情報処理学会論文誌 Vol.53 No.1 39–48 (Jan. 2012)

ユーザの翻訳精度判定に対する既判定精度提示の影響

宮部 真衣[1,a)] 吉野 孝[2,b)]

受付日 2011年4月11日, 採録日 2011年10月3日

概要：機械翻訳を利用した多言語間コミュニケーションにおいて不正確な翻訳文を用いると，意思疎通が困難となる．しかし，ユーザは翻訳精度を正しく判断できない場合があるため，翻訳精度に関する気付きを与える必要がある．本研究では，他者やシステムによって判定された精度を提示した場合，ユーザ自身の翻訳精度判定にどのような影響があるかを検証する．本研究の貢献は次の点にまとめられる．(1) 誤った精度を既判定精度として提示した場合でも，既判定精度の影響を受ける場合があることを示した．また，正しい精度（本来の精度）を提示した場合，提示した精度との一致率が最も高くなることを示した．(2) 本来の精度と隣接した評価値を既判定精度として提示した場合，本来よりも高い隣接評価値との一致率の方が，本来よりも低い隣接評価値との一致率よりも高い傾向があることを示した．

キーワード：多言語間コミュニケーション，機械翻訳，翻訳精度，精度判定

Influence of Objective Accuracy Indication in Users' Judgment of Translation Accuracy

Mai Miyabe[1,a)] Takashi Yoshino[2,b)]

Received: April 11, 2011, Accepted: October 3, 2011

Abstract: In communication using machine translation, inaccurate translations impede mutual understanding between communicating individuals. However, people often make inaccurate judgment of translation accuracy. Therefore, it is necessary to consider a method for preventing them from judging accuracy inaccurately. In this study, we verify the influence of objective accuracy indication in their judgment of translation accuracy. The contribution of this study is as follows: (1) Even if an inaccurate indication was shown, the indication influences their judgment. When an accurate indication was shown, the indication influences subjects' judgment strongly. (2) When a higher adjacent evaluation of accurate evaluation was provided to subjects, the evaluation was more influential on them than a lower adjacent evaluation of accurate evaluation.

Keywords: multilingual communication, machine translation, translation accuracy, judgment of translation accuracy

1. はじめに

近年，世界規模のインターネットの普及により，電子メールや掲示板，チャットなどのコミュニケーションツールが広く利用されている．また，インターネットの普及にともなったインターネット上の使用言語の多様化により，ネットワークを介した多言語間コミュニケーションの需要も高まっている．しかし，一般に多言語を十分に習得することは容易ではない．母語以外の言語を用いて十分なコミュニケーションを行うことは困難であり，相互理解ができない可能性が高い [1], [2]．そのため，母語でのコミュニケーションを支援するために，機械翻訳技術を用いた支援が行われている [3], [4]．

近年，機械翻訳技術は急速に進展しているが，高精度な翻

1 東京大学知の構造化センター
Center for Knowledge Structuring, The University of Tokyo, Bunkyo, Tokyo 113–8656 Japan
2 和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University, Wakayama 640–8510 Japan
a) mai.miyabe@gmail.com
b) yoshino@sys.wakayama-u.ac.jp

訳を行うことは困難である．精度の低い文章は話者間の相互理解を困難にし，円滑なコミュニケーションの妨げとなる．また，機械翻訳を介したコミュニケーションでは，低精度な文章を用いることによって思い違いが高頻度で発生すると報告されている [5], [6]．したがって，円滑にコミュニケーションを行うためには，精度の高い文章を作成しなければならない．翻訳精度を向上させるための手法に，翻訳リペアがある [7]．翻訳リペアは，文章中の不適切な翻訳箇所を減少させるために，翻訳結果を確認しながら，ユーザ自身が入力文を書き換えていく作業を指す．折り返し翻訳[*1]を利用することにより，母語のみを用いて作成した翻訳文の精度を確認することができる．

これまでに，翻訳リペア作業において，修正のためのルール教示を行うことによる影響が検証されている [8]．この研究では，事前にユーザに対して修正のためのルールを教示し，翻訳回数を制限した翻訳リペア実験を行っている．しかし，翻訳精度は，翻訳する言語ペアや翻訳システムに依存しており，同じテキストを翻訳しても，同じ結果を得ることができるとは限らない [9]．したがって，ルール教示によって書き換えたテキストの翻訳精度についても，翻訳する言語ペアや翻訳システムによって異なると考えられ，ルール教示に基づく修正によって翻訳精度が改善するとは限らないという問題がある．また，翻訳リペアを適用する場面によって，要求される翻訳精度は異なると考えられる．高精度な翻訳結果が必要な場面では，ルール教示がされなくなるまで修正する，という利用方法が想定され，「ルール教示がされなくなる」ということがユーザにとっての修正終了の判断指標となる．一方，ある程度の翻訳精度で許容される場面であれば，ルール教示がされている場合であっても，必要とされる翻訳精度が確保されていれば，修正をせずにそのまま利用してもよいと考えられる．この場合，修正終了の判断指標がないため，ユーザ自身が翻訳精度を判断し，修正を終了することになる．しかし，これまでの研究において，ユーザの不正確な翻訳精度の判断（不正確判定）が発生しうることが示されている [10]．この実験では，翻訳リペア作業を実施した際，修正すべき文のうち，平均 7%，最大 23%の文に対し不正確判定が発生し，修正が行われなかったと述べられている．この結果は，本来修正が必要な翻訳結果を見た際，ユーザがまったく修正をしなかった文について分析されたものであり，修正作業を行った際の修正終了判断における不正確判定（修正を行ったものの，精度が十分に向上していない段階で修正終了してもよいと判断したもの）については考慮されていないため，実際にはより多くの不正確判定が発生していると考えられる．不正確判定は，低精度なメッセージの利用につながり，コミュニケーションにおいて思い違いを引き起こす可能性が高い．そのため，ある程度の翻訳精度で許容される場面では，ユーザの不正確判定を防ぐ仕組みが必要となる [10]．

