情報処理学会論文誌 Vol. 52 No. 1 131–139 (Jan. 2011)

用例の正確性評価を目的とした用例評価手法の比較

福島 拓[†1] 吉野 孝[†2]

現在，在日外国人数や訪日外国人数は増加傾向にあり，多言語によるコミュニケーションの機会が増加している．多言語環境支援の一方法として，用例を正確に多言語に翻訳した用例対訳が用いられている．用例対訳は正確な多言語間コミュニケーション支援が可能なため，医療分野などの正確性が求められる分野で多く利用されており，動的な用例対訳収集も行われている．しかし，これまで用例対訳を含めた用例収集の取り組みにおいて用例の正確性評価が行われてこなかった．用例の正確性評価がなされていない用例は，医療などの正確性を求められる分野で使用することはできない．そこで我々は，用例の正確性評価手法の確立を目指して複数の評価手法の比較評価を行った．本論文では，比較実験の結果から次の知見を得た．(1) 用例の正確性評価実験の結果，評価軸を提示していない評価手法と提示した評価手法の間に相関関係がみられなかった．用例の正確性評価には，評価軸を明確に提示した評価手法が必要である．(2) 用例の評価において，評価者は詳細な評価を行うことを好んだ．このため，2 値の評価段階よりも複数の評価段階を選択可能とした評価手法が評価者から支持された．

Comparison of Methods Used for Accurate Evaluation of Example Sentences

Taku Fukushima[†1] and Takashi Yoshino[†2]

Recently, there has been an increase in the number of foreign nationals residing in Japan as well as the number of foreigners visiting the country. Consequently, there may be increased communication among people speaking different native languages. A parallel text that combines example sentences and their accurate translation is used in a multilingual environment, for example, hospitals where the staff members and patients speak different languages. Some researchers are dynamically collecting example sentences or parallel texts. However, they have not yet evaluated the accuracy of the collected sentences. An unevaluated example sentence cannot be used at needed accurate communication fields. Therefore, we evaluated evaluation methods to evaluate example sentence. We observed the following from the experiments in which evaluation methods were compared. (1) From the result of the experiment, there was no correlative relationship between having evaluation axes and not having evaluation axes. It is necessary to indicate the evaluation axis explicitly when using this method. (2) Evaluators prefer detailed evaluation methods, and hence, they supported those methods in which the example sentences are evaluated in multiple steps.

1. はじめに

現在，在日外国人数や訪日外国人数は増加傾向にあり[1),2)]，多言語によるコミュニケーションの機会も増加している．しかし，在日外国人や訪日外国人の中には，日本語を理解できない人が多数存在している[3)]．一般に多言語を十分に習得することは非常に難しく，母語以外の言語によるコミュニケーションは困難である[4)–6)]．このため，用例対訳や機械翻訳などの言語資源を組み合わせて利用できる仕組みである言語グリッドの活動が広がるなど[7),8)]，言語の壁を越える活動が活発化している．

日本語を理解できないことの影響が顕著に現れる分野の 1 つに医療がある．日本語が通じない外国人患者と日本人医療従事者間とのやりとりは，意思の疎通を十分に行うことができない．言葉が通じない環境では医療従事者が外国人患者の症状を十分に理解することが難しく，その結果として医療ミスが発生する恐れがある．現在，日本語を理解できない外国人の支援は医療通訳者が行っているが，医療通訳者は慢性的な人員不足となっている．また，通訳者の身分保障や通訳者自身のメンタルケアなどの問題が存在している[9)]．このため，医療現場を訪れるすべての外国人患者の支援を医療通訳者が行うことは難しい．医療通訳者が仲介しない場合，医療従事者が日本語を話すことができない外国人患者の症状を十分に理解することは難しく，医療ミスにつながる可能性がある．

情報技術を利用した医療分野の支援として，多言語医療受付支援システム M^3（エムキューブ）[10)] がある．M^3 は，正確な用例対訳を使用して医療受付での応対や問診の支援を行っている．用例対訳とは，用例を多言語に翻訳した多言語コーパスのことを指す．なお，本論文では医療分野などの正確性が求められる分野で使用する文章を「用例」，用例を多言語に翻訳してひとまとまりにして管理したものを「用例対訳」とする．

我々は Web 上での用例対訳の収集，共有，提供を目的とする多言語用例対訳共有システ

†1 和歌山大学大学院システム工学研究科
Graduate School of Systems Engineering, Wakayama University
†2 和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University

ム TackPad（タックパッド）の開発を行い，試用実験で有用性を確認した[11]．本システムの利用者は医療従事者や患者，医療通訳者などが参加しており，医療現場で使用する用例対訳の収集・共有を行っている．本システムでは利用者が用例対訳の作成を行うが，用例対訳の正確性は用例作成者に依存している．このため，用例対訳の正確性評価を行う必要がある．

また，本システムにおける用例対訳の作成は次の流れで行っている．

(1)　医療現場で使用される用例を本システムに提案する．

(2)　提案された用例を各言語に翻訳する．

このとき，(1) で提案された用例が医療現場では使用されないものであったり，不正確であったりした場合，(2) で用例が翻訳されても使用されない用例対訳や不正確な用例対訳が作成されるこのため，(1) の段階で用例作成者以外が正確性評価を行う必要があると考えられる．なお，本論文では (1) の段階の正確性を「用例の正確性」とする．

