情報処理学会論文誌　データベース　Vol.5　No.2　1–10 (June 2012)

多対多の関係性を持つ多言語用例対訳グラフにおけるメタノード作成手法

福島 拓[1,a)]　吉野 孝[2,b)]

受付日 2011年12月16日, 採録日 2012年4月7日

概要：世界的なグローバル化を背景に，我々は多言語間コミュニケーション支援を目的とした多言語用例対訳共有システムの開発を行っている．共有対象の用例対訳は，各言語間で 1 対 1 に対応する必要があるが，用例対訳には各言語間で 1 対 1 に対応しない言葉の組合せが存在している．このような多言語間の言葉の多様性への対応が求められているが，単純な用例間の意味のつながり情報のみでは対応することができていなかった．そこで，本論文では，1 対多，多対多の関係にある用例対訳グラフを提供可能な用例対訳とするために，メタノード作成手法を提案する．また，メタノード作成手法適用により顕在化したリンク不足の問題を解決するために，新たな用例間リンクを発見する方法についても提案する．

キーワード：用例対訳，メタノード，多対多関係

Meta-node Composition Method on Multilingual Parallel-text Graph Having Many-to-many Relationship

Taku Fukushima[1,a)]　Takashi Yoshino[2,b)]

Received: December 16, 2011, Accepted: April 7, 2012

Abstract: Recently, there is increasingly globalized world. We developed a multilingual parallel-text sharing system for multilingual communication. A parallel-text requires one-on-one combination among each language. However, parallel texts often have example sentences of many-to-many combination among each language. A multiplicity of expression causes this problem. It is difficult to solve this problem only information of association between example sentences. Therefore, we proposed a meta-node composition method for parallel-text graph having many-to-many combination. Moreover, we proposed a new parallel-text link discovery method to solve latent problems.

Keywords: parallel text, meta node, many-to-many relation

1. はじめに

近年の世界的なグローバル化により多言語間コミュニケーションの機会が増加している．しかし，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なこともあり [1], [2], [3]，日本語を理解できない外国人と日本人とのコミュニケーションは十分に行うことができない．このため，形態素解析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 [4], [5]，言語の壁を越える活動が活発化している．

情報技術を利用した正確性が求められる分野の多言語支援として，正確性の確保が可能な用例対訳が多く用いられている．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．用例対訳は言語間の用例の変換を目的としているため，各言語に一意に変換可能であ

[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, Wakayama 640–8510, 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)] fukushima@yoslab.net
[b)] yoshino@sys.wakayama-u.ac.jp

るという特徴がある．例として，「耳鳴りがします．（日本語）」「My ears are ringing.（英語）」「Sentindo um tinido nos ouvidos.（ポルトガル語）」という用例対訳をあげる．このような用例対訳を利用すると，日本語しか理解できない人でも「耳鳴りがします.」と翻訳先言語を選択するだけで，正確な英語やポルトガル語の言葉を利用することができる．用例対訳を利用した医療分野のシステムとして，多言語医療受付支援システム M^3 [6] や，多言語問診票作成システム [7] がある．M^3 は，用例対訳を用いて医療受付での応対や問診の支援を行っている．また，多言語問診票作成システムは，用例対訳や機械翻訳を用いて医療受付で記入を行う問診票の多言語化を可能としている．また，防災分野でも用例対訳による外国人支援が行われている [8]．

用例対訳は前述のとおり，ある言語の用例を他の言語の用例に一意に変換可能であるという特徴がある．しかし，多言語の言葉には各言語間で 1 対 1 に対応せず，1 対多，多対多の組合せとなるものが存在している．また，使用する相手やニュアンスの違いによって同じ意味の複数の言葉が存在する場合も多い．このため，用例対訳も 1 対多，多対多の関係になりうるものが存在している．**1 対多，多対多の関係にある用例対訳はグラフデータ構造で表すことができるが，用例対訳の特徴である多言語間の言葉の変換を一意に行うことができない．**例としては，「Yes（英語）」と「はい（日本語）」「ある（日本語）」という用例対訳が 1 対多の関係となっている．この用例対訳の場合，「Do you smoke? – たばこは吸いますか？」という質問には「Yes – はい」という回答を，「Are you currently taking medication? – 現在服用している薬はありますか？」という質問の場合は「Yes – ある」という回答をそれぞれ選択する必要があり，一意に変換を行うことができない．このような一意変換できない用例対訳は，多言語対話システムで使用することができない．既存システムでの多対多の関係にある用例対訳は，便宜的に 1 対 1 の別の用例対訳として扱うことが多いが，用例が重複するため管理が煩雑となっている．このように，多対多の関係にある用例対訳は 1 対 1 の関係にある用例対訳と比べて少ない存在ではあるが，用例対訳の利用において大きな問題を引き起こしている．

そこで本論文では，1 対多，多対多の関係にある用例対訳グラフを提供可能な用例対訳とするために，メタデータを利用したデータ構造を提案する．なお，用例対訳グラフとは，用例対訳中の用例や用例間の関係リンクをグラフ構造として表したものを指す．また，本論文で使用するメタデータを「メタノード」とする．メタノードは，同じ意味の用例を関連づける役割を担う．その際，1 対多，多対多の関係にある用例対訳を 1 対 1 の関係にすることで，一意に変換可能な用例対訳を実現する．また，複雑化した用例対訳グラフは，各用例間の関係リンクが不足している場合も発見が難しいという問題がある．このため，用例対訳グラフとメタノードを利用し，不足している各用例間の関係リンク発見手法について検討を行う．

2. 関連研究

用例対訳の需要増加を背景に，用例対訳の収集が行われている．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 [9]．また，Bond らは田中コーパス [10] を基に，用例対訳の収集プロジェクトを行っている [11]．このプロジェクトは TATOEBA プロジェクトという名前で活動が行われており，日常的に使用する用例の収集を，日本語，英語，フランス語，中国語，ドイツ語など様々な言語で行っている．Utiyama らも「みんなの翻訳」というプロジェクトで様々な文書翻訳を Web 上で行っており，合わせて翻訳支援のためのツールの提供も行っている [12]．我々も正確性が求められる医療分野の用例対訳の収集，共有を目的とした，多言語用例対訳共有システム TackPad の開発を行っている [13]．TackPad では，(i) 医療従事者や患者などが必要な用例をシステムに登録，(ii) 翻訳者が (i) で登録された用例を各言語に翻訳，の手順で，医療現場で求められている用例対訳の収集を Web 上で行っている．収集した用例対訳は他の多言語対応の医療支援システムに提供を行うことで，医療分野における多言語間コミュニケーションの支援を目指したシステムである．なお，多言語対応システムとは，用例対訳を使用した多言語支援システムを指す．このように用例対訳の収集は多く行われているが，これらの研究では多言語間に存在する 1 対多や多対多の関係について考慮していない．本論文では，多言語用例対訳における 1 対多や多対多の関係を考慮したメタノード作成手法を提案する．