本研究では，高精度な翻訳結果が要求される場面だけではなく，ある程度の精度であれば許容される場面への翻訳リペアの適用を想定し，ユーザの不正確判定を防ぐ仕組みを検討する．不正確判定はユーザ自身の精度の判断基準が他のユーザと異なることが原因で発生していると考えられる．そこで，不正確判定を防ぐための仕組みとして，客観的精度の提示を考える．本論文では，翻訳文に対する客観的精度（**既判定精度**）をユーザに提示することによる影響を検証する．

2. 精度表示によるユーザへの影響

我々はこれまでに，不正確判定の減少を目的とし，翻訳自動評価手法を用いた翻訳精度表示手法の提案およびその効果の検証を行ってきた [11]．この研究においては，翻訳精度に関する気付きを与えることによって，ユーザの判断に影響を与え，不正確判定の減少ができる可能性があるのではないかと考え，客観的な指標として翻訳精度を提示した場合の効果を検証している．しかし，実験の結果，精度表示を行うことによる不正確判定の減少効果は見られなかった．一方，アンケートにおいて，精度表示は役に立つ可能性があるものの，表示された精度が信用できないというコメントが得られた．この実験においては，十分な精度になるまで翻訳文を修正する作業を行っている．また，低精度な文を「高精度である」と判断した場合を不正確判定とし，不正確判定の発生数についての検証を行った．しかし，翻訳精度を提示しているものの，その文に対するユーザの評価については検証していない．つまり，翻訳精度を提示することが，本当にユーザに対して影響を与えていなかったのかどうかは検証できていない．

そこで，本論文では，翻訳精度を提示した場合，ユーザ自身の翻訳精度判定にどのような影響があるかを検証する．

3. 実験

既判定精度提示の影響を検証するために，既判定精度提示による翻訳精度評価実験を行った．

実験の被験者は，大学生および大学院生 30 名である．被験者の年齢は 18 歳から 25 歳（平均 22 歳）である．

3.1 検証項目

本実験では，以下の項目を明らかにする．

［検証項目 1］ 既判定精度の提示は，ユーザの評価結果に影響を与えるか？

［検証項目 2］ 既判定精度の提示は，評価時間に影響を与えるか？

[*1] 折り返し翻訳とは，他言語への翻訳結果を再度原言語へと翻訳することである．

表 1 実験に用いたテキストの一部
Table 1 Examples of sentences used in the experiment.

テキストセット	原文	折り返し翻訳文
精度 1 のテキスト	彼の感動はだんだん静まっていった。	ただ彼の興奮はだんだんに高まることです。
	ポーターはいませんので台車をお使いください。	運搬夫の姿は台車を使ってもらうことに見えないためです。
精度 2 のテキスト	おじいさんは入れ歯の手入れをした。	おじいさんはこの入れ歯条件です。
	どうやったら私にそれが快速急行だとわかりますか。	私のいわゆる高速の急行はそれに分かりません。
精度 3 のテキスト	このオレンジは新鮮でないこともあって余りおいしくない。	このオレンジは新鮮でないのもあって，残っておいしくない。
	だいたいどのあたりでなくしたか見当はつきますか。	大体どこでなくしたので、知っていますか？
精度 4 のテキスト	あまりの混雑にけが人が出たという例さえある。	非常に行き過ぎた混雑にケガ人が出現した例まである。
	ロサンゼルス行きの列車の座席を予約したいんですけど。	ロサンゼルスの到着する列車の席を予約したいです。
精度 5 のテキスト	彼は最後の一球に悔いを残した。	彼は最後に投げた球に対し後悔を残した。
	もしもし、こちらは七百四号室のジョンソンです。	こんにちは、704 番の室のジョンソンです。

精度 1 のテキスト，精度 2 のテキスト，精度 3 のテキスト，精度 4 のテキスト，精度 5 のテキストは，それぞれ事前に評価した結果が 1，2，3，4，5 のテキストを抽出したものである．

3.2 実験内容

被験者は，2 つの文（原文および折り返し翻訳文）を比較し，折り返し翻訳文が原文と同じ意味になっているかどうかを評価する．評価指標については，Walker らの適合性評価（5 段階評価）[12] を用いた．

適合性評価の評価基準を以下に示す．

5： All（完全に一緒）
4： Most（文法など少し問題はあるが，まあまあ一緒）
3： Much（意味は大体つかめる）
2： Little（雰囲気は残っているが，もとの意味は分からない）
1： None（まったく駄目）