そこで本論文では，用例の正確性評価のための評価軸，評価段階が異なる 3 つの評価手法を用いて比較実験を行い，用例の正確性評価が可能な評価手法の要件抽出を行う．適用先は，要件抽出後の長期実験が可能である TackPad とする．なお，本論文で評価対象とする用例は，医療や司法など，正確性が求められる分野で使用するものを対象とする．

2. 関 連 研 究

円滑な多言語間コミュニケーション支援を目指して，機械翻訳を用いた支援技術の研究が行われている[12]．しかし，機械翻訳は正確性が求められる医療分野で利用可能な精度には達していない[13]．また，機械翻訳はルールや対訳データに従って計算機を使った動的な翻訳を行うため[14]，あらかじめすべての出力文の正確性を確保することはできない．

そこで現在，用例対訳による支援が行われている．用例対訳はあらかじめ正確性を評価することができる．また，紙媒体や電子媒体など様々な媒体で利用可能であるため，正確な多言語間コミュニケーションが求められる分野で使用されている．用例対訳を利用したシステムとして，“日本語でケアナビ” ⋆1がある．“日本語でケアナビ” は，介護に関する日本語と英語の用例対訳を約 8000 文提供している．しかし，これらの用例対訳の作成には多くの時間がかかっている[15]．

このため，用例対訳を Web 上で収集する取り組みが行われている．用例対訳の収集プロジェクトとして，Bond らの TATOEBA プロジェクトがある[16]．TATOEBA プロジェクトは田中コーパス[17] を基礎として，日本語，英語，フランス語，中国語，ドイツ語など様々な言語の用例対訳を収集している．また，Chen らは Web 上にある用例対訳を自動的に収集する試みを行っている[18]．

しかし，従来の用例対訳を収集する取り組みでは，個々の用例対訳を評価する仕組みを用意していない．このため，用例対訳の正確性の確保ができていないという問題点がある．正確性の確保が行われていない用例対訳は，高い正確性が必要な分野の支援に用いることができない．そこで本論文では，用例対訳の作成に必要な用例の正確性確保を行う評価手法を提案する．また，実験を通して用例の評価手法の要件抽出を行う．その際，用例登録者が Web 上でお互いに用例を評価する仕組みを用意することで，用例の正確性の確保を目指す．

Web 上での評価においては，様々なアプローチがなされている．Sen らは，映画推薦システムで付与されるタグを評価するために「賛成」「反対」の 2 値で評価している[19]．また，Yahoo!ニュース⋆2や newsing ⋆3では，利用者の評価をもとに注目されている記事をページの上に表示する取り組みが行われている．これらのニュースサイトにおける評価は，「おすすめ」「がっかり」の 2 値評価を提示したり,「びっくりした」「興味深い」などの評価項目を総合した点数を提示したりしている．しかし，これらの評価は記事そのものの正確性評価には利用されていない．

また，商品の評価機能の例として，Amazon ⋆4や価格.com ⋆5などがある．これらのサイトでは，販売している商品の評価や感想をその商品の購入者や利用者が行っている．実際に商品を購入，所有している人の評価や感想は，その商品の購入を思案している別の利用者にとって非常に説得力のある情報であり[20]，評価の有用性判別の研究も行われている[21]．

しかし，本システムで評価を行うべき対象は単語や文である．これらは医療従事者と患者の間のコミュニケーションに用いられるため，平易な言葉が使用されている．このため，医療分野に精通していない人でもほとんどの用例を評価することができる．それに対して，商品の評価は評価者が購入もしくは所有していないと評価することができない．これらのことから，用例の評価は商品の評価よりも 1 人あたりの評価可能対象物の数が多いため，評価者の負荷が高くなると考えられる．また，評価に関する評価者の負担は評価者数と評価対象物数のバランスも関係していると考えられる．今後，評価者数が評価対象用例数と比較して

⋆1 http://nihongodecarenavi.jp/

⋆2 http://headlines.yahoo.co.jp/hl

⋆3 http://newsing.jp/

⋆4 http://www.amazon.co.jp/

⋆5 http://kakaku.com/

多くなる可能性もあるが，現時点では用例数に対して評価者数が十分であるとはいえない．

商品の評価では段階評価と自由記述の組合せが多く用いられている．しかし，用例の評価にそのまま適用すると評価者の負荷が高くなり，十分な用例評価ができない可能性が高くなると考えられる．このため，本論文では記事や商品の評価ではなく用例の正確性評価を目的とした評価手法の確立を目指す．

なお，アンケート調査，社会調査の分野ではカテゴリ尺度や両極尺度，一対比較や自由回答など多くの評価手法がすでに提案されている[22),23)]．しかし，これらの評価手法は実際に使用したり体験したりした，過去に行われた内容に対して多く利用されている．これに対し，用例の評価は自分自身が使ったことがない用例も評価する必要がある．また，用例の評価対象の数は一般的なアンケートの項目数よりも多くなり，評価者の負荷も高いと考えられる．このため，次章の評価手法の設計では，通常のアンケートで使用されている評価手法の中から評価者の負荷が少ないものを選定し，利用することとする．

3. 評価手法の設計

本章では構築した評価手法と適用先である多言語用例対訳共有システムについて述べる．まず，3.1 節で本システムの概要を述べた後，3.2 節で評価手法の設計について述べる．

3.1 システム概要

本システムは，多言語用例対訳を収集するため画面インタフェースを多言語としている．本システムの画面例を**図 1** に示す．本システムは，日本語と実質の世界標準語である英語，在日外国人の上位 3 国の言語である中国語，韓国・朝鮮語，ポルトガル語の用例を収集している[1)]．また，利用者の要望をもとにスペイン語，ベトナム語，タイ語，インドネシア語も収集しており，現在 9 言語の用例対訳を収集している．なお，PHP と MySQL を使用して

図 1　TackPad の画面例
Fig. 1　Screenshot of TackPad.

