社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2004－NL－164（9）
2004／11／4

可読性の向上を目的とした片仮名表記の外来語に関する換言について

吉田 辰巳† 遠間 雄二† 増山 繁† 酒井 浩之†

† 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘1-1，豊橋技術科学大学知識情報工学系
E-mail: ††{nerusyu,sakai}@smlab.tutkie.tut.ac.jp, †††masuyama@tutkie.tut.ac.jp

あらまし 近年，公的文書等における難解な片仮名語の乱用が問題視されている．片仮名語は従来日本語で用いられてきた語彙に換言可能な場合も多い．換言するための知識は人手で作成することもできるが，実用的な質と量を満たすためには，多大な労力を要する．そのため，換言知識は可能なかぎり自動的に獲得することが望ましい．以上より，本研究では，難解な片仮名語をより理解しやすい語へと言い換えるための換言知識を自動的に獲得する手法を提案，その有用性を確認した．

キーワード 換言，外来語，片仮名語，可読性向上

Knowledge Acquisition on Paraphrasing Katakana Words Adopted from English for Improving Readability

Tastumi YOSHIDA†, Yûji TÔMA†, Shigeru MASUYAMA†, and Hiroyuki SAKAI†

† Dept.of Knowledge-based Information Eng., Toyohashi University of Technology, Toyohashi, Aichi, 441-8580 Japan
E-mail: ††{nerusyu,sakai}@smlab.tutkie.tut.ac.jp, †††masuyama@tutkie.tut.ac.jp

Abstract In this paper, we report about knowledge acquisition of paraphrasing katakana words adopted from English for improving readability. Recently, the problem of using difficult katakana words in public documents has been drawing much attention. Mostly these words are from English. Moreover, they often have synonyms in Japanese. However, it is impossible to obtain rules manually because of time and effort. Therefore, we propose a method of paraphrasing katakana to Japanese words automatically. Experimental results show that recall 82.8% and precision 81.1% are attained for English words restoration, recall 14.6% and precision 70.8% are attained for acquisition of paraphrasing.

Key words paraphrasing, foreign word, kanakana, readability.

1. はじめに

近年，日本語の公的文書において片仮名語が多用されるようになり，文書可読性の低下が問題となっている．国立国語研究所は，自治体の公報紙等で用いられることの多い難解な外来語である片仮名語に対して，より理解しやすい語への換言例を作成し，文書可読性の向上を推奨している(注1)．換言とは，ある言語表現の意味内容を保持したまま，別の表現へと言い換えることであり[10]，その有用性が最近注目されている[4][11]．同所は，今後も半年に50語程度ずつ換言例を追加するとしている．このような手順を取っている理由としては，一度に全ての換言知識を整備することが困難であること，外来語である片仮名語には新語が多いため，ある時点で換言知識を十分に整備したとしても，次の年には知識の追加や変更が必要になることなどが挙げられている．2004年9月現在で，提案されている語は109語であり，日常的に用いられる片仮名語の数に比べると，はるかに少ない．より大量の換言知識を整備するために，全ての換言規則を人手で作成するのではなく，ある程度自動的に取得することが望ましい．

片仮名語からの換言知識の獲得を目的とした研究として，宮木らの名詞間類似度を用いた研究[7]が挙げられる．名詞間類似度は，用いられる文脈が似ている語は意味も似ているという仮定に基づき，共起する名詞を要素とするベクトル間の余弦によって定義される．

宮木らの研究では，名詞間類似度と換言可能性を同一のものとして扱っているが，実際には異なる場合も多い．例えば，

(注1)：http://www.kokken.go.jp/public/gairaigo/

「アップル」と「ぶどう」は，用いられる文脈が似ているため，名詞間類似度は高いが換言可能ではない．宮木らの手法では，上位概念語，関連語，反意語なども換言可能と見なされてしまうからである．特に，宮木らは上位概念語を換言可能であるとしているが，この判断は必ずしも適切ではないと考える．確かに，ある語からその上位概念語への換言を行った場合，より広範な意味を持つ語になるため，間違った詳細情報が付加されることはない．しかし，換言処理の応用を考えると，一般には情報の変化だけでなく，情報の欠落も起こらないことが望ましい．上位概念語への換言は，原文の情報が抽象的になることで，詳細な情報の欠落が起こる場合がある．また，情報欠落の点を除いても，文脈によっては上位概念語に換言することが不適切な文もあることから，上位概念語は必ずしも換言の正解例とすべきではないといえる．