評価の際は，1 組の文に対して 30 秒以内で評価するものとした[*2]．

3.3 影響要因

本研究では，評価手法として Walker らの適合性評価（5 段階評価）を用いる．5 段階評価のうち，どの評価値を既判定精度として提示するかによって，効果は変わると考えられる．また，用いるテキストの本来の精度も，結果に影響すると考えられる．そこで，本実験では，以下のような実験条件において検証を行う．

(1) 用いるテキストの本来の精度
5 種類（本来の精度評価値が 1，2，3，4，5）のテキストを用いる．本論文では，本来の精度評価値が 1，2，3，4，5 のテキストを，それぞれ**「精度 1 のテキスト」「精度 2 のテキスト」「精度 3 のテキスト」「精度 4 のテキスト」「精度 5 のテキスト」**と呼ぶこととする．

(2) 提示する既判定精度
既判定精度を 6 種類（非表示，1，2，3，4，5）とし，提示する．

被験者は，各条件を組み合わせた 30 種類（5 種類のテキスト × 6 種類の既判定精度）の状態において，それぞれ 20 文ずつ評価を行う．

3.4 利用テキスト

3.3 節で述べたように，本実験においては，すでに評価が行われているテキストを用いる必要がある．そこで，これまでに行ってきた実験 [7], [10] において，すでに精度が評価されたテキスト[*3]を用いることとした．これらの実験においては，3 名の評価者により，Walker らの適合性評

[*2] 30 秒という制限時間については，被験者が長時間考え込まないようにするための目安として設定した．非母語話者の日本語読解に関する先行研究 [13] によると，上級の非母語話者は 1 秒あたり約 3.8 文字の読解が可能である．また，非母語話者は母語話者よりも読解所要時間が長いと述べられている．先行研究より，非母語話者は 30 秒で 114 文字程度を読解でき，母語話者である日本人はさらに多くの文字を読解できると考えられる．3.4 節で示すように，今回の実験で被験者が比較する 2 つの文の合計文字数の最大値は 82 文字であり，30 秒以内に十分読解可能であると考えられる．

[*3] これらの研究で用いられたテキストは，機械翻訳試験文 [14] および会話表現データベース [15] の一部である．

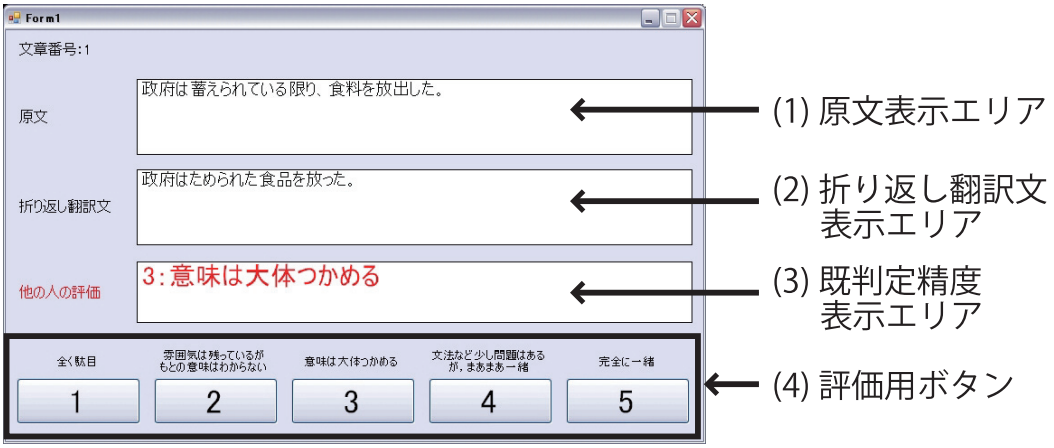


図 1 評価用ツール
Fig. 1 Screenshot of the experimental tool.

価 [12] を用いた精度評価が行われている．なお，この評価は主観評価であるため，同じ文に対する評価結果が，3 名の評価者間で異なる場合もある．3 名の評価が一致しているものについては，ばらつきが小さく，一般的に同じ評価がされやすいと考えられる[*4]．そこで，3 名の評価が一致しているテキストを各評価値につき 120 文抽出し，利用することとした．

実験で利用したテキストの一部を**表 1** に示す．表 1 における精度 1 のテキスト，精度 2 のテキスト，精度 3 のテキスト，精度 4 のテキスト，精度 5 のテキストは，それぞれ評価結果が 1，2，3，4，5 のテキストを抽出したものの一部である．入力文の最小文字数は 15 文字，最大文字数は 29 文字であり，入力文に対する折り返し翻訳文の最小文字数は 9 文字，最大文字数は 53 文字である．また，比較する 2 文の合計文字数は，最小で 25 文字，最大で 82 文字であった．

3.5 実験ツール

本実験で用いた実験ツールの画面を**図 1** に示す．原文表示エリア（図 1 (1)）に原文が，折り返し翻訳文表示エリア（図 1 (2)）に評価する折り返し翻訳文が表示される．被験者は評価用ボタン（図 1 (4)）を用いて評価を行う．既判定精度表示エリア（図 1 (3)）には，5 段階評価におけるいずれかの評価値または「他の人の評価はありません」というテキストが表示される．なお，実験ツールでは，必ず確認しなければならない原文と折り返し翻訳文と，評価用ボタンの間に既判定精度表示エリアを配置し，既判定精度については原文・折り返し翻訳文よりも大きなフォントで赤字表記し，強調表示するようにした．