Web 上での用例対訳の収集を可能としている．

本システムの主要機能は以下に示す 3 つである．

(1)　用例の提案
医療従事者や患者などが他の言語に翻訳してほしい用例を提案する機能である．実際に用例を使用する利用者がそれぞれの立場から提案するため，必要な用例を集めることができる．また，本機能は翻訳作業が不要なため，理解できる言語が 1 言語の利用者も用例対訳の収集，共有に貢献することが可能である．

(2)　対訳の作成
「用例の提案」で提案された用例を翻訳者が翻訳する機能である．医療分野では正確な翻訳が必要なため，本機能は翻訳者のみが利用する．

(3)　用例対訳の検索
本システム内の用例対訳を検索する機能である．本機能は，医療従事者や患者，翻訳者などすべての利用者が利用可能である．

3.2 評 価 機 能

本節では，構築した 3 つの用例評価機能の設計について述べる．

用例の評価機能は，評価の目的である正確性に関する情報を得る必要がある．また，2 章で述べたとおり，用例の評価において，評価者が理解可能な言語の用例はすべて評価が可能であり，評価者に対する負担が大きくなると考えられる．このため，用例の評価機能は下記の要件を満たす必要がある．

評価機能の要件 1　用例の正確性の確保に必要な情報が得られる．
評価機能の要件 2　評価者の負荷が少ない．

また，用例対訳の評価には，用例そのものが適切かどうかを判断する「用例の評価」と，対訳が用例の翻訳として適切かを判断する「用例対訳の評価」がある．本論文では，1 章でも述べたとおり，「用例の評価」に関する手法の確立を目指す．

なお，「用例の評価」では「評価機能の要件 1」の用例の正確性の確保に必要な情報として，下記の評価を行う必要があると考えられる．

評価条件 1　登録された用例は実際に用例を使用する現場で使用されるかどうかの評価
評価条件 2　用例の表現や言葉遣いは適切かどうかの評価

これらをふまえて，用例の評価を目的とした 3 つの評価手法を用意した．各評価手法の画面例を**図 2** に示す．本論文では，各評価手法をそれぞれ評価手法 A，評価手法 B，評価手法 C とする．今回用意した評価手法は「評価機能の要件 2」（評価者の負荷が少ない）を満

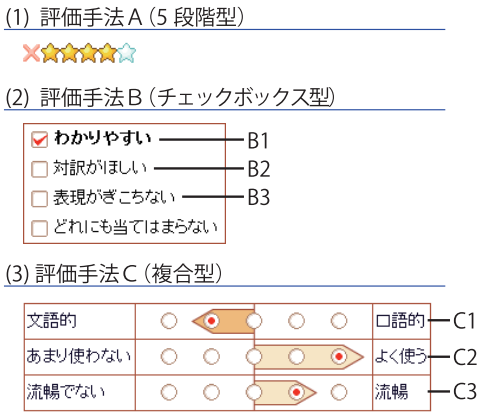

図 2　3 つの評価手法の画面例
Fig. 2　Screenshots of three evaluation methods.

表 1　評価手法と評価段階，評価軸の関係
Table 1　Relation among the evaluation methods, evaluation steps and evaluation axes.

		評価段階	
		2 段階	5 段階
評価軸	単一（未提示）	——	評価手法 A
	複数	評価手法 B	評価手法 C

表 2　評価手法ごとの特徴
Table 2　Features of the evaluation methods.

	評価手法 A	評価手法 B	評価手法 C
クリック回数	1 回	複数回	複数回
評価軸	単一（未提示）	複数	複数
評価の度合い	あり（5 段階）	なし（2 段階）	あり（5 段階）

たすために，自由記述などを設けず，1 回のクリックで評価を可能としている．ただし，各評価手法の評価回数，評価段階は異なっている．このため，各評価手法は 1 回のクリックで評価を可能としているが，評価手法間での負荷の差は存在している．

構築した評価手法と評価段階，評価軸の関係を**表 1** に，評価手法ごとの特徴を**表 2** にそれぞれ示す．評価手法 A は表 2 のクリック回数を見ると，表 1 で評価軸が複数となっている他の評価手法よりも少なく，評価者の負荷が少ないため有利である．しかし，評価手法 A は評価軸を提示していない．このため，前述した「評価機能の要件 1」（用例の正確性確保に必要な情報が得られる）を満たさない可能性がある．ただし，評価手法 A は Amazon や YouTube[⋆1] などの Web サイトで採用されているうえ，評価者の負荷も少ない．評価手法 A に関しては，次章の実験で「評価機能の要件 1」を満たすかどうかの確認を行う．なお，評価手法 B と評価手法 C に関しては，用例の正確性の確保に必要な情報を評価軸とすることで「評価機能の要件 1」をすでに満たしていると考えられる．また，評価手法 A と評価手法 C は，評価の度合いを選択可能な点が評価手法 B と異なっていることが分かる．

