”, Journal of Experimental Psychology: Applied, Vol.5, No.4, pp.379–392, (1999).

[14]<http://khc.sourceforge.net/>, 参照 2022 年 1 月 23 日

[15] 宮本友樹, 片上大輔, 重光由加, 宇佐美まゆみ, 田中貴紘, 金森等, “ボライトネス・ストラテジーに基づく会話エージェントの言語的な振る舞いの違いが人との関係性構築にもたらす効果～初対面における冗談の心理効果～”. 知能と情報, vol. 30, No. 5, pp. 753-765, (2018)