換言知識の自動獲得の研究としては，他にもパラレルコーパスを用いたもの[1]がある．しかし，一般に大量のパラレルコーパスを取得することは困難であり，獲得される換言知識の量は少ない．特に片仮名語に関しては，パラレルコーパス自体が入手困難であるため，本研究では考慮していない．また，久光ら[2]は，括弧表現を利用して換言事例を自動獲得している．この方法も獲得される知識の量は多くないが，単一のコーパスからも知識獲得が可能であることや，獲得された知識の正解率が高いこと等の利点もある．そのため，本研究ではこの研究に基づく手法も併用している．

2.1節に示すように，換言の必要がある片仮名語の大部分は，英語に基づく外来語である．つまり，その片仮名語が，日本語の語彙の中では新語・未知語だったとしても，元となった英単語は，英語の語彙中で既知の場合が多い．そのため，日英翻訳のための辞書を用いて，英単語から日本語訳を獲得することができる．そこで，片仮名語から元になった英単語を復元する．一般に，1つの英単語に相当する日本語訳は，数十語程度獲得されるので，その中から最も相応しい一語に絞りこむために，統計的手法を併用する．また，難解な片仮名語は新聞記事中で補足説明されている場合がある．本手法では，これらの情報も利用し，換言知識獲得の手助けとした．

以上を踏まえ，本研究では，片仮名語からの換言知識を自動獲得するための手法を考案した．その際に，片仮名語の性質に特化した手法を用いることで性能の向上を試みた．

2. 片仮名語の性質

本研究では，換言知識獲得の対象を片仮名語に限定している．名詞一般と，片仮名語とを比べると，若干性質に違いがある．片仮名語に特有の性質を利用することで，換言知識獲得の性能を向上させることができる可能性がある．以下，経験的な観察や予備実験の結果明らかになった片仮名語の性質を説明する．

2.1 片仮名語の分類

コーパスより無作為に抽出した500語の片仮名語を分類したところ，表1のようになった．種類毎の出現割合と，換言可能かどうか，換言知識を自動獲得する必要性があるかどうかを合わせて示す．

表1 片仮名語の分類

Table 1 Classification of Katakana Words.

種類	割合 (%)	換言可能性	自動獲得の必要性
英単語	55.0	可能	必要
英語複合語	3.6	可能	必要
英語略語	4.0	可能	必要
和製英語	1.4	可能	不要
和語	4.6	可能	不要
その他外来語	2.6	可能	必要
固有名詞	23.6	一部可能	不要
単位表現	1.4	一部可能	不要
オノマトペ	1.0	不可能	不要
その他	2.8	不可能	不要

以上より，本研究の対象は，英単語または英語の複合語とする．これらは，片仮名語の大部分を占め，換言可能な例が多く，かつ換言知識の自動獲得が有効である．

2.2 元になった英単語と片仮名語との違い

英単語に基づく片仮名語が，日本語の語彙において新語であったり，多くの日本語使用者にとって馴染の少ないものであったとしても，その元となった英単語は，英語の語彙において良く知られた語である場合が多い．例えば，「チュートリアル」は，いくつかの片仮名語辞書には記載されていない．しかし，元となった英単語は，簡易な英語辞書にさえ記載されている．

このことから，英単語に基づく片仮名語の換言候補を得る場合には，日本語の言語資源のみではなく，英語の言語資源を活用することが有効であると言える．

2.3 多義性・曖昧性の少なさ

一般に，使用頻度の低い単語ほど，多義性や曖昧性が少ない傾向にある．片仮名語は，日本語の名詞一般に比較すると，使用頻度の低いものが多く，そのため，多義性や曖昧性も少ない．また，片仮名語と元になった英単語とを比較しても，言葉の多義性や曖昧性は少ないものが多い．例えば，英語の「bank」には，「土手」，「(組織としての) 銀行」，「(建物としての) 銀行」，「湖畔」，「塊」等，様々な日本語訳が存在する．しかし，片仮名語としての「バンク」に相当するのは「(組織としての) 銀行」のみであり，言葉の多義性や曖昧性が減少している．