各評価手法の詳細を以下に示す．

(1)　評価手法 A（5 段階型）（図 2-(1)）
星をクリックすることによって評価が可能な評価手法である．この評価手法は，Amazon や YouTube などの Web サイトで採用されている．

(2)　評価手法 B（チェックボックス型）（図 2-(2)）
あらかじめ用意した文に対して同意できるかどうかを，チェックボックスで行う評価手法である．チェックボックスを使用することで，複数の項目にチェックすることも可能とした．評価手法 A では 1 つの評価軸のみで評価を行うが，評価手法 B では複数の評価軸による評価を可能とした．今回は，「分かりやすい」「対訳がほしい」「表現がぎこちない」の 3 つの項目を用意した．なお，本論文ではそれぞれ B1，B2，B3 とする．B2 は「評価条件 1」の評価を，B1 と B3 は「評価条件 2」の評価をそれぞれ目的としている．図 2-(2) では 4 つ目の項目として「どれにもあてはまらない」があるが，考察時には使用していない．

(3)　評価手法 C（複合型）（図 2-(3)）
あらかじめ用意した対義語の組に対して，5 段階評価を行う評価手法である．評価手法 B は評価の度合いを表現できないが，評価手法 C は可能とした．また，評価手法 A と同様の 5 段階評価だが，対義語による評価軸を提示している点が評価手法 A とは異なる．今回は，「文語的–口語的」「あまり使わない–よく使う」「流暢でない–流暢」を用意した．なお，本論文ではそれぞれ C1，C2，C3 とする．C2 は「評価条件 1」の評価を，C1 と C3 は「評価条件 2」の評価をそれぞれ目的としている．

なお，評価手法 B，評価手法 C ともに 1 つの評価条件に対して複数の評価軸をあてはめている．これは，次章の実験で正しく評価されているかどうかの確認を行うために行ってい

⋆1 http://www.youtube.com/

る．また，評価手法 A，評価手法 C の評価段階は 5 段階を採用した．これは，(1) 評価手法 C は対義語を配置し，どちらに近いかを評価するため中央値である「どちらともいえない」にあてはまる項目が必要，(2) 3 段階評価の場合，評価手法 B との十分な差異がない，(3) 7 段階評価以上の場合，評価者の負荷が増大する[24]，という理由から 5 段階評価としている．

4. 比較実験

本章では，評価手法の比較実験について述べる．本実験の目的は，評価軸と評価段階が異なる 3 つの評価手法のうち用例の正確性評価が可能な評価手法についての知見を得ることである．また，評価手法 A が「評価機能の要件 1」（用例の正確性確保に必要な情報が得られる）を満たすかどうかの確認も行う．

4.1 評価手法の比較実験

本節では，3.2 節で設計した評価手法の比較実験について述べる．本実験では，本システムに登録されていた 30 文の日本語の用例を用いた．これは，下記の条件を満たす用例である．

(1)　被験者が実験対象用例を作成していない．
(2)　被験者が実験対象用例をすでに評価していない．
(3)　被験者（患者）が使用する用例（医療従事者が使用する用例は用いない）．

このようにして用意した用例は，10 文を 1 組として 3 つの用例群（用例群 1，用例群 2，用例群 3）に分けた．被験者は 1 つの評価手法に対し，用例群のうち 1 つ（用例群 1，用例群 2，用例群 3 のいずれか）の 10 文の用例を評価した．被験者別の評価手法と用例群の関係を**表 3** に示す．表 3 のように，実験では同一の用例を複数回評価を行わないようにしている．被験者は実験全体で計 30 文の用例の評価を行った．なお，評価手法の評価順序，用例の出現順序はそれぞれ被験者ごとに順序交換を行っている．

表 3　評価者別の評価手法と用例群の関係
Table 3　Example allocated relation among the evaluation methods, example sentences groups and evaluators.

	評価手法 A	評価手法 B	評価手法 C
評価者 α	用例群 1	用例群 2	用例群 3
評価者 β	用例群 1	用例群 3	用例群 2
評価者 γ	用例群 2	用例群 1	用例群 3
評価者 δ	用例群 2	用例群 3	用例群 1
評価者 ϵ	用例群 3	用例群 1	用例群 2
評価者 ζ	用例群 3	用例群 2	用例群 1

・表中の用例群 1，用例群 2，用例群 3 は，それぞれ用例 10 文を 1 組とした用例群を指す．
・被験者は上記のうちどれかの組合せで用例の評価を行っている．

また，本実験ではすべての用例に対して必ず評価を付けるように依頼した．評価手法 B については 1 つの用例に対して 1 つ以上の項目にチェックを付けるように，評価手法 C については 3 つの評価軸すべてを評価するように依頼した．評価手法 B の評価項目がどれもあてはまらない場合,「どれにもあてはまらない」にチェックを付けるように依頼した．この項目は，被験者がすべての用例に対して評価を行ったかを確認するために用意している．また，実験後にアンケートへの回答を依頼した．