これまでに行った精度提示による文章修正実験 [11] では，システムによる自動判定精度が提示されると説明した．この実験では，「誤った精度が表示されたため，表示を無視した」という意見が見られており，自動判定精度はユーザに信頼されていなかったと考えられる．そこで本実験では，提示評価に関する信頼性を高めるために，既判定精度表示エリアに表示される内容は，他の人が事前に評価した結果であると説明した．

なお，文章はランダムに表示されるようになっており，評価する文の順番は各被験者によって異なる．

4. 実験結果

4.1 精度評価結果

本実験では，各テキストセット（120 文）の評価において，6 種類の既判定精度提示を行っている．また，各提示において，被験者は 20 文ずつ評価を行っている．そこで，各提示を行った際の各評価値に対する評価結果の該当文数を調べた．**表 2** に各既判定精度提示時における評価結果の平均該当文数[*5]を示す．表 2 において，テキストセットが「精度 1 のテキスト」，被験者の評価結果が「1」，提示した既判定精度が「精度 1」の場合の値である 14.6 は，本来の精度が 1 である文に対する既判定精度として「1」を提示した場合に，その文の精度を被験者が「1」であると評価した数の平均値を意味する．また，表 2 において，被験者の評価結果が本来の評価値と同じであった場合の該当文数を太字で示す．

表 2 より，各テキストセットにおいて，本来の精度評価値と同じ評価を行った場合に，該当数が多くなる傾向が見られた．

また，各提示を行った際の各評価値に対する評価結果の該当文数に関して，いくつかの条件において有意差が見られる．そこで，表 2 において，6 種類の提示間で有意差の見られたものについて，多重比較を行った．多重比較の結

[*4] 本実験で用いた Walker らの適合性評価は，原則として 2 名以上で行うものである．そのため，3 名の評価者による評価が一致している場合は，Walker らの適合性評価における最小人数（2 名）での評価の一致よりも，信頼性が向上すると考えられる．

[*5] 表 2 における値は，被験者 30 名の平均値である．

表 2 各既判定精度提示時における評価結果の平均該当文数
Table 2 Average number of each evaluated sentence in each experimental condition.

テキストセット	被験者の評価結果	提示した既判定精度						有意確率
		非表示（文）	精度 1（文）	精度 2（文）	精度 3（文）	精度 4（文）	精度 5（文）	
精度 1 のテキスト	1	12.3	**14.6**	11.1	11.3	11.0	11.7	0.000*
	2	6.7	4.7	8.2	7.1	7.8	7.1	0.000*
	3	0.8	0.6	0.6	1.4	1.0	1.0	0.037*
	4	0.2	0.1	0.1	0.1	0.2	0.2	0.595
	5	0.0	0.0	0.0	0.1	0.0	0.0	0.221
精度 2 のテキスト	1	3.2	4.7	2.3	2.4	2.7	2.5	0.000*
	2	9.5	8.9	**10.8**	8.5	8.6	8.3	0.025*
	3	5.0	4.5	4.9	6.8	5.8	5.9	0.001*
	4	1.6	1.4	1.6	1.7	2.6	2.3	0.002*
	5	0.7	0.5	0.4	0.6	0.4	1.0	0.279
精度 3 のテキスト	1	0.5	0.8	0.2	0.4	0.3	0.2	0.002*
	2	3.5	4.3	4.7	2.9	2.5	2.7	0.000*
	3	8.4	8.3	7.5	**9.6**	7.7	7.6	0.003*
	4	5.9	4.9	6.0	5.6	7.6	6.7	0.046*
	5	1.8	1.7	1.6	1.5	1.8	2.8	0.021*
精度 4 のテキスト	1	0.2	0.0	0.0	0.1	0.1	0.1	0.398
	2	0.6	1.0	1.4	0.9	0.6	0.6	0.018*
	3	4.4	4.6	4.3	4.9	3.7	3.4	0.021*
	4	8.0	8.1	7.7	8.5	**9.9**	7.9	0.103
	5	6.9	6.2	6.6	5.6	5.6	8.0	0.017*
精度 5 のテキスト	1	0.0	0.0	0.0	0.0	0.0	0.0	0.549
	2	0.1	0.3	0.2	0.1	0.1	0.0	0.099
	3	0.8	0.9	0.7	1.0	0.6	0.6	0.623
	4	3.7	3.0	3.8	3.8	4.5	2.8	0.016*
	5	15.3	15.8	15.2	15.1	14.8	**16.6**	0.026*

*：有意差あり（Friedman 検定）$p<0.05$
表中の値は，被験者 30 名の平均値である．

果，有意差のあった提示内容の組合せを**表 3** に示す．それぞれのテキストセットにおける多重比較結果を以下に示す．

(1) 精度 1 のテキスト
表 3 より，評価結果 1 の該当数に関して，既判定精度 1 と，その他の既判定精度との間に有意差が見られる．表 2 を見ると，精度 1 のテキストにおいて既判定精度 1 を提示した場合，被験者が評価結果を 1 と判断した数は平均 14.6 文であり，他の条件と比較して有意に多いことが分かる．また，評価結果 2 の該当数については，既判定精度 1 と，4 つの既判定精度（非表示，2，3，4）との間に有意差が見られた．表 2 より，既判定精度 1 を提示した場合の評価結果 2 の該当数は平均 4.7 文，その他の条件では平均 6.7 文以上となっている．既判定精度 1 を提示した場合，被験者が評価結果 2 と判断した数が少ないことが分かる．このことから，本来の精度が 1 のテキストに対して，誤った精度を提示すると，本来の精度よりも高い評価をする数が多くなる傾向が見られた．