メタデータの概念を言葉に適用したものとしてシソーラスがある．梶らはコーパスからのシソーラスの自動生成を行っている [14]．また，李は名詞概念と動詞概念をリンクで結合したハイパー・シソーラスを提案している [15]．福井らはシソーラスを用いた多言語翻訳を提案している [16]．これらは，シソーラスの特徴である上位概念と下位概念を結んだ木構造を基本として研究が行われている．しかし，本論文で扱う多言語用例対訳は用例間の意味のつながり，つまりグラフデータ構造が基本となっている．このため，本論文のメタノード作成手法はグラフデータ構造への適用を行うため，シソーラスとは本質的に異なる．

用例対訳以外では，グラフデータ構造のグループ化が行われている．小島らはハイパーリンクを用いた Web ページ群のグループ化を行っている [17]．また，岡田らはソーシャルネットワークや Web への適用を目的としたコミュニティ発見手法を提案している [18]．我々もこれらの研究と同様に，グラフデータ構造である用例対訳の分類を行う．その際，メタノードを用いて用例対訳を分類する．さらに，メタノードを用いた用例対訳の提供時における問題につい

ても解決を目指す．

3. 多言語用例対訳グラフにおけるメタノード作成手法

本論文で提案するメタノード作成手法は，1 対多，多対多の関係性を持つ用例対訳グラフを，多言語対応システムで利用可能にするために用いる．まず，3.1 節で多言語用例対訳における 1 対多，多対多の問題について述べた後，3.2 節で本手法と，作成したメタノードの提供方法について述べる．さらに，3.3 節で不足している用例間関係リンク発見手法について述べる．なお，本手法は多言語用例対訳共有システム TackPad [13] に適用する．

3.1 多言語用例対訳の 1 対多，多対多における問題点

用例対訳の例を**図 1** に示す．用例対訳は，(1) 元となる用例を準備する，(2) 用意された (1) の用例を各言語に翻訳する，という流れで作成することが一般的である．しかし，(2) で作成された用例間で意味が同じであるという担保が行われない．

図 1 を例に説明する．図 1 は元となる用例が日本語で，英語，中国語，韓国朝鮮語の各用例が作成された例である．このとき，意味が同じであると担保されているのは「日本語–英語」「日本語–中国語」「日本語–韓国朝鮮語」の用例間のみであり，「英語–中国語」「英語–韓国朝鮮語」「中国語–韓国朝鮮語」の用例間は意味が同じとは限らない．このため，先ほどの手順の後に，「(2) で作成された翻訳用例間で意味が同じであることを確認する」という手順が必要であると考えられる．このことを明確にするために，我々が開発している TackPad では意味が同じである用例間にリンクの作成を行っている．本論文では，このリンクを「用例間リンク」とする．**用例間リンクがつけられた用例は意味が同じであるため，用例対訳での多言語間の言葉の変換で使用することができる．**しかし，実際の用例対訳は 1 対多や多対多の関係を持つものが存在している．**多対多の関係性を持つ用例対訳は，一意に言葉の変換を行うことができない．**多対多の関係を持つ用例対訳の例を**図 2** に示す．図 2 中の二重線は，用例間リンクを示す．なお，図 2-(1) と図 2-(7) の間に用例間リンクを作成することが可能であるが，以降の説明の都合上，用例間リンクがない例を使用している．

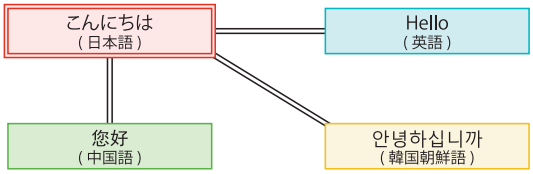


図 1　用例対訳の例
Fig. 1　Example of parallel texts.

図 2-(5) の韓国朝鮮語はどの時間帯でも使用できる挨拶である．このため，図 2 では各言語の朝の挨拶や昼の挨拶である 7 つの用例と用例間リンクでつながっている．図 2 の用例対訳の場合，図 2-(5) の韓国朝鮮語とリンクしている用例が言語ごとに一意でない（英語の 3 用例，日本語と中国語の 2 用例ずつとつながっている）ため，簡単には多言語間の言葉の変換を行うことができない．

簡単に多言語変換ができないことは，用例対訳を多言語対応のシステムで使用する際に問題となる．例として多言語対話システムをあげる．多言語対話システムは対話のフローをあらかじめ用意している場合が多い．対話フローの作成には，質問と回答の対を用意する必要がある．質問と回答の対の例として，「この病院は初めてですか？」という質問と「はい」「いいえ」の回答の組合せがある．この用例を元に，他の言語へ変換することで多言語対応を行っている．このとき，日本語に対して同じ言語の複数の用例がリンクされている場合，その用例対訳は言葉の一意変換ができない．結果的に，1 対多，多対多の関係を持つ用例対訳は多言語対話システムで使用できない．

そこで本論文では，「メタノード」を用意することで，1 対多，多対多の関係を持つ用例対訳を一意に変換可能な 1 対 1 の用例対訳への変換を行う．

3.2 メタノードの作成手法

本節では，前節で述べた問題の解決に用いるメタノードの作成手法について説明する．

本論文で作成するメタノードは次の特徴を持つ．

- メタノードは同じ意味の多言語用例を含む．
- 1 つのメタノードには同一言語の用例は複数含まない．

このことから，本論文では用例対訳グラフの中でクリーク（すべてのノードが相互にリンクされている部分グラフ）となっているものをメタノードとする．なお，他のクリークに包含されるクリークはメタノードとはしないこととする．また，メタノードは，利用者が作成する用例間リンクを元に動的に生成するものとする．このことで，用例対訳の作成者には従来から行っている「用例の作成」と「用例間リンクの作成」以外の作業の負担を求めない形で，提供

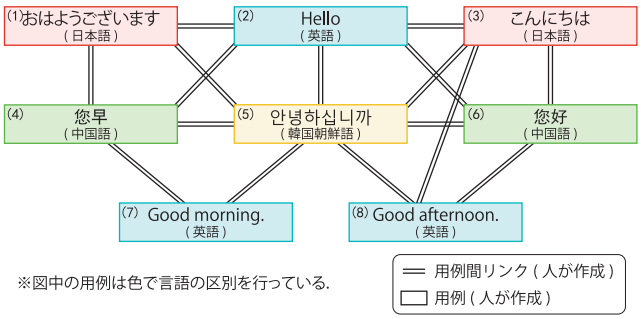


図 2　用例対訳の用例間リンク例
Fig. 2　Example of parallel-text links.