被験者は，本システムの利用者である和歌山大学システム工学部所属の大学生・大学院生 33 人（男性 17 人，女性 16 人）である．また，今回の被験者は患者の立場で用例の評価を依頼した．本システムの利用者は用例の登録者や翻訳者であるが，病気になった場合は用例を使用する患者になりうる．このため，評価用例はすべての本システム利用者が評価可能な患者の立場で使用可能なものとした．ただし，医療従事者の立場の用例が含まれていないため実験結果は限定的である可能性がある．

また，4.1 節の比較実験で同一の評価手法と用例群で評価した人数を**表 4** に示す．評価手法と用例群の組合せは 9 つあるが，すべての組合せで 10 人以上が評価している．なお，表 4 の人数が異なっているのは一部の被験者の評価ログ取得に失敗したためである．

4.2 インタフェース比較実験

4.1 節で行った実験では，評価手法 A と評価手法 C を同じ 5 段階評価という分類で行った．しかし，評価手法 A と評価手法 C の評価インタフェースが異なっていたため，評価インタフェースの違いが評価結果に影響を及ぼす可能性が考えられた．このため，4.1 節の実験後に評価インタフェースを統一した評価手法の比較実験を行った．

比較実験で用いた評価インタフェース例を**図 3** に示す．図 3 は，評価軸，評価段階は評価手法 A，インタフェースは評価手法 C を採用している．なお，図 3 を評価手法 A' とする．

この評価手法 A' を用いて，和歌山大学の学生 17 人に用例の評価を依頼した．なお，被

表 4　各評価手法の評価人数
Table 4　Number of evaluators in each evaluation method.

	用例群 1	用例群 2	用例群 3
評価手法 A	11 人	11 人	11 人
評価手法 B	10 人	12 人	11 人
評価手法 C	12 人	10 人	11 人

・表中の用例群 1，用例群 2，用例群 3 は，それぞれ用例 10 文を 1 組とした用例群を指す．

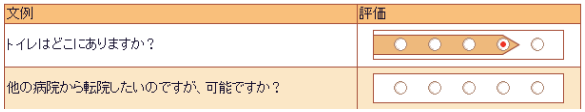

図 3　評価手法 A' の画面例
Fig. 3　Screenshot of evaluation method A'.

験者は 4.1 節で用例の評価を行った人に依頼している．また，評価対象用例は 4.1 節の実験時に被験者が評価手法 A で評価した 10 個の用例を用いた．通常の実験の場合，同一の被験者に同一の用例を短い期間で使用すると，1 回目の実験内容が 2 回目の実験内容に影響を及ぼす可能性がある．しかし，4.1 節の実験と本実験の間は 4 カ月以上離れているため，被験者が同一の用例を評価することについては影響が少ないと考えられる．

5. 実験結果と考察

5.1 評価手法の比較実験の結果と考察

本節では，4.1 節で述べた比較実験の結果と考察について述べる．まず，5.1.1 項で評価手法の相関について考察を行い，5.1.2 項で被験者が回答したアンケートから考察を行う．

5.1.1 評価手法の相関

本項では，各評価手法による用例の評価結果について考察する．考察するにあたり，用例ごとに各評価手法による評価結果の平均を求めた．その後，平均を用いて評価手法間で相関係数を求めた．各評価手法間の評価結果の相関を**表 5** に示す．なお，評価手法 A は 5 段階評価（1〜5）の平均，評価手法 B は（チェックされた数/評価者数），評価手法 C は左端の値を 1，右端を値を 5 に変換した 1〜5 の値の平均を用いている．本項では，表 5 の相関係数の結果から考察を行う．

評価手法 B と評価手法 C の相関関係

表 5 より，評価手法 B1（分かりやすい）と評価手法 C の各評価軸との間には正の相関関係があることが分かる．特に，評価手法 B1 と評価手法 C2（あまり使わない–よく使う），評価手法 C3（流暢でない–流暢）のように同傾向の評価軸は同様の評価が行われている．

また，評価手法 B3（表現がぎこちない）と評価手法 C の各評価軸との間には負の相関関係があることが分かる．これは，評価手法 C2 の「あまり使わない」，評価手法 C3「流暢でない」が評価手法 B3 の「表現がぎこちない」と同様の評価であることが示されている．

これらのことより，評価軸を用意した評価手法 B と評価手法 C は同傾向の評価が行われていると考えられる．

表 5　各評価手法間の相関係数
Table 5　Correlation coefficient among evaluation methods.

	A	B1	B2	B3	C1	C2	C3
A	1.00						
B1	−0.11	1.00					
B2	−0.21	*−0.50*	1.00				
B3	−0.05	*−0.86*	0.40	1.00			
C1	−0.21	**0.52**	−0.19	*−0.61*	1.00		
C2	−0.32	**0.69**	−0.30	*−0.57*	**0.61**	1.00	
C3	−0.19	**0.69**	−0.38	*−0.69*	**0.64**	**0.62**	1.00

・表中の A，B，C はそれぞれ評価手法 A（5 段階型），評価手法 B（チェックボックス型），評価手法 C（複合型）を指す．
・B1 は「分かりやすい」，B2 は「対訳がほしい」，B3 は「表現がぎこちない」が評価軸である．「どれにもあてはまらない」は省略している．
・C1 は「書き言葉–話し言葉」，C2 は「あまり使わない–よく使う」，C3 は「流暢でない–流暢」が評価軸である．
・相関係数が 0.5 以上のものを太字に，相関係数が −0.5 以下のものを斜体にしている．