中には，「ケース」のように複数の語義を持つ片仮名語も存在する．しかし，このような語は数が少なく，また，日本語に十分に浸透しているため，和語や漢語に換言する必要のないものがほとんどである．そのため，本研究では片仮名語の多義性に関しては言及しない．換言知識として獲得されるのは，1語の片仮名語に対する0または1語の和語もしくは漢語とする．複数の語義を持つ場合はその中の1つが得られれば良いとする．

3. 手　法

3.1 英単語の復元

音素と書記素の情報を併用して，片仮名語から英単語を復元する．なお，1つの片仮名語から複数の英単語が得られた場合は，最も出現頻度の高い英単語を優先的に採用する．出現頻度はEDR辞書の英単語辞書に記載されている．片仮名語からの

英単語の復元に関しては先行研究としていくつか知られているが [12] [6]，網羅性と性能を考慮して，今回は独自のヒューリスティックスを用いた．

まず、英単語復元に用いた英単語辞書上に存在する全ての英単語に対して，発音表記，書記表記に基づき片仮名語候補を作成した．その上で，コーパス中の片仮名語と比較し，一致した先の英単語を復元英単語の候補とする．

本論文中に全ての規則を記述できないので，Web (注2)で公開した．なお，表 2，3 に一部例示する．

使用した辞書は，日本語への換言を考慮し，EDR 辞書の英語単語辞書を用いた．

表 2　発音情報に基づく変換規則の例
Example of Rules for Phonetic Signs.

母音が単体で存在した場合	
a	ア
各子音との組合せ	
ts	チャ
b	バ
dy	ジャ
d	ダ
f	ファ
g	ガ
h	ハ
j	イエ，イェ，エ

表 3　書記情報に基づく変換規則の例
Example of Rules for Spelling.

母音が単体で存在した場合	
a	ア，エイ，エー
各子音との組合せ	
b	バ，ベイ，ベー
c	カ，ケイ，ケー
d	ダ，デイ，デー
f	ファ，フェイ，フェー
g	ガ，ゲイ，ゲー
h	ハ，ヘイ，ヘー
j	ジャ，ジェイ，ジェー
ph	ファ，フェ，フェー

3.1.1　発音情報に基づく変換

換言知識の獲得を行う全ての片仮名語に関して，英単語への復元を試みた．片仮名語は，基本的に英単語の発音に従って作られている．そのため，まず英単語の発音を片仮名表記へと変換する規則を作成した．なお，発音情報は，IPA(国際発音記号) [5] に従って与えられることを前提としている．この規則は，英語の音素の全子音と全母音との組合せについて，それがどのような片仮名語の文字に相当するかを記述したものである．基本的に，片仮名語は英語の音素を日本語のローマ字へ置き換え，子音と母音の組合せによって片仮名へと変換したものが大部分であるため，この規則は有用である．そして，多重母音に関する例外や曖昧母音等の，複数存在する候補にも対応している．

3.1.2　書記情報に基づく変換

多くの片仮名語は英語の発音に従っている．しかし，書記 (つづり) に従っているものも多数存在する．この問題にも対応するため，音素ではなく，書記素と片仮名語との対応関係を認定する規則も作成した．音素に基づいた方法で，候補の英単語が見つからなかった場合に，この書記素に基づく方法を用いる．

3.1.3　長音と促音の扱い

英語には長音と促音に対応した発音表記や，書記表記は存在していないと言える．なぜならば，長音と促音は特定の音素に対応しているわけではないからである．そのため，発音表記等から長音と促音を獲得するのは困難である．本研究では，片仮名語の中の長音と促音は無視することにした．例えば，「ブクマーカー」も「ブックマーカ」も同一のものとして扱う．

上記を踏まえ，英単語復元に用いた英単語辞書 (EDR 辞書) 上に存在する全ての英単語に対して，発音表記，書記表記に基づき片仮名語候補を作成した．その上で，コーパス中の片仮名語と比較し，一致した先の英単語を復元英単語の候補とする．

3.2　複 合 語

英単語が復元されなかった場合には，複合語の可能性があるとして，語の分割を行う．本研究では，コーパス (4.1 節で後述の日経記事) から無作為に抽出した 200 の複合語について検討した結果得た，以下の 2 つの仮定に基づいた復元手法を用いている．