(2) 精度 2 のテキスト
表 3 より，評価結果 1 の該当数に関して，既判定精度 1 と，4 つの既判定精度（2，3，4，5）との間に有意差が見られる．表 2 を見ると，精度 2 のテキストにおいて既判定精度 1 を提示した場合，被験者が評価結果を 1 と判断した数が有意に多いことが分かる．このことから，本来の精度が 2 のテキストに対して，本来よりも低い精度を提示すると，本来の精度よりも低い評価をする数が多くなる傾向が見られた．

(3) 精度 3 のテキスト
表 3 より，評価結果 2 の該当数に関して，既判定精度 1，2 と，既判定精度 3，4，5 との間にそれぞれ有意差が見られる．表 2 を見ると，精度 3 のテキストにおいて既判定精度 1，2 を提示した場合，被験者が評価結果を 2 と判断した数は平均 4.3 文および 4.7 文である．一方，既判定精度 3，4，5 を提示した場合は平均 2.9 文以下であり，既判定精度 1，2 を提示した場合の方が有意に多いことが分かる．このことから，本来の精

表 3 各既判定精度提示時における評価結果の該当文数に関する多重比較結果
Table 3 Results of multiple comparison between differences of average number of sentences in each experimental condition.

テキストセット	被験者の評価結果	有意差のあった既判定精度提示の組合せ
精度 1 のテキスト	1	非表示—1，1—2，1—3，1—4，1—5
	2	非表示—1，1—2，1—3，1—4
	3	なし
精度 2 のテキスト	1	1—2，1—3，1—4，1—5
	2	2—3
	3	非表示—3，1—3，2—3
	4	1—4，1—5
精度 3 のテキスト	1	なし
	2	1—3，1—4，1—5，2—3，2—4，2—5
	3	2—3
	4	なし
	5	なし
精度 4 のテキスト	2	非表示—2
	3	なし
	5	なし
精度 5 のテキスト	4	4—5
	5	4—5

組合せにおける各数値は，提示した既判定精度を意味する．

表 4 平均評価時間
Table 4 Average time of evaluation.

テキストセット		提示した既判定精度						有意確率
		非表示（秒）	精度 1（秒）	精度 2（秒）	精度 3（秒）	精度 4（秒）	精度 5（秒）	
精度 1 のテキスト	平均	8.3	7.9	8.3	8.4	8.2	8.7	0.006*
	標準偏差	3.1	2.9	2.6	2.8	2.7	2.7	
精度 2 のテキスト	平均	8.4	8.1	8.4	8.3	8.7	8.6	0.110
	標準偏差	2.9	3.0	3.0	3.0	3.2	2.8	
精度 3 のテキスト	平均	8.1	8.4	8.6	8.0	8.1	8.3	0.350
	標準偏差	2.6	2.6	3.2	2.8	2.3	2.8	
精度 4 のテキスト	平均	8.0	7.9	7.8	7.8	7.8	7.8	0.479
	標準偏差	2.5	2.0	2.3	2.8	2.7	2.6	
精度 5 のテキスト	平均	7.1	7.1	7.1	7.0	7.1	6.3	0.001*
	標準偏差	2.4	2.4	2.4	2.4	2.3	2.2	

*：有意差あり（Friedman 検定）$p<0.05$

度が 3 のテキストに対して，本来よりも低い精度（1 および 2）を提示すると，本来の精度よりも低い評価をする数が多くなる傾向が見られた．

(4) 精度 4 のテキスト
精度 4 のテキストについては，評価結果 2，3，5 において有意差が見られるが，多重比較の結果，特徴的な差異は見られなかった．

(5) 精度 5 のテキスト
表 3 より，評価結果 4，5 の該当数に関して，既判定精度 4 と，既判定精度 5 との間に有意差が見られる．表 2 を見ると，精度 5 のテキストにおいて既判定精度 4 を提示した場合，被験者が評価結果を 4 と判断した数は，既判定精度 5 を提示した場合よりも多くなり，被験者が評価結果を 5 と判断した数は，既判定精度 5 を提示した場合よりも少なくなっていた．

4.2 評価時間

各条件における，1 文あたりの平均評価時間を**表 4** に示す．表 4 より，本実験ではどの実験条件においても平均 7 秒～9 秒で評価が行われていた．

4.3 アンケート結果

実験後に実施したアンケートの結果を**表 5** に示す．表 5 の各質問に対する評価は，1：強く同意しない，2：同意し

表 5 アンケート結果
Table 5 Results of questionnaire.