なお，「評価条件 2」を用いて設計した評価手法 B1 と評価手法 B3 の間，評価手法 C1 と評価手法 C3 の間はそれぞれ相関関係がみられており，適切に評価されたと考えられる．しかし，「評価条件 1」を用いて設計した評価手法 B2 と評価手法 C2 の間には相関関係がみられなかった．アンケート結果から「どういうときに対訳が欲しいのかよく分からなかった」という意見が得られていたことから，評価手法 B2 の「対訳がほしい」という項目については再考する必要があると考えられる．

評価手法 A と他の評価手法の相関関係

表 5 より，評価手法 A は他の評価手法と相関関係がないことが分かる．特に，評価手法 A と評価手法 C2 との間には弱い負の相関がみられる．評価手法 A と評価手法 C2 のそれぞれで行った用例評価の平均値の分布を**図 4** に示す．図 4 より，評価手法 C2 で「あまり使わない」と評価された用例が，評価手法 A で高い評価になっている用例（図 4 左上）が存在する．また，逆に評価手法 C2 で「よく使う」と評価された用例が，評価手法 A で低い評価になっている用例（図 4 右下）も存在していることが分かる．

また，評価手法 A の平均と評価手法 C2 の平均が大きく異なった用例を**表 6** に示す．表 6 の 1 と 2 は，評価手法 C2 で「よく使う」と評価されたが評価手法 A の平均が低かったもの，表 6 の 3 と 4 は，評価手法 C2 で「あまり使わない」と評価されたが評価手法 A の平均が高かったものの例である．表 6 より，評価手法 C2 での評価結果の平均が高い値になっ

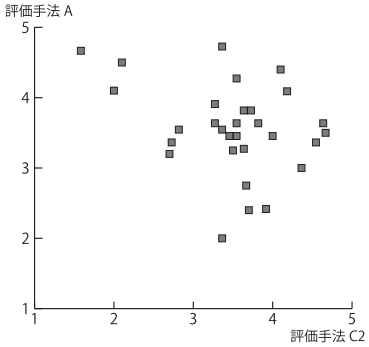

図 4　評価手法 A と評価手法 C2 での用例評価の平均値の分布
Fig. 4　Distribution of the average of example evaluation between evaluation methods A and C2.
・縦軸が評価手法 A での評価，横軸が評価手法 C2 での評価，各点が用例である．
・評価手法 C2 は左端の値を 1，右端の値を 5 としている．

表 6　評価手法 A と評価手法 C2 の評価が大きく異なった用例とその評価値
Table 6　Example sentences and their evaluation values for greatly different evaluations using evaluation methods A and C2.

	用例	A	C2	A-C2
1	治療費はどれくらいかかりますか？	2.82	4.67	−1.85
2	保険証をわすれました．	3.45	4.73	−1.27
3	寒くなると膝がしくしく痛みます	4.00	1.58	2.42
4	障害年金はうけられますか？	4.55	2.00	2.55

・評価手法 C2 の評価軸は「あまり使わない–よく使う」である．
・“A” “C2” の列は，それぞれ評価手法 A と評価手法 C2 の平均である．
・“A-C2” の列は，評価手法 A の平均から評価手法 C2 の平均を引いたものである．

た場合は病院でよく使うと思われる用例が，評価手法 C2 での評価結果の平均が低い値になった場合は病院での使用頻度が低いと思われる用例が選出されていることが分かる．このことより，評価手法 C2 ではほぼ適切に評価されているが，評価手法 A では逆の評価がなされていると考えられる．

用例の評価では前述のとおり「実際に使用するか」「表現や言葉遣いは適切か」の判断を行う必要がある．しかし，評価手法 A は用例の評価が可能な評価手法 B や評価手法 C の評価と相関関係がなく，逆の評価が行われた用例も存在する．このため，評価手法 A は「評価機能の要件 1」（用例の正確性確保に必要な情報が得られる）を満たさず，用例の評価手法として適切でないと考えられる．

表 7　評価の行いやすさに関するアンケート結果（5 段階評価）
Table 7　Questionnaire result of the usability level of evaluation (5-point Likert scale).

	評価手法 A	評価手法 B	評価手法 C
平均	3.73	3.09	3.52
標準偏差	0.88	1.07	0.94

・質問：「この評価のやり方は，文例の評価を行いやすかった」
・評価段階：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

表 8　評価の行いやすさに関するアンケート結果（順位付け）
Table 8　Questionnaire result of the usability level of evaluation (ranking).