可能な用例対訳を作成する．なお，一般的には 3 つ以上のノードを保持するものをクリークと呼ぶ場合が多いが，本論文では 2 つ以上のノードを保持するものをクリークとし，本論文で提案するメタノード作成に用いる．これは，用例対訳が 2 つ以上の用例から構成されるためである．

次に，メタノード作成手法のアルゴリズムについて述べる．本手法は利用者によって用例間リンクが作成された際に適用を行う．以下に，用例 α と用例 β の間に用例間リンクが作成されたときの，メタノード作成手法のアルゴリズムを示す．なお，最悪計算時間は $O(n^2)$ である．

(1) 用例 α と用例 β が元々持っている，用例間リンク群をそれぞれ抽出し，重複していた用例を選ぶ．本論文では重複していた用例群を重複用例群とする．
(2) 重複用例が 1 つもない場合は，(7)-(a) の処理を行う．
(3) 用例 α，用例 β からつながっている，メタノード群をそれぞれ抽出する．
(4) 用例 α の各メタノードに含まれる用例すべてが，用例 α，もしくは重複用例群に含まれる用例であった場合，走査していたメタノードに用例 β を追加する．
(5) 用例 β のメタノード群についても，(4) と同様の操作を行う．
(6) (4) または (5) でメタノードへ用例の追加が行われていた場合，下記の操作を行う．
  (a) メタノードどうしで含有関係が発生した場合，それらのメタノードを統合する．
  (b)「メタノード」に追加された用例を，「重複用例群」から削除する．
(7) この時点での「重複用例群」の数に合わせて，下記の操作を行う．
  (a) 重複用例群の数が 0 個，かつ，(4) または (5) でメタノードへの用例の追加が行われていない場合，用例 α と用例 β を新規のメタノードに追加する．
  (b) 重複用例群の数が 1 個の場合，用例 α，用例 β と重複用例の用例を新規のメタノードに追加する．
  (c) 重複用例群の数が 2 個以上の場合，重複用例群の間で用例間リンクが存在しているかを調べる．存在している場合，用例 α，用例 β と重複用例群の用例を新規のメタノードに追加する．存在していない場合，用例 α，用例 β と重複用例群の用例それぞれを新規のメタノードに追加する．

図 2 に上記のアルゴリズムを適用してメタノードを付与した例を**図 3** に示す．図 3 の黒丸が作成したメタノードである．メタノード ⟨i⟩，⟨ii⟩，⟨iv⟩ は，完全グラフとなっていた 4 つの用例を，メタノード ⟨iii⟩ は，完全グラフとなっていた 3 つの用例をそれぞれ含んでいる．

このようにして作成されたメタノードは，多言語対応システムに用例対訳として提供するときに使用する．たとえば，図 3 のメタノード ⟨i⟩ を利用すると，朝の挨拶の用例対訳が，図 3 のメタノード ⟨iv⟩ を利用すると，昼の挨拶の用例対訳がそれぞれ利用できる．

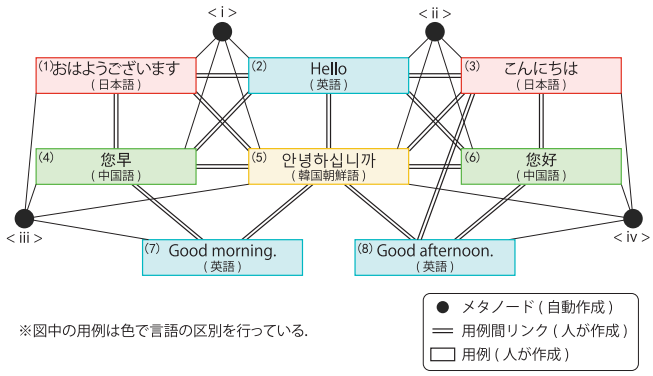


図 3　用例対訳のメタノードの作成例
Fig. 3　Example of meta nodes for parallel texts.

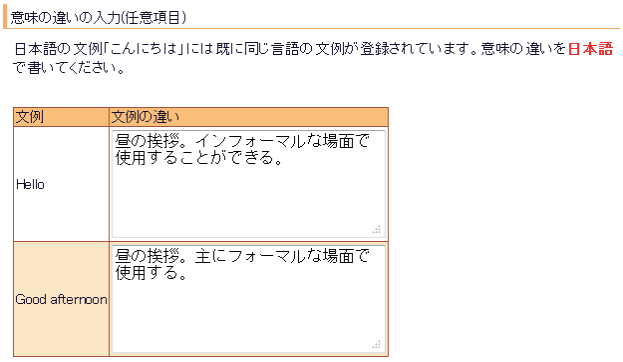


図 4　メタノードに付与する意味の入力画面例
Fig. 4　Screenshot of giving meaning to meta node function.