- 複合語は 2 語から成るものが大部分
- 複合語を構成する単語は独立した形でも使用

2 語からなる複合語は約 7 割の 146 語，そのうち 125 語において双方の片仮名語がコーパス中に存在していた．

手法としてはまず，与えられた片仮名文字列を分割可能な箇所のうち、ある 1 箇所で分割する．その結果得られた片仮名文字列が，コーパス中で独立して用いられている頻度を調べる．もし，どちらかが一度もコーパス中で使用されていなかった場合，その区切り位置は不適切であると見なす．これを全ての分割位置において実行する．また，複数の区切り位置が見つかった場合には，2 語のそれぞれの出現頻度の調和平均が最大になる区切り位置のみを適切であると見なす．

このようにして，複合語の区切り位置を獲得し，それを構成するそれぞれの語を 3.1 節に示す方法で英単語に復元する．

3.3　日本語の取得

復元された英単語に対応する日本語を獲得するために，EDR 辞書を用いた．

EDR 辞書では，言語に依存しない中間的な概念を定義し，英語と日本語の全ての語について，対応する概念の識別子が記述されている．大部分の語は，多義または広義であり，複数の概念を有する．本研究では，復元された英単語に相当する全ての概念を獲得し，さらにそれらに相当する全ての日本語訳を獲得している．そのため，一般には，1 つの英単語に対して，数十程度の日本語訳が獲得される．

(注2)：http://www.smlab.tutkie.tut.ac.jp/research/NL/DATA/KatakanaTranslated.html

3.4 統計手法による名詞間類似度

2.3節で述べた通り，多くの片仮名語は，元の英単語が持つ複数の語義のうち，1つだけに対応している．そのため，獲得した複数の日本語訳から，換言知識として相応しい1語を絞りこむ必要がある．本研究では，統計情報を用いて名詞間の類似度を判定し，片仮名語との類似度が高い日本語訳を換言知識として獲得する．名詞間類似度は，ベクトル空間法[9]によって計算している．それぞれの名詞毎に，コーパス中の同一文中に出現する名詞の意味素性に基づくベクトルを作成する．そして，2名詞間の類似度を，それぞれのベクトルの余弦によって決定する．これは，使用される文脈が似ている語ほど，類似度が高いという仮定に基づいている．例として「バンク」という片仮名語を考えた場合，英単語復元では「bank」が，日本語訳の獲得では「土手」や「銀行」等の複数候補が獲得される．日本語のコーパス中で，「土手」が用いられる文脈と，「バンク」が用いられる文脈は大きく異なるため，類似度が低いとされ，換言知識としては棄却される．一方，「銀行」は用いられる文脈が似ているため，類似度が高いとされ，換言知識として獲得される．

なお，本研究では，シソーラスの構造として木構造を想定している．そのため，葉の部分に対応する意味素性は，互いに完全に独立ではない．ある名詞が持つ意味素性と，別の名詞が持つ意味素性が異なっていたとしても，その先祖に対応する意味素性の大部分が共通しているならば，これら2つの名詞は比較的近い意味を持っているとするのが自然である．

宮木らの研究[7]では，シソーラス上で葉に対応する意味素性だけでなく，節点に対応する全ての意味素性をベクトルの要素とすることで，この問題に対処している．例えば，「コンクリート」という名詞は「セメント」という子の位置の意味素性を持っているが，これは「石材等」という，より抽象的な意味素性の子に相当する．さらに，「資材」，「人工物」，「無生物」，「具体物」，「具体」，「名詞」の順で親が存在する．コーパス中に「コンクリート」が出現した場合は，シソーラス上で「コンクリート」以上の先祖全てが出現したと見なしている．本研究でも，意味素性の頻度はこの方法に従って決定している．