順位	評価手法 A	評価手法 B	評価手法 C
1	15 人	7 人	11 人
2	9 人	8 人	16 人
3	9 人	18 人	6 人
合計	60	77	61

・3 つの評価手法の順位付けを被験者が行った結果である．
・表中の「合計」は評価手法ごとに「順位×人数」を合計した結果である．

評価手法 A と評価手法 B，評価手法 C の間の相関関係がないことの原因として，評価手法 A は評価軸を提示していないため，評価者それぞれが自分で評価軸を決めている可能性が考えられる．このため，用例の正確性評価には評価手法 A のような曖昧さを含む評価手法ではなく，評価手法 B や評価手法 C のように評価軸を明確に示した評価手法を用いる必要があると考えられる．

5.1.2　被験者による評価

本項では，被験者が回答したアンケート結果から考察を行う．

被験者が各評価手法の評価の行いやすさを 5 段階で評価した結果を**表 7** に，3 つの評価手法を評価しやすかった順に順位付けした結果を**表 8** にそれぞれ示す．なお，表 7 の平均は値が大きいものが，表 8 の合計は値が小さいものが，それぞれ支持されていることを示す．

表 7，表 8 より，評価手法 A（5 段階型）と評価手法 C（複合型）が，評価手法 B（チェックボックス型）より比較的評価しやすい手法であることが分かる．ただし，表 7 の結果を用いてフリードマン検定を行った結果，有意確率は 0.066 となり有意差はみられなかった．

アンケートの自由記述欄から得られた評価手法 A に関する好意的な意見として，「入力の負荷が少ない」「単純な操作で，率直な評価ができた」などが得られた．しかし，「自分の評価基準を決めるためやりにくかった」「基準があいまいな分，信憑性に欠ける」など評価基

表 9　評価手法 A’ と他の評価手法間の相関係数
Table 9　Correlation coefficient among evaluation method A’ and the other methods.

	A	B1	B2	B3	C1	C2	C3
A’	0.08	0.32	0.18	0.07	0.30	−0.11	0.11

・表中の A，B，C はそれぞれ評価手法 A（5 段階型），評価手法 B（チェック型），評価手法 C（複合型）を指す．
・表中の A’ は，評価手法 A’（評価軸が評価手法 A，インタフェースが評価手法 C）を指す．
・B1 は「分かりやすい」，B2 は「対訳がほしい」，B3 は「表現がぎこちない」が評価軸である．「どれにもあてはまらない」は省略している．
・C1 は「書き言葉–話し言葉」，C2 は「あまり使わない–よく使う」，C3 は「流暢でない–流暢」が評価軸である．

準を指定していないことによる曖昧さを問題視する意見もみられた．

アンケートの自由記述欄から得られた評価手法 C に関する好意的な意見として，「詳細に評価できるので微妙なニュアンスが伝わりやすい」「軸や基準が決められているので評価しやすかった」などが得られた．しかし，「話し言葉と書き言葉の判別はむずかしい」，また「流暢かどうかの判定が分かりにくかった」という意見も得られたため，評価に使用する対義語を再考する必要があると考えられる．

評価手法 B に関しては，「設問のまま素直に回答できる」という好意的な意見もみられたが，「あてはまるかあてはまらないかだけで，微妙な判断ができない」という評価の度合いを選択できないことを問題視する意見が得られた．このため，評価者の意図を汲み取るために，2 値の評価ではなく評価の度合いが選択可能な評価手法が必要であると考えられる．

また，評価手法 B に関して「もう少し選択肢を増やしたほうがいい」という意見が多くみられた．これは，評価手法 B では自分の評価を適切に表現しきれなかったことが理由に考えられる．しかし，評価項目を増やすと評価するために必要な時間が増え，評価者の負荷も増えるため，評価項目を増やすことは難しいと考えられる．評価手法 A や評価手法 C ではこのような意見がみられなかったことから，用例の評価においては，評価の度合いを選択可能にする必要があると考えられる．

5.2　インタフェース比較実験の結果と考察

本節では，4.2 節の実験結果と考察について述べる．

4.2 節の実験結果から，5.1.1 項の評価手法の相関と同様に相関係数を求めた．評価手法 A’ と他の評価手法との相関係数を**表 9** に示す．表 9 より，一部に弱い相関がみられるものが存在しているが，評価手法 A’ と他の評価手法との間に強い相関関係がみられなかった．また，単一の評価軸で 5 段階の評価段階である評価手法 A と評価手法 A’ との間にも相関関係はみられなかった．これは，評価手法 A，評価手法 A’ ともに評価軸を提示していないことが原因であると考えられる．このため，評価手法 A のインタフェースを変更した評価手法 A’ も，用例の正確性評価には不向きであると考えられる．

6. おわりに

本論文では，用例の正確性評価を目的として評価軸，評価段階が異なる 3 つの評価手法を用いて比較実験を行い，用例の評価手法の要件抽出を行った．

本実験の結果として，以下の知見を得た．

(1)　用例の正確性評価実験の結果，評価軸を提示していない評価手法と提示した評価手法の間に相関関係がみられなかった．用例の正確性評価には，評価軸を明確に提示した評価手法が必要である．

(2)　用例の評価において，評価者は詳細な評価を行うことを好んだ．このため，2 値の評価段階よりも複数の評価段階を選択可能とした評価手法が評価者から支持された．

以上の結果より，用例の正確性評価を行う評価手法は「複数の評価段階」かつ「評価軸を明確」にする必要があると考えられる．

本論文では評価軸の数や評価項目を十分に検証していない．このため，今後，評価項目の選定作業を行い，評価軸の数も含めた用例の評価手法の確立を目指す．また，評価軸の明確化や複数の評価段階を利用可能にすることにより，評価軸や評価段階が存在しない評価手法よりも評価者の負荷が増加する可能性がある．このため，今後結論に基づいた評価手法を本システムに実装し，評価者の負担に関する長期実験を行う．

謝辞　本研究の一部は，戦略的情報通信研究開発推進制度（SCOPE）および科研費基盤研究（B）（19300036），基盤研究（B）（22300044）の助成を受けたものである．