しかし，日本語しか分からない利用者は，メタノード ⟨ii⟩ とメタノード ⟨iv⟩ のどちらの「こんにちは」を利用すべきか判断することが難しい．このため，運用上の解決策として，メタノードに意味を持たせることを考える．本手法の適用先である TackPad では，1 つの用例に対して，同じ言語の複数の用例が付与されたときに，意味の違いの入力を促している．画面例を**図 4** に示す．図 4 は，日本語と英語を理解できる翻訳者が，すでに「Hello（図 3-(2)）」が付与されている用例「こんにちは（図 3-(3)）」に対して，「Good afternoon（図 3-(8)）」を付与したときの例である．この状態では，「Hello」と「Good afternoon」の違いが分からない日本語話者はどちらを選択すべきかの判断が難しい．しかし，図 4 のように意味の違いを日本語で記述することで，日本語話者は使用場面に合わせて用例対訳の選択を行うことができる．また，意味の違いをそれぞれのメタノードに付与することで，各言語でメタノードに意味を持たせることができる．このことで，利用者はメタノードの意味を見て，どのメタノード（用例対訳）を使用するか判断することができる．ただし，ほぼ同義の文のために違いの説明が難しい文が存在しているため，本システムでは任意入力の項目としている．

3.3　不足用例間リンクの発見手法

1 対多，多対多の関係にある用例対訳においては，関係が複雑になっているため用例間リンクが不足している場合も発見することが難しい．しかし，本論文で提案しているメタノード作成手法ではクリークを基にメタノード作成しているため，1 つでもリンクが不足している場合は適切なメタノードを形成できない．このため，本節では不足している用例間リンクの発見手法について述べる．本手法で発見した不足用例間リンクは，用例対訳作成システムの利用者に提示する．このようにすることで，新たな用例間リンクの作成補助ができると考えられる．

本手法の手順を以下に示す．

(1) ある用例 A について着目する．
(2) 次の条件すべてに合うメタノード α が存在した場合，「用例 A と用例間リンクを介してつながれている用例」と「メタノード α に属している用例」で，共通している用例の数を調べる．
  (a) メタノード α には用例 A と用例間リンクで結合された用例の一部（もしくはすべて）が含まれる．
  (b) メタノード α には用例 A は属していない．
  (c) メタノード α に用例 A と同じ言語の用例が属していない．
(3)「(メタノード α と用例 A に共通の用例数)/(メタノード α に属する用例数)」を求める．本論文ではこの値を「メタノードへの結合可能性」とする．メタノードへの結合可能性が大きいほど用例 A はメタノード α に属する可能性が高くなる．

上記の内容を図 3 を用いて説明する．図 3-(1) は用例間リンクにより用例 (2)，(4)，(5) とつながっている．用例 (2)，(4)，(5) が属していて用例 (1) が属していないメタノードは，⟨ii⟩，⟨iii⟩，⟨iv⟩ である．しかし，メタノード ⟨ii⟩，⟨iv⟩ には，用例 (1) と同じ言語である日本語の用例がすでに属しているため，用例 (1) は属することができない．このため，メタノード ⟨iii⟩ について確認する．「用例 (1) と用例間リンクを介してつながれている用例」と，「メタノード ⟨iii⟩ に属している用例」とで共通している用例は用例 (4) と (5) である．これらのことから，用例 (1) のメタノード ⟨iii⟩ への結合可能性は $(2/3) = 67\%$ であると求められる．メタノードへの結合可能性が高いものが不足している用例間リンクである可能性が高いため，効率的に利用者に不足用例間リンクの確認を依頼することができる．今回の例の場合，用例 (1) と用例 (7) に用例間リンクの作成される可能性が利用者に提示されることとなる．

なお，本手法で 0%より大きい値が示された場合，それらの用例は結合の可能性がある．このため，利用者に用例間リンクの作成可能性について提示することとする．ただし，値が低い場合は結合の可能性が低いため，値が高いものを優先的に利用者に提示する必要があると考えられる．

4.　多言語用例対訳への適用

本章では，メタノード作成手法を多言語用例対訳に適用する．本論文では，メタノード作成手法を次に示す用例対訳に適用した．

(1) 多言語用例対訳共有システム TackPad [13] で収集済みの用例対訳
(2) 多言語問診票作成システム [7] で使用している用例対訳

TackPad の収集言語は日本語，英語，中国語，韓国朝鮮語，ポルトガル語，スペイン語，ベトナム語，タイ語，インドネシア語の 9 言語である．また，収集済みの用例数は 14,487 件，用例間リンク数は 18,285 件であった．多言語問診票作成システムで使用している用例対訳は，日本語，英語，中国語，韓国朝鮮語，ポルトガル語，ベトナム語の 6 言語である．また，用例数は 2,480 件，用例間リンク数は 5,278 件であった．

用例対訳ごとの言語別の用例数を**表 1** に示す．TackPad の用例対訳は，日本語が他の言語の 2 倍以上収集されていることが分かる．これは，TackPad の利用者は日本語話者が多いことが理由として考えられる．また，多言語問診票作成システムの用例対訳はベトナム語を除いてほぼ同数の用例数であることが分かる．なお，TackPad の約 1 万件の用例対訳と，多言語問診票作成システムのすべての用例対訳はあらかじめ用意したものを直接データベースに挿入している．

4.1　メタノード作成手法の適用結果と考察

本節では，2 つの多言語用例対訳にメタノード作成手法を適用した結果とその考察について述べる．

4.1.1　メタノード作成結果

メタノードに結合された用例数を**表 2** に示す．TackPad の用例対訳は，2 つの用例が結合した用例対訳が最も多いという結果になった．これは，用例の新規作成をともなわ

表 1　各言語の用例数
Table 1　Number of parallel texts in each language.

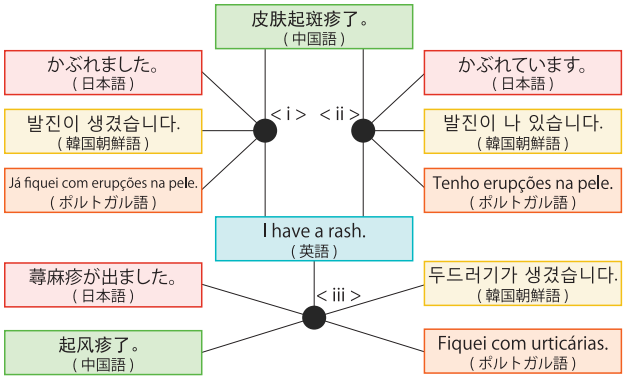
言語	用例数	
	TackPad	問診票
日本語	4,567	508
英語	2,427	473
中国語	2,330	484
韓国朝鮮語	2,136	436
ポルトガル語	2,248	475
スペイン語	706	0
ベトナム語	26	104
タイ語	38	0
インドネシア語	9	0
合計	14,487	2,480

・表中の「問診票」は多言語問診票作成システムを指す．

表 2　メタノードに結合された用例数
Table 2　Number of example sentences in meta node.