質問番号	質問	評価値（人）					中央値	最頻値
		1	2	3	4	5		
1	表示されている他の人の評価結果を見た.	1	5	6	12	6	4	4
2	他の人の評価結果を参考にして評価を行った.	7	9	4	10	0	2	4

5 段階評価の評価値：1：強く同意しない，2：同意しない，3：どちらともいえない，4：同意する，5：強く同意する

ない，3：どちらともいえない，4：同意する，5：強く同意する，の 5 段階評価によって行った．

表 5 の質問 1 より，「表示されている他の人の評価結果を見た」という質問項目に対して「5：強く同意する」と回答した被験者が 6 名，「4：同意する」と回答した被験者が 12 名となっており，被験者の半数程度は評価結果を確認していたことが分かる．一方，質問 2 を見ると，「他の人の評価結果を参考にして評価を行った」という質問項目に対して「1：強く同意しない」と回答した被験者が 7 名，「2：同意しない」と回答した被験者が 9 名となっており，16 名の被験者が提示された評価結果を参考にしていなかったと回答していた．

5. 考察

5.1 既判定精度提示の影響

本節では，**［検証項目 1］既判定精度の提示は，ユーザの評価結果に影響を与えるか？**について議論する．

実験におけるアンケート結果（表 5）から，被験者の半数程度は提示された評価結果を参考にしていなかったと回答していた．しかし，自由記述では，「赤く大きな文字で提示されており，目立つので目に入った」といった記述が見られた．被験者本人は参考にしていなかったと回答していても，無意識のうちに影響を受けていた可能性もあり，本当に影響されていなかったのかを検証することは難しい．

今回の実験では，既判定精度を提示していない状態（非表示）でも評価を行った．表 3 より，精度 1 のテキストセット，精度 2 のテキストセット，精度 4 のテキストセットについては，いくつかの提示条件において非表示での結果との有意差が見られており，既判定精度を提示することにより，何らかの影響を与える可能性があると考えられる．そこで，本研究では，提示した既判定精度と同じ評価を行った場合，影響を受けたと見なすこととし，既判定精度の影響について検証する．

提示した既判定精度と被験者の評価結果との一致数を**表 6** に示す．表 6 における一致率は，30 名の合計を 30 名の全体の評価数（600 文）で割ったものである．

表 6 より，すべてのテキストセットにおいて，本来の精度を既判定精度として提示した場合の一致率が最も高い．また，すべてのテキストセットにおいて，本来の精度と隣接した評価値（**隣接評価値**）を提示した場合の一致率が比較的高くなっている．

精度 1 のテキストおよび精度 5 のテキストについては，隣接評価値（精度 1 のテキストについては既判定精度 2，精度 5 のテキストについては既判定精度 4）の一致率が 2 番目に高い．一方，隣接していない評価値を提示した場合の一致率は，すべて 10%以下である．

精度 2 のテキスト，精度 3 のテキスト，精度 4 のテキストについては，本来の精度よりも 1 つ高い精度を提示した場合の一致率が 2 番目に高く，本来の精度よりも 1 つ低い精度を提示した場合の一致率が 3 番目に高い．

精度 1 のテキスト，精度 5 のテキストと精度 2 のテキスト，精度 3 のテキスト，精度 4 のテキストを比較すると，本来の精度を既判定精度として提示した場合の一致率が前者は 70%以上であるのに対し，後者は 50%程度である．3.2 節で述べたように，本研究では 5 段階の適合性評価を用いている．5 段階評価の指標を見ると，1 と 5 については「まったく駄目」「完全に一緒」となっており，比較的判断しやすいのに対して，2，3，4 については判断が曖昧になる可能性があると考えられる．精度 1 のテキスト，精度 5 のテキストは，それぞれ本来の翻訳精度が 1（まったく駄目）および 5（完全に一緒）の文を集めたものであり，比較的判断しやすく，本来の精度との一致率が高くなった可能性があると考えられる．

また，精度 1 のテキスト，精度 5 のテキストについては，隣接評価値がそれぞれ 1 つである．一方，精度 2 のテキスト，精度 3 のテキスト，精度 4 のテキストについては，隣接評価値がそれぞれ 2 つ（本来よりも低い評価値および高い評価値）ずつ存在する．精度 2 のテキスト，精度 3 のテキスト，精度 4 のテキストにおける隣接評価値の一致数を見ると，どのテキストセットについても，本来よりも高い評価値の一致数が多い．また，表 2 を見ても，同様の傾向が見られる．

以上のことから，正しい精度（本来の精度）を提示した場合の一致率が最も高いものの，誤った精度を既判定精度として提示した場合でも，既判定精度の影響を受ける場合もあることが分かった．また，隣接評価値を既判定精度として提示した場合，本来よりも高い隣接評価値との一致率の方が，本来よりも低い隣接評価値との一致率よりも高い

表 6 提示した既判定精度と被験者の評価結果が一致した数
Table 6 Number of sentences which evaluation is coincident with an indication.