参 考 文 献

1) 法務省:平成 21 年末現在における外国人登録者統計について，法務省（オンライン），入手先〈http://www.moj.go.jp/nyuukokukanri/kouhou/nyuukokukanri04_00005.html〉（参照 2010-07-20）．
2) 法務省：平成 21 年における外国人入国者数及び日本人出国者数について（確定版），法務省（オンライン），入手先〈http://www.moj.go.jp/nyuukokukanri/kouhou/press_100312-2.html〉（参照 2010-07-20）．
3) 田村太郎：多民族共生社会ニッポンとボランティア活動，明石書店 (2000).
4) Takano, Y. and Noda, A.: A temporary decline of thinking ability during foreign language processing, *Journal of Cross-Cultural Psychology*, Vol.24, pp.445–462 (1993).

5) Aiken, M., Hwang, C., Paolillo, J., et al.: A group decision support system for the Asian Pacific rim, *Journal of International Information Management*, Vol.3, pp.1–13 (1994).
6) Kim, K.J. and Bonk, C.J.: Cross-Cultural Comparisons of Online Collaboration, *Journal of Computer Mediated Communication*, Vol.8, No.1 (2002).
7) Ishida, T.: Language Grid: An Infrastructure for Intercultural Collaboration, *IEEE/IPSJ Symposium on Applications and the Internet* (*SAINT-06*), pp.96–100 (2006).
8) Sakai, S., Gotou, M., Tanaka, M., et al.: Language Grid Association: Action Research on Supporting the Multicultural Society, *International Conference on Informatics Education and Research for Knowledge-Circulating Society* (*ICKS-08*) (2008).
9) 高嶋愛里：在日外国人支援活動：京都における「医療通訳システムモデル事業」，国際保健支援会 2 (2005).
10) 宮部真衣，吉野　孝，重野亜久里：外国人患者のための用例対訳を用いた多言語医療受付支援システムの構築，電子情報通信学会論文誌，Vol.J92-D, No.6, pp.708–718 (2009).
11) 福島　拓，宮部真衣，吉野　孝，重野亜久里：医療分野を対象とした多言語用例対訳収集 Web システム TackPad の開発，マルチメディア，分散，協調とモバイル（DICOMO2008）シンポジウム，pp.1030–1036 (2008).
12) Yoshino, T., Fujii, K. and Shigenobu, T.: Availability of Web Information for Intercultural Communication, *Pacific Rim International Conference on Artificial Intelligence* (*PRICAI-08*), pp.923–932 (2008).
13) 林田尚子，石田　亨：翻訳エージェントによる自己主導型リペア支援の性能予測，電子情報通信学会論文誌，Vol.J88-D1, No.9, pp.1459–1466 (2005).
14) 塚田　元，渡辺太郎，鈴木　潤，永田昌明，磯崎秀樹：統計的機械翻訳，NTT 技術ジャーナル，Vol.19, No.6, pp23–25 (2007).
15) 上田和子，ジョイ・デヴェラ，水野真木子，角南北斗，原田マリアフェ：『日本語でケアナビ』と実践的コミュニティー，国際交流基金関西国際センター日本語教育シンポジウム（2008 年 3 月 8 日），パネルディスカッション資料，泉南郡田尻町 (2008).
16) Bond, F., Nichols, E., Appling, D.S., et al.: Improving Statistical Machine Translation by Paraphrasing the Training Data, *Proc. IWSLT 2008*, pp.150–157 (2008).
17) Tanaka, Y.: Compilation of a multilingual parallel corpus, *Proc. PACLING 2001*, pp.265–268 (2001).
18) Chen, J., Chau, R. and Yeh, C.H.: Discovering Parallel Text from the World Wide Web, *ACSW Frontiers '04: Proc. 2nd Workshop on Australasian Information Security, Data Mining and Web Intelligence and Software Internationalisation*, Vol.32, pp.157–161 (2004).
19) Sen, S., Harper, F.M., LaPitz, A. and Riedl, J.: The Quest for Quality Tags, *Proc. 2007 International ACM Conference on Conference on Supporting Group Work* (*GROUP '07*), pp.361–370 (2007).
20) 渡辺弘美：ウェブを変える 10 の破壊的トレンド，ソフトバンククリエイティブ (2007).
21) 山澤美由起，吉村宏樹，増市　博：Amazon レビュー文の有用性判別実験，情報処理学会研究報告，自然言語処理研究会，2006-NL-173-(3), pp.15–20 (2006).
22) 酒井　隆：図解アンケート調査と統計解析がわかる本，日本能率協会マネジメントセンター (2003).
23) 内田　治，醍醐朝美：実践アンケート調査入門，日本経済新聞社 (2001).
24) シストラットコーポレーション：戦略理論・調査手法＞ワーディング 1 つ 1 つに理由があります，シストラットコーポレーション（オンライン），入手先〈http://www.systrat.co.jp/theory/reading/QNTY01wording.html〉（参照 2010-07-20）.

(平成 22 年 4 月 19 日受付)
(平成 22 年 10 月 4 日採録)

福島　拓（学生会員）
1986 年生．2008 年和歌山大学システム工学部デザイン情報学科中退．2010 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．現在，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在学中．多言語間コミュニケーション支援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）．コラボレーション支援の研究に従事．