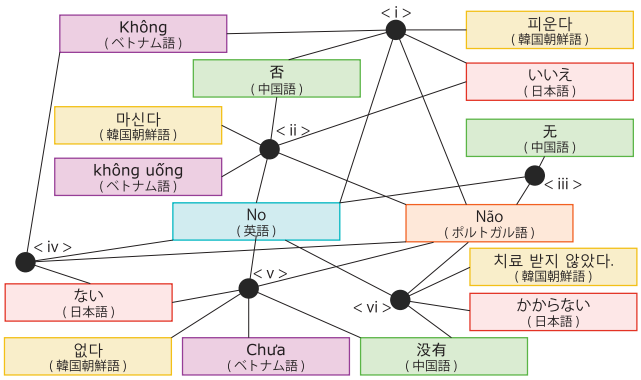
結合された用例数	メタノード数	
	TackPad	問診票
2	4,252	0
3	37	2
4	5	73
5	1,399	336
6	0	108
合計	5,693	519

・表中の「問診票」は多言語問診票作成システムを指す．



※図中の各用例は，メタノードで結合された用例同士が同じ意味であることを示す．

図 5　TackPad から抽出されたメタノードの例
Fig. 5　Example of extracted meta node from TackPad.



※図中の各用例は，メタノードで結合された用例同士が同じ意味であることを示す．

図 6　多言語問診票作成システムから抽出されたメタノードの例
Fig. 6　Example of extracted meta node from multilingual interview-sheet composition system.

表 3　各グラフ中の用例数と多対多関係を含むグラフ数
Table 3　The number of parallel texts in each graph and graphs including many-to-many relations.

グラフ中の用例数	TackPad		問診票	
	グラフ数	多対多	グラフ数	多対多
1〜2	2,411	-	4	-
3〜5	1,478	8	341	0
6〜8	686	61	99	16
9〜20	39	39	20	20
合計	4,614	108	464	36

・表中の「問診票」は多言語問診票作成システムを指す．
・表中の「多対多」は，1 対多，多対多の関係を持つ用例が存在していたグラフの数を指す．

ない用例間リンク作成機能の提供を現時点で行っておらず，3 つ以上の用例を含むメタノードを人手では作成できない状態であったためであると考えられる．また，5 つの用例が結合したメタノードも比較的多いという結果になった．これは，収集済みの用例のうち，1 万件あまりの用例はあらかじめ用意した用例対訳をデータベースに直接挿入していることが影響していると考えられる．多言語問診票作成システムの用例対訳は，あらかじめ用例間リンクを多く配置していたため，メタノードに結合された用例数が多くなる傾向にあった．

作成されたメタノードの一例を**図 5** と**図 6** に示す．図中の各用例は，メタノードで結合された用例どうしが同じ意味であることを示す．なお，用例間リンクは省略している．図 5 の用例対訳群は，TackPad から抽出されたもので，英語 1 用例，中国語 2 用例，日本語，韓国朝鮮語，ポルトガル語が各 3 用例が結合されていた．図 5 のメタノード ⟨iii⟩ には病名が含まれた用例が結合されている．また，メタノード ⟨i⟩，⟨ii⟩ にはニュアンスが多少異なる症状の用例が結合されており，メタノードによる分類を行えていることが分かる．図 6 の用例対訳群は多言語問診票作成システムから抽出されたもので，英語，ポルトガル語が各 1 用例，日本語，中国語，ベトナム語が各 3 用例，韓国朝鮮語が 4 用例結合されていた．図 6 のメタノード ⟨i⟩，⟨ii⟩，⟨v⟩ は 6 言語の用例がリンクされているが，メタノード ⟨iii⟩，⟨iv⟩，⟨vi⟩ はそれぞれ 3 言語，4 言語，5 言語の用例のみしかリンクされていないことが分かる．これらの用例対訳に対しては，3.3 節で述べた不足用例間リンクの発見手法を用いると，メタノードに含まれる用例の数が増えると考えられる．

4.1.2　1 対多および多対多の用例対訳

本項では，1 対多，多対多の関係にあった用例対訳について考察する．1 つのグラフに含まれていた用例の数と，多対多関係の用例対訳を含むグラフの数を**表 3** に示す．なお，用例を 3 文以上含むグラフは 1 対多，多対多関係になる可能性がある．表 3 より，1 対多，多対多の関係になりうる大きさのグラフのうち，TackPad の用例対訳は 4.9%（= 108/2,203），多言語問診票作成システムの用例対訳は 7.8%（= 36/460），1 対多，多対多の関係を持つ用例対訳が含まれていたことが分かる．この結果から，1 対 1 の関係にある用例対訳と比べると少ない存在ではあるが，1 対多，多対多の関係を持つ用例対訳は一定数存在していることが分かる．また，用例対訳の収集が進むと，場面や言い回しなどが多少異なる類似用例対訳も増加することが考えられ，1 つのグラフに含まれる用例対訳の数も増えると考えられる．表 3 では，グラフ中の用例対訳数が増加すると，1 対多，多対多の関係を持つ用例対訳を含む割合も

増加していることを示している．このため，1 対多，多対多の関係を持つ用例対訳は今後さらに増加することが考えられ，本手法のような 1 対 1 の関係に用例対訳を分類する手法の重要度は今後さらに増すと考えられる．

なお，本論文で述べたメタノード作成手法はクリークの概念を用いてメタノードの作成を行っている．言い換えると，メタノードに含まれる用例群は相互に用例間リンクで結ばれた状態となっている．用例間リンクは同一の意味であることを翻訳者が示したものである．このため，同一の意味ではない用例間に用例間リンクは付与されないことから，メタノードに含まれる用例対訳はすべて同一の意味であることが保証されている．ただし，基準となる用例や用例間リンクは人手で作成を行っているため，人為的な間違いが含まれる可能性がある．このため，複数人で用例や用例間リンクの正確性の確認を行うなど，間違いを検出する仕組みが別途必要であると考えられる．

4.1.3　メタノードの作成時間

本項では，メタノードの作成に必要な時間について調査し，その考察を行う．なお，計算機は CPU が Intel Xeon 3.16 GHz，メモリが 4 GB のものを使用した．また，適用データは表 1 の TackPad および多言語問診票作成システムの用例対訳である．本手法は，用例間リンクの作成時に適用する．このため，本実験ではメタノード作成手法を用例間リンクの数だけ繰り返している．