テキストセット	提示した既判定精度	評価結果の一致数			
		平均（文）	標準偏差（文）	30 名の合計（文）	一致率（%）
精度 1 のテキスト	精度 1	14.6	4.3	439	73.2
	精度 2	8.2	4.5	247	41.2
	精度 3	1.4	1.9	41	6.8
	精度 4	0.2	0.6	7	1.2
	精度 5	<0.1	0.2	1	0.2
精度 2 のテキスト	精度 1	4.7	3.3	142	23.7
	精度 2	10.8	3.8	324	54.0
	精度 3	6.8	3.2	205	34.2
	精度 4	2.6	2.1	78	13.0
	精度 5	1.0	1.5	30	5.0
精度 3 のテキスト	精度 1	0.8	1.1	24	4.0
	精度 2	4.7	3.4	141	23.5
	精度 3	9.6	3.7	288	48.0
	精度 4	7.6	3.9	227	37.8
	精度 5	2.8	2.4	83	13.8
精度 4 のテキスト	精度 1	<0.1	0.2	1	0.2
	精度 2	1.4	1.5	43	7.2
	精度 3	4.9	3.2	147	24.5
	精度 4	9.9	3.5	298	49.7
	精度 5	8.0	5.1	240	40.0
精度 5 のテキスト	精度 1	<0.1	0.2	1	0.2
	精度 2	0.2	0.5	6	1.0
	精度 3	1.0	1.1	29	4.8
	精度 4	4.5	3.2	135	22.5
	精度 5	16.6	2.7	497	82.8

傾向が見られた．

5.2 評価時間への影響

本節では，**［検証項目 2］既判定精度の提示は，評価時間に影響を与えるか？**について議論する．実験では，「正しい精度」および「誤った精度」をそれぞれ既判定精度として提示した．誤った精度を提示した場合，ユーザの直感と合わず，評価時間が長くなる可能性があると考えられる．

表 4 より，精度 1 のテキストおよび精度 5 のテキストにおいて，6 種類の提示間で有意差が見られた．そこで，精度 1 のテキストおよび精度 5 のテキストにおける評価時間に関して，多重比較を行った[*6]．多重比較の結果を**表 7** に示す．多重比較の結果，精度 1 のテキストについては既判定精度 1 と 5 の間に有意差が見られた．また，精度 5 のテキストについては，既判定精度 5 とその他 5 種類の既判定精度との間に有意差が見られた．

表 4 より，精度 5 のテキストにおける既判定精度「5」提示時の平均評価時間は 6.3 秒である．一方，その他の精度を提示した場合，平均評価時間は 7.0 秒あるいは 7.1 秒となっている．したがって，精度 5 のテキストにおいて，既判定精度として「5」を提示した場合，評価時間が有意に短くなっていることが分かった．

[*6] ホルムの方法 [16] により多重比較を行った．

表 7 平均評価時間に関する多重比較結果
Table 7 Results of multiple comparison between differences of average time of evaluation.

テキストセット	有意差のあった既判定精度提示の組合せ
精度 1 のテキスト	1—5
精度 5 のテキスト	非表示—1，1—5，2—5，3—5，4—5

組合せにおける各数値は，提示した既判定精度を意味する．

一方，その他の提示に関しては，正しい精度を提示した場合および誤った精度を提示した場合の評価時間に有意差が見られなかった．そのため，既判定精度の提示は，評価時間に大きな影響を与えていないと考えられる．

5.3 今後の課題

5.3.1 実験条件について

4.3 節で述べたように，アンケートの結果，被験者の半数程度は，「評価結果を参考にしていなかった」と回答していた．また，アンケートの自由記述においては，「明らか

に間違っている評価が表示されていた」という記述が多く見られた．今回の実験では，誤った精度を提示することによる影響の検証を行った．一方，正しい提示をすることによって，異なる効果が得られる可能性がある．今後，正しい精度を提示することによる効果について検証を行う必要がある．

5.3.2 提示手法について

既判定精度提示の最終的な目的は，不正確判定防止手法として用いることである．本実験の結果，既判定精度を提示することによる影響は見られたものの，被験者が完全に影響されることはなかった．不正確判定防止手法として用いるためには，ユーザの判断に対しての影響を強める必要がある．

ユーザ自身の意見へ影響を与える現象として，「同調」と呼ばれるものがある [17], [18]．同調とは，集団に所属した場合に，自分自身の意見を曲げて，多数派に従ってしまう現象である．今回の実験では，他の人の評価結果として，1つの評価のみを提示した．そのため，強い同調圧力は発生していなかったと考えられる．不正確判定防止手法を実現するためには，正確な精度判定手法を構築し，また提示精度に関して同調圧力を与えるような仕組みやインタフェースを検討する必要があると考えられる．

6. おわりに

本論文では，翻訳リペアにおけるユーザの不正確判定防止手法として，客観的精度の提示を想定し，客観的精度の提示効果を検証するために，既判定精度の提示による効果の検証を行った．本研究の貢献は以下の 3 点にまとめられる．

(1) 正しい精度（本来の精度）を提示した場合，提示した精度との一致率が最も高くなることを示した．また，誤った精度を既判定精度として提示した場合でも，既判定精度の影響を受ける場合があることを示した．
(2) 本来の精度と隣接した評価値を既判定精度として提示した場合，本来よりも高い隣接評価値との一致率の方が，本来よりも低い隣接評価値との一致率よりも高い傾向があることを示した．
(3) 既判定精度の提示による，評価時間に対する大きな影響はないことを示した．