本研究で用いた名詞間類似度に用いたベクトルの計算式は，以下の通りである．

$$\vec{V_i} = (v(k_i, f_0), ..., v(k_i, f_j), ..., v(k_i, f_{M-1})),$$

$$v(k_i, f_j) = FF(k_i, f_j) \log \frac{N}{SF(f_j)}$$

$\vec{V_i}$： 片仮名語もしくは換言候補 k_i が使用される文脈を反映したベクトル

f_j： 文章中に出現する j 番目の種類の名詞意味素性

N： コーパス中の文の数

$v(k_i, f_j)$： $\vec{V_i}$ の j 番目の要素

$FF(k_i, f_j)$： f_j またはその子孫の意味素性を持つ名詞が k_i と同一文中に出現する頻度

$SF(f_j)$： コーパス中で f_j またはその子孫の意味素性を持つ名詞が出現する文の数

最も類似度が高いとされた名詞のコサイン値が0.5に満たない場合，その片仮名語には換言候補が存在しないとする．

3.5 補足表現からの換言知識獲得

統計情報を用いた手法は，低頻度語に対して信頼性が低い．そこで，低頻度の片仮名語に関する換言知識を獲得するため，コーパス中の括弧表現を利用した．

新聞記事などのコーパス中では，アクティブクロス（活力拠点）のように，片仮名語の直後に日本語訳を示していることがある．本手法では，このような場合に「アクティブクロス」を「活動拠点」へと換言可能であると見なしている．ただし，セガ・エンタープライゼス（本社東京）のように，括弧内が換言知識ではない例もある．これらを区別するために，以下の規則を用いている．

括弧内の表現のうち，以下のいずれかの条件に該当する場合は，換言知識ではないとする．

- 括弧内が文である．
  - 区点，読点が存在
  - 提題表現（〜は）が存在
  - 末尾の文節が述語
- 括弧内が和語，漢語ではない．
  - 数詞が存在
  - 英字が存在
- 直前の片仮名語が会社名である．
  - 固有名詞辞書に会社名として記載
  - 語句の先頭が「代表」または「本社」
  - 語句が接頭辞または接尾辞を含む

いずれの条件にも当てはまらない場合，さらに統計情報による判定を行う．各変数を以下のように定義する．

K_i： ある片仮名語

R_j： ある括弧表現

$F(K_i)$： K_i が括弧表現とともに出現する回数

$P(R_j|K_i)$： K_i が括弧表現とともに出現した条件の下で，括弧内が R_j であるような条件付確率

換言知識かどうかを判定するための条件は，以下の通りである．なお，若い番号の条件ほど優先順位が高い．

（1） K_i に3.1節の方法を適用して得られた換言候補に，R_j があれば換言知識

（2） $F(K_i)$ が5未満ならば換言知識ではない

（3） $P(R_j|K_i)$ が0.2以下ならば換言知識ではない

（4） それ以外は換言知識

この手法は語源となる単語表現等に依存しないため，略語やその他外来語が元となる片仮名語，英語単語辞書にとっての新語に対する換言知識の獲得にも，ある程度の効果があるといえる．

4. 実験と結果

4.1 手法の実装

本手法を実装して，片仮名語からの換言知識獲得を行った．コーパスは，日経新聞記事1990年1月1日から1996年12月31までの7年間分を採用した．日英翻訳用の辞書として，EDR辞書の英語単語辞書，日本語単語辞書，概念辞書の3つを使用

した．シソーラスとしては日本語語彙大系[3]を，形態素解析器として JUMAN version 3.5(注3)を，構文解析器として KNP version2.0b6(注4) をそれぞれ使用した．

4.2 評価実験

計算機を利用した手法の評価を行う場合，基本的には正解例との比較によって行うことが望ましい．なぜなら，手法を変更した場合や，まったく異なる手法と比較するときに，同じ正解例を用いて再評価することが容易だからである．この評価を，本稿では評価法Aと呼ぶことにする．

しかし，換言知識の獲得に関しては，あらかじめ十分な正解例を作成しておくことは困難である．なぜならば，ある片仮名語から換言可能な和語，漢語が複数通り考えられるためである．特に，複合語や，日本語において一単語で表現できる単語が存在しない場合には，換言事例の数が莫大になる可能性があり，それらを完全に列挙することは難しい．そこで，上記の評価方法に加え，本研究では，手法の出力を見て，人手で換言知識としてふさわしいかどうかを判断するという評価を合わせて行った．この評価を，本稿では評価法Bと呼ぶことにする．

英単語の復元 (出力は出現頻度順)，複合語の分割 (出力は出現頻度順)，名詞間類似度のみによる知識獲得 (換言候補は全出現名詞)，括弧表現を利用した知識獲得，すべてを併用した総合的な知識獲得に関して，評価法Aで評価を行った．また，英単語の復元，名詞間類似度のみによる知識獲得，括弧表現を利用した知識獲得，すべてを併用した総合的な性能に関しては，評価法Bでも評価を行った．