メタノードの作成時間を測定した結果，TackPad の用例対訳（用例間リンク数：18,285 件）は 2,394.6 秒（1 用例間リンクあたり約 0.13 秒），多言語問診票作成システムの用例対訳（用例間リンク数：5,278 件）は 555.8 秒（1 用例間リンクあたり約 0.11 秒），それぞれかかるという結果となった．これらから，利用者が用例間リンクを作成したときに，メタノード作成手法は毎回 0.11〜0.13 秒程度の時間が必要となることが分かる．ただし，この処理時間は用例間リンク作成後にかかるサーバの処理時間であるため，この処理に関係する利用者の待ち時間は発生しない．また，メタノード作成手法は同一のグラフ内の用例（類似した用例群）のみを走査対象としているため，手法で扱う用例数が極端に多くなることは少ない．そのうえ，1 つのメタノードに属する用例数の最大は用例対訳コーパスの対応言語数であり，最大クリークがシステムによって処理できない大きさになることはない．これらのことから，本手法は実運用に耐えることができると考えられる．

4.2　不足用例間リンク発見手法の適用結果と考察

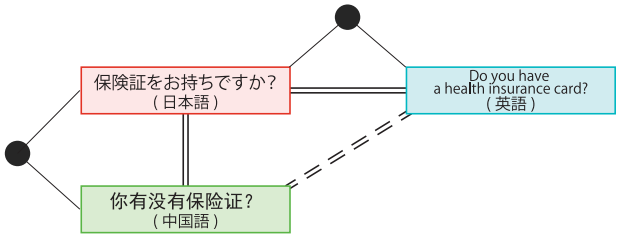
不足用例間リンク発見手法で抽出された用例のメタノードへの結合可能性を**表 4** に示す．なお，用例は複数の用例間リンクが不足している場合があるため，表 4 の用例数は延べ数である．また，表 4 の値は，用例とメタノードごとに計算しており，実際に必要な用例間リンクの数を求めて

表 4　用例のメタノードへの結合可能性

Table 4　The connected possibility of parallel texts and meta nodes.

メタノードへの結合可能性	用例数	
	TackPad	問診票
80%	0	4
75%	4	12
67%	4	4
60%	0	4
50%	13,551	11
40%	0	11
33%	27	2
30%	4	4
20%	6	7
合計	13,596	59

・表中の「問診票」は多言語問診票作成システムを指す．
・用例は複数の用例間リンクが不足している場合があるため，表中の用例数は延べ数である．



※図中の二重実線は作成済みの用例間リンクを，二重破線は不足用例間リンクを示す．

図 7　不足用例間リンクの例

Fig. 7　Example of an insufficient parallel-text link.

いるわけではない．このため，実際に必要な用例間リンクの数は表 4 の値よりも少なくなる．

表 4 のメタノードへの結合可能性をみると，TackPad の用例対訳は，多くの用例に用例間リンクが足りていない可能性が考えられる．用例の新規作成をともなわない用例間リンク作成機能の提供を行い，さらに不足用例間リンクの情報を利用者に提示することで，必要な用例間リンクの増加やメタノードに含まれる用例数の増加が行われると考えられる．また，表 4 より，TackPad の用例対訳はメタノードへの結合可能性が 50%となっているものが多かったことが分かる．メタノードへの結合可能性が 50%となっていた用例対訳の例を**図 7** に示す．図 7 では，英語と中国語の用例を用例間リンクで結合すると新たに 3 つの用例を持つメタノードを作成できることが，本手法の適用により可視化できる．また，多言語問診票作成システムの用例のメタノードへの結合可能性をみると，少ないながらも用例間リンクが足りていない可能性がみられる．特に，結合可能性が 80%（4/5 一致）や 75%（3/4 一致）のものが存在するなど，結合の可能性が高い用例を発見できていることが分かる．

なお，本手法は不足している可能性を利用者に提案するものであり，必ずしも正確ではない．このため，本手法

表 5　不足用例間リンクの妥当性
Table 5　The adequacy of the provided insufficient parallel texts relation.

言語対	候補対数	不正確	
		推薦間違い	軽微な用例の間違い
日–英	37	9	3
日–中	51	4	3
日–韓	46	7	14
日–葡	80	17	1
日–西	10	0	0
英–中	100	12	2
英–葡	99	7	1
英–西	100	2	1
西–葡	99	10	2
西–韓	99	6	17
合計	721	74	44

・表中の「推薦間違い」は不足用例間リンク発見手法で発見された用例対が不正確であったことを，「軽微な用例の間違い」は用例そのものの間違いをそれぞれ示す．

の結果の妥当性について調査を行った．調査対象は，表 4 の TackPad で 50%の結合可能性が示されたデータである．今回はそのうち，日–{英，中，韓，葡，西}，英–{中，葡，西}，西–{韓，葡} の 10 言語対[*1]，計 721 対を調査対象とした[*2]．上記の結合可能性がある用例対に関して，両言語が理解できる翻訳者各 1 名に評価を依頼した．評価は Walker らの 5 段階適合性評価 [19] を用いた[*3]．また，3 以下の評価をつけた場合は理由を併記するように依頼した．本評価では，3 以下であった場合を不正確と判定した．

評価結果を**表 5** に示す．表 5 より，本手法で発見した多くの用例対は正確であったものの，一部の用例対は不正確であったことが分かる．本手法による用例対の提示が不正確であったものの理由としては，用例中の用語の過不足や，類似した用語の使用，一人称が三人称に変化したものなどがあった．これらは，用例の翻訳に意訳が含まれていることや，日本語の主語抜けが存在していることが理由として考えられる．このため，不足用例間リンク発見手法で発見された用例間リンク候補は，正確な場合が多いものの，人手による最終判定が必要であると考えられる．また，評価結果には用例そのものの間違いが含まれていた．これは，正確性評価を用例や用例間リンクに対して適用した後に，本手法を適用すべきものであるが，本評価実験はまだ評価が行われていないものに対しても適用したためであると考えられる．このため，今後は正確性評価を十分に行った用例対に対して再適用が必要であると考えられる．

なお，本論文で提案した不足用例間リンクの発見手法は，調査対象となる用例数が増えると計算時間が増える恐れがある．しかし，不足用例間リンク発見手法の適用は，用例間リンクが作成されたときのみ行う．このため，1 度に行う処理も少なくなることから，不足用例間リンク発見手法の処理時間は実運用では問題にならないと考えられる．

*1　“葡” はポルトガル語を，“西” はスペイン語をそれぞれ示す．また，評価可能な言語対は 15 言語対であるが，英–韓，中–韓，中–葡，中–西，韓–葡，の 5 言語対は，今回の評価において日本国内での翻訳者の確保ができなかったため評価を行っていない．
*2　各言語対に対して最大 100 文をランダムに抽出した．ただし，日本語を含む言語対はすべて 100 文以下であったため，すべての結合可能性のある用例対に関して評価を行った．なお，評価対象外の用例が英–葡，西–韓，西–葡の言語対の各 1 対に含まれていたため，評価対象から除外している．
*3　評価基準は，1：まったく違う意味，2：雰囲気は残っているが元の意味は分からない，3：意味はだいたいつかめる，4：文法などに多少問題があるがだいたい同じ意味，5：同じ意味，である．

表 6　各管理手法の比較
Table 6　The comparison of each method.