本研究は機械翻訳を介した多言語間コミュニケーションにおける翻訳リペアの適用を想定した実験を行い，既判定精度を提示することにより，ユーザの判定に影響を与えうることを示した．本研究により得られた結果は，翻訳リペアにおける不正確判定の防止のために利用できる．また，近年，ユーザによる評価を採用した Web 上のサービスも増加してきているが，そのようなユーザが評価行為を行う場面のあるシステムを設計する際に，本研究の知見が参考になる可能性があると考えられる．

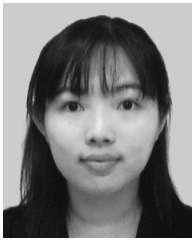
今回は，誤った精度の提示による効果を検証した．今後，正しい精度を提示することによる効果について検証を行う必要がある．また，ユーザの判断に影響を与えやすいインタフェースの検討を行っていく必要がある．

謝辞 本研究の一部は，日本学術振興会科学研究費基盤研究（B）（22300044）の補助を受けた．

参考文献

[1] Aiken, M.: Multilingual Communication in Electronic Meetings, *ACM SIGGROUP, Bulletin*, Vol.23, No.1, pp.18–19 (2002).
[2] Tung, L.L. and Quaddus, M.A.: Cultural differences explaining the differences in results in GSS: Implications for the next decade, *Decision Support Systems*, Vol.33, No.2, pp.177–199 (2002).
[3] 藤井薫和，重信智宏，吉野 孝：機械翻訳を用いた異文化間チャットコミュニケーションにおけるアノテーションの評価，情報処理学会論文誌，Vol.48, No.1, pp.63–71 (2007).
[4] Inaba, R.: Usability of Multilingual Communication Tools, Lecture Notes in Computer Science 4560, pp.91–97 (2007).
[5] 山下直美，石田 亨，平田圭二：機械翻訳を用いた対話における思い違いに関する分析，情報処理学会論文誌，Vol.47, No.1, pp.112–120 (2006).
[6] 宮部真衣，吉野 孝：機械翻訳を介したチャットコミュニケーションにおける精度判定に基づく送信拒否の適用可能性，情報処理学会論文誌，Vol.51, No.3, pp.784–795 (2010).
[7] 宮部真衣，吉野 孝，重信智宏：折返し翻訳を用いた翻訳リペアの効果，電子情報通信学会論文誌，Vol.J-90-D-I, No.12, pp.3142–3150 (2007).
[8] 山下直美，坂本知子，野村早恵子，石田 亨，林 良彦，小倉健太郎，井佐原均：機械翻訳へのユーザの適応と書き換えへの教示効果に関する分析，情報処理学会論文誌，Vol.47, No.4, pp.1276–1286 (2006).
[9] Ogura, K., Hayashi, Y., Nomura, S. and Ishida, T.: User Adaptation in MT-mediated Communication, *The 1st International Joint Conference on Natural Language Processing* (*IJCNLP-04*), pp.596–601 (2004).
[10] Miyabe, M., Yoshino, T. and Shigenobu, T.: Effects of Repair Support Agent for Accurate Multilingual Communication, Lecture Notes in Computer Science 5351, pp.1022–1027 (2008).
[11] 宮部真衣，吉野 孝：機械翻訳を用いた高精度な文章作成のための翻訳精度表示の影響，情報処理学会研究報告，デジタルドキュメント研究会，Vol.2010-DD-77, No.5, pp.1–7 (2010).
[12] Walker, K., Bamba, M., Miller, D., Ma, X., Cieri, C. and Doddington, G.: Multiple-Translation Arabic (MTA) Part 1, Linguistic Data Consortium (LDC) catalog number LDC2003T18 and ISBN 1-58563-276-7.
[13] 柳澤絵美，大木理恵，鈴木美加：アイカメラを使って観察した日本語学習者の読みの特徴—レベルの違いから見えてくるもの，東京外国語大学留学生日本語教育センター論集（36），pp.1–12 (2010).
[14] NTT Natural Language Research Group, available from ⟨http://www.kecl.ntt.co.jp/icl/mtg/resources/index.php⟩
[15] ATR 音声翻訳通信研究所：会話表現データベース，available from ⟨http://www.atr-p.com/sdb.html⟩

[16] 入戸野宏：心理生理学データの分散分析，生理心理学と精神生理学，Vol.22, No.3, pp.275–290 (2004).
[17] Asch, S.E.: Effects of group pressure upon the modification and distortion of judgment, *Groups, leadership and men: Research in human relations*, Guetzkow, H. (Ed.), pp.177–190, Carnegie Press (1951).
[18] 池田謙一，唐沢 穣，工藤恵理子，村本由紀子：社会心理学，有斐閣 (2010).



宮部 真衣 （正会員）

1984 年生．2006 年和歌山大学システム工学部デザイン情報学科中退．2008 年同大学大学院システム工学研究科システム工学専攻博士前期課程修了．2011 年同大学院システム工学研究科システム工学専攻博士後期課程修了．博士（工学）．現在，東京大学知の構造化センター特任研究員．多言語間コミュニケーション支援，災害時のコミュニケーションに関する研究に従事．

吉野 孝 （正会員）

1969 年生．1992 年鹿児島大学工学部電子工学科卒業．1994 年同大学大学院工学研究科電気工学専攻修士課程修了．現在，和歌山大学システム工学部デザイン情報学科准教授．博士（情報科学）．コミュニケーション支援の研究に従事．