なお，評価実験のために本手法を適用したのは，コーパスから無作為に抽出した 500 語の片仮名語である．

4.3 評価

本手法の部分的な性能，および総合的な性能を評価するために，標本抽出した 500 の片仮名語に関して，以下の 3 種類の正解例を作成した．

正解例 E　片仮名語の元となった英単語
正解例 S　複合語の区切り位置
正解例 P　片仮名語から換言可能な語

なお，正解例 E と正解例 S は工学部大学院生 1 人で，正解例 P は工学部大学院生 4 人が協議の上，手法の出力とは独立に作成した．正解例 E と S はほぼ自明であり，作成者による違いがほとんど生じない．そして，以下のように再現率と精度を計算した．

$$\text{再現率} = \frac{\text{正解と一致する出力の数}}{\text{正解の数}} \times 100\ [\%],$$

$$\text{精度} = \frac{\text{正解と一致する出力の数}}{\text{出力の数}} \times 100\ [\%],$$

以上の評価実験の結果を，表 4 に示す．ここで，英単語の復元に関しては，手法が第 1 位として出力したもののみを正解と見なす評価の他に，上位 3 つの英単語の中に 1 つでも正しい語があれば正解と見なす評価も合わせておこなった．

(注3)：http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html
(注4)：http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html

表 4　正解例との比較による評価 (評価法 A) の結果
Result by Estimation Method A.

評価対象	用いた正解例	再現率 (%)	精度 (%)
英語復元による知識獲得	正解例 P	27.6	2.3
英単語復元 (1 位のみ)	正解例 E	81.1	82.2
英単語復元 (3 位以内)	正解例 E	88.8	89.3
複合語分割	正解例 S	93.8	95.2
名詞間類似度のみ	正解例 P	12.8	16.7
括弧表現のみ	正解例 P	6.3	64.3
全てを併用した総合性能	正解例 P	14.6	70.8

さらに，本手法の出力が換言知識として正しいかどうか人手で検証した．その結果を表 5 に示す．

$$\text{精度} = \frac{\text{正解と見なされた数}}{\text{出力の数}} \times 100\ [\%]$$

表 5　人手での検証による評価 (評価法 B) の結果
Result Obtained manually (Estimation Method B).

評価対象	精度 (%)
英語復元に基づく知識獲得	3.6
名詞間類似度のみによる知識獲得	28.6
括弧表現からの知識獲得	92.9
全てを併用した総合性能	74.5

4.4 手法の出力の例

本手法によって獲得された換言知識の正解例を表 6 に，不正解例を表 7 に示す．

表 6　正　解　例
Sample of Correct Answers.

片仮名語	換言事例
ベース	歩調
コンサート	演奏会
ショック	衝撃
ボーナス	賞与金
メカニズム	からくり
プレッシャー	重圧
アーキテクチャー	設計思想
アメニティー	快適性
アモルファス	非晶質
インフラストラクチャー	社会的生産基盤
レプリカ	複製品
リニアモーターカー	磁気推進式列車

5. 考　察

英単語復元の部分では，第一位のみを正解とした場合でも再現率 82.2%，精度 81.1% という高い性能を得ており，この処理を片仮名語の換言知識獲得に利用することは有効であると言える．第三位までを有効とした場合も再現率，精度ともに大差なく，換言候補が膨れあがることを考慮すると，第一位のみを英単語復元の正解とした方が有用である．

換言候補獲得における，英単語復元のみによる精度は 2.3%，

表7 不正解例

Sample of Incorrect Answers.

片仮名語	獲得された知識	誤りの原因
カード	名刺	狭義語を獲得
パック	駐車場	誤った英単語 (park) を復元
ムード	気味	言葉の機微が異なる
プラットホーム	駅頭	あまり使われない日本語が獲得
クラリネット	–	対応する日本語がない
イントラネット	–	英単語としても新語

統計手法に基づく名詞間類似度を用いた処理の精度は 16.7% であるが，これらを併用した総合的な手法の精度は，70.5%であった．名詞間類似度のみで換言可能性を判定した場合の誤りは，上位概念語，関連語，反意語などを獲得してしまうことである．一方，辞書から日本語訳を得る方法の場合の誤りは，「バンク」に対する「土手」のように，不適切な語義に対応した日本語訳を獲得してしまうことである．両手法の誤り方は大きく異なっているため，これらの共通集合のみを採用することによって，いずれの誤りをも軽減されているのだと考える．