比較項目	M_p	M_{e1}	M_{e2}
(1) 用例作成時のコスト	大	大	小
(2) 不足用例間リンク発見手法のコスト	小	中	大
(3) データベースのサイズ	大	中	中
(4) 1 対多関係の管理の煩雑さ	小	大	小
(5) 用例対訳利用システム側の管理の煩雑さ	小	小	大

・表中の「M_p」はメタノードで用例対訳を管理した手法（提案手法）を，「M_{e1}」は用例対訳を 1 対 1 の形で管理した手法を，「M_{e2}」はメタノードがない管理手法をそれぞれ示す．

4.3　従来手法との比較

本節では，提案手法と従来手法の計算コストや利点，欠点の比較を行う．本節では下記の 3 手法の比較を行う．また，以降の各項で比較を行う項目とその結果を**表 6** に，手法の模式図とデータベース構造を**表 7** にそれぞれ示す．

M_p（提案手法）　メタノード作成手法を適用し，メタノードで用例対訳を管理した手法である．図 3 のように，用例，用例間リンク，メタノードから構成されている．

M_{e1}　用例対訳を 1 対 1 の形になるように管理した手法である．用例とメタノードから構成された構造とほぼ同じである．図 3 から，用例間リンクを取り除いた構造を持っている．既存の用例対訳利用システム[*4]で多くとられてきた構造である．

M_{e2}　メタノードがない管理手法である．旧 TackPad で用いられていた構造である．図 2 のように，用例と用例間リンクから構成されている．

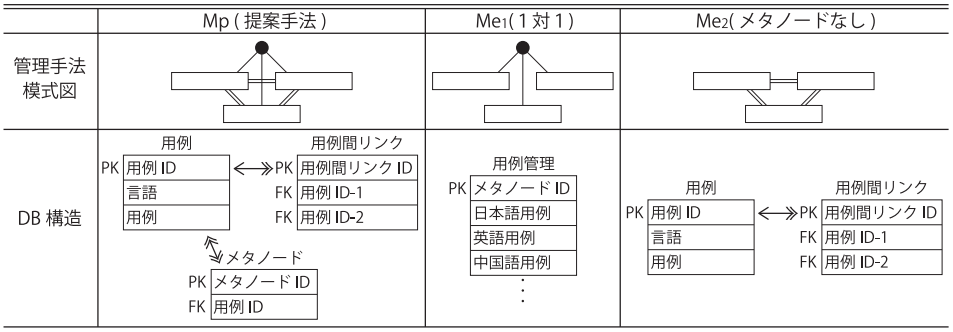
用例対訳の作成のための計算コスト

本項では用例対訳の作成時にかかる計算コストに関して考察を行う．なお，対訳の関係がない（用例間リンクが存在しない）用例作成時の計算コストはすべて同一であるため，対訳関係がある用例の作成時に関して考察する．

対訳関係がある用例の作成時には，人手によって用例と用例間リンクが作成される．その後，メタノードが存在する M_p と M_{e1} は，本論文で述べたメタノード作成を行う必要がある．しかし，M_{e2} の場合はメタノードが存在しないため，コストは用例や用例間リンクの保存コストのみとな

*4　1 章で述べた，文献 [6]，[7] などのシステム．

表 7　各管理手法のデータベース構造
Table 7　The database structures of each method.



・模式図の四角は用例を，二重線は用例間リンクを，黒丸はメタノードをそれぞれ示す．
・DB 構造の PK は主キーを，FK は外部キーを，矢印はテーブル間の関係（1 対多，多対多）をそれぞれ示す．

る．このため，用例対訳の作成時は M_{e2} の計算コストが最も少なくなる（表 6-(1)）．

不足用例間リンク発見手法の計算コスト

本項では不足用例間リンク発見手法の計算コストに関して考察を行う．なお，本手法はメタノードに属する可能性がある用例の発見を目的としている．M_{e1} に関しては用例間リンクの概念が存在しないが，この目的を満たす方法を用いて比較を行う．以下に，各手法の不足用例間リンク発見手法の計算ステップを示す．なお，各項目のギリシャ文字は同一の値となる．

M_p　(1) 調査対象用例と用例間リンクでつながった用例群（α 個）を取得
(2) (1) の属するメタノード群（γ 個）を取得
(3) (2) のメタノードに属する用例と (1) の用例間をすべて走査（$\alpha \times \gamma$ 回）

M_{e1}　(1) 調査対象用例の属するメタノード群を取得，メタノードに属する用例群（α 個）を取得
(2) (1) のメタノードに属する用例が別に属するメタノード群（γ 個）を取得
(3) (2) のメタノードに属する用例と (1) の用例間をすべて走査（$\alpha \times \gamma$ 回）

M_{e2}　(1) 調査対象用例と用例間リンクでつながった用例群（α 個）を取得
(2) (1) の用例それぞれと用例間リンクでつながった用例群を取得
(3) (2) の用例それぞれと用例間リンクでつながった用例群（β 個）と (1) の用例間をすべて走査（$\alpha \times \beta$ 回）

このとき，$\gamma \leq \beta$ である．すべての手法とも $O(n^2)$ であるが，(3) の処理は M_{e2} が多くなる傾向にあることが分かる．また，M_{e1} は (1) で 1 度メタノードを取得してから α 文の用例群を取得している．このことから，不足用例間リンク発見手法は M_p の計算コストが最も少ないという結果となる（表 6-(2)）．

各手法の利点および欠点

すべてを 1 対 1 の用例対訳として用いる従来手法 M_{e1} は，表 7 のようにデータベース構造が単純で分かりやすいという利点がある．しかし，1 対多や多対多の関係を保存する場合，用例間リンクを暗黙的に用いる必要がある．つまり，プログラム上で構造を管理する必要が出てくるため，プログラムが煩雑になる欠点がある．また，用例間リンクを用いていないため，多対多関係の用例の場合，同一の用例がデータベースの複数のカラムに保存される．この点においても管理が煩雑になり，M_{e1} の欠点となる（表 6-(4)）．

従来の TackPad で用いてきた従来手法 M_{e2} は，用例間リンクの概念があるため M_{e1} で生じた管理の煩雑さは発生しない．また，メタノードの作成をともなわないため処理時間も短い．しかし，メタノードがないため，用例対訳の提供時に問題が生じる．M_p や M_{e1} の場合，メタノード ID のみを用例対訳利用システムが保持すればよい（表 7 の場合，1 つのメタノード ID）．しかし，M_{e2} の場合，複数の用例 ID を用例対訳利用システムが保持する必要がある（表 7 の場合，3 つの用例 ID）．このため，用例対訳利用システムが管理すべき用例が不必要に多くなり，煩雑になる問題が生じる．また，用例対訳に属する用例が変更された場合に用例対訳利用サービス側では簡単に把握できない．たとえば，用例 ID が 1 の「おはよう」と用例 ID が 2 の「Good morning」から構成される用例対訳があるとする．この用例対訳中の「おはよう」が用例 ID が 3 の「おはようございます」に変更された場合，メタノードを利用した M_p や M_{e1} はすぐに反映可能であるが，M_{e2} は用例 ID を保持しているため，書き換えが必要になる．このように，M_{e2} は用例対訳の変更に弱い構造である点も欠点となる（表 6-(5)）．

最後に，提案手法 M_p について考察する．本手法は他の手法と比較して，用例，用例間リンク，メタノードと保存すべき項目が多いため，表 7 のようにデータベースの保持項目数（表 6-(3)）や処理時間が多くかかる欠点がある．しかし，他の手法で生じた管理の煩雑さは生じない．また，

4.1 節や 4.2 節で述べたとおり，処理時間の大小は大きな問題にならない可能性が高い．これらのことから，管理の煩雑さ（表 6-(4)，(5)）が少ない手法が重要となるため，提案手法が用例対訳の管理に適していると考えられる．

5. おわりに

本論文では，1 対多，多対多の関係にある用例対訳グラフを提供可能な用例対訳とするために，メタノードを利用したデータ構造を提案した．また，メタノード作成手法適用により顕在化したリンク不足の問題を解決するために，新たな用例間リンクを発見する方法について提案，実装を行った．今後は，実システム上での継続的な運用と，メタノードで結合された用例対訳の提供を行う．

謝辞　本研究の手法検討において，ご助言をいただいた大阪大学林良彦教授に感謝する．また，本研究の一部は，科研費基盤研究（B）（22300044）の助成を受けたものである．

参考文献

[1] 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).
[2] Aiken, M., Hwang, C., Paolillo, J. and Lu, L.: A group decision support system for the Asian Pacific rim, *Journal of International Information Management*, Vol.3, No.2, pp.1–13 (1994).
[3] Kim, K.J. and Bonk, C.J.: Cross-Cultural Comparisons of Online Collaboration, *Journal of Computer Mediated Communication*, Vol.8, No.1 (2002).
[4] Ishida, T.: Language Grid: An Infrastructure for Intercultural Collaboration, *IEEE/IPSJ Symposium on Applications and the Internet* (*SAINT-06*), pp.96–100 (2006).
[5] Sakai, S., Gotou, M., Tanaka, M., Inaba, R., Murakami, Y., Yoshino, T., Hayashi, Y., Kitamura, Y., Mori, Y., Takasaki, T., Naya, Y., Shigeno, A., Matsubara, S. and Ishida, T.: Language Grid Association: Action Research on Supporting the Multicultural Society, *International Conference on Informatics Education and Research for Knowledge-Circulating Society* (*ICKS-08*), pp.55–60 (2008).
[6] 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).
[7] 福島　拓，吉野　孝，重野亜久里：用例対訳を用いた多言語問診票作成システムの開発と評価，情報処理学会研究報告，グループウェアとネットワークサービス研究会，Vol.2011-GN-78, No.14, pp.1–7 (2011).
[8] Hasegawa, S., Sato, K., Matsunuma, S., Miyao, M. and Okamoto, K.: Multilingual disaster information system: information delivery using graphic text for mobile phones, *AI & Society*, Vol.19, No.3, pp.265–278 (2005).
[9] Chen, J., Chau, R. and Yeh, C.-H.: Discovering Parallel Text from the World Wide Web, *ACSW Frontiers'04*, Vol.32, pp.157–161 (2004).
[10] Tanaka, Y.: Compilation of a multilingual parallel corpus, *Proc. PACLING 2001*, pp.265–268 (2001).
[11] Bond, F., Nichols, E., Appling, D.S. and Paul, M.: Improving Statistical Machine Translation by Paraphrasing the Training Data, *Proc. IWSLT 2008*, pp.150–157 (2008).
[12] Utiyama, M., Abekawa, T., Sumita, E. and Kageura, K.: Minna no Hon'yaku: A website for hosting, archiving, and promoting translations, *Translating and the Computer 31 Conference* (2009).
[13] 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).
[14] 梶　博行，森本康嗣，相薗敏子，山崎紀之，飯田恵子，内田安彦：コーパス対応の関連シソーラスナビゲーション，情報処理学会研究報告，データベース・システム研究会，Vol.1999-DBS-118, pp.97–104 (1999).
[15] 李　航：ハイパー・シソーラスとその学習，情報処理学会研究報告，自然言語処理研究会，Vol.1992-NL-92, pp.81–88 (1992).
[16] 福井健司，柏岡秀紀：対訳文選択のための用例翻訳用シソーラスの構築，情報処理学会研究報告，自然言語処理研究会，Vol.2006-NL-124, pp.47–54 (2006).
[17] 小島秀一，高須淳宏，安達　淳：Web ページ群の構造解析とグループ化，*NII Journal*, Vol.4, pp.23–35 (2002).
[18] 岡田直樹，谷川恭平，土方嘉徳，西田正吾：交グラフと意味的解析を利用したコミュニティ発見手法，Web とデータベースに関するフォーラム（WebDB Forum 2010）(2010).
[19] Walker, K., Bamba, M., Miller, D., Ma, X., Cieri, C. and Doddington, G.: Multiple-Translation Arabic (MTA) Part 1, *Linguistic Data Consortium, Philadelphia* (2003).

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）．コミュニケーション支援の研究に従事．

（担当編集委員　前田 亮）