さらに，「クラリネット」のような英単語復元はできても換言候補が存在しない片仮名語が精度の限界を作り上げていると言える．そのため，固有名詞の排除，もしくは変換不要語の認定をする必要があるのではないかと考える．

手法の再現率は，総合的な処理の場合でも，14.6% と低い値であった．英単語復元のみによる知識獲得の再現率が 27.6% であるために，手法全体の再現率を低下させる大きな原因になっていると考える．そして，正解例として挙げられた語の中には，辞書やコーパス中に存在しないものが多数含まれていたため，このような結果となった．

人手で片仮名語からの換言知識を作成する場合，既存の語を組み合わせて新しい語を作る場合がある．例えば，「インフォームド・コンセント」という片仮名語に対して，国立国語研究所が提唱している言い換え例は，「納得診療」である．しかし，この語は実験で使用した辞書やコーパス中に一度も出現していない．本手法は，既存の語の中から換言事例を見付ける手法であるため，このような新しい語を獲得できていない．

換言知識として新しく作られる語の多くは，既存の語句を組み合わせた複合語である．今後，辞書にない語を換言知識として獲得するためには，新しい複合語を生成する技術が必要となる．複合語の多くは，上位概念語の前に説明が付加された形になっている．例えば，「納得診療」は「診療」の一種であり，「納得」がより詳細な意味を説明している．現在，シソーラスの自動構築などで語の上位概念を自動獲得する手法は多く提案されている．そのため，今後は詳細説明の部分をいかにして獲得するかが課題であると考える．

評価方法に関しては，正解例との比較 (評価法 A) と，出力の人手での検証 (評価法 B) の 2 種類を行った．その結果，手法の総合的な性能に関して，評価法 A の精度は 70.8%，評価法 B の正解率は 74.5% であった．この 2 つの値の差は，人が正解例を作成したときに思い付かなかったような語が，本手法によって取得されたことが原因である．これは一般に，換言知識獲得の評価は，正解例との比較だけでは不十分であることを示している．

そして，新語に関しては，いかに精度よく固有名詞等を認定，排除できるかが重要な課題と言える．なぜならば，新語が固有名詞か否かを固有名詞辞書から獲得することは困難だからである．

6. まとめ

片仮名語からの換言知識を自動的に獲得する手法を提案した．手法には，従来の名詞間類似度のほか，片仮名語の性質に特化した手法を併用した．その結果，より高い精度を得ることができたが，再現率はさほど向上しなかった．これは，人手による正解例の中に，辞書やコーパス中にない語が多数含まれていることと換言不可能語が原因である．今後は，既存の語を組み合わせて新しい語を自動的に生成し，それを換言候補とするような工夫と，換言不可能語の排除が必要になると考えている．

謝辞

言語データとして，日本経済新聞 CD-ROM 版の使用を許可して頂いた日本経済新聞社に深謝する．

文献

[1] R. Barzilay, K. R. Mckeown, "Extracting Paraphrases from a Parallel Corpus," Proc. of ACL2001, pp. 50-57, 2001.

[2] T. Hisamitsu, Y. Niwa, "Extracting useful terms from parenthetical expressions by combining simple rules and statistical measures," Didier Bourigault, Christian Jacquemin and Marie-Claude L'Homme(eds.), Recent Advances in Computational Terminology, John Benjamins Publishing Company, pp. 209-224, 2001.

[3] 池原悟，宮崎正弘，白井諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦 (編)，"日本語語彙大系，" 岩波書店，1997.

[4] 乾健太郎：言語表現を言い換える技術，言語処理学会年次大会第 8 回年次大会チュートリアル資料，pp. 1-21 (2002)

[5] International Phonetic Association, Handbook of the International Phonetic Association. A Guide to the Use of the International Phonetic Alphabet, Cambridge University Press, Cambridge (1999)

[6] Kevin Knight, Jonathan Graehl, "Machine Transliteration," Computational Linguistics, pp.599-612, 1998.

[7] 宮木衛，酒井浩之，吉田辰巳，増山繁，"カタカナ語からの換言候補の自動獲得"，情報アクセスのためのテキスト処理シンポジウム発表論文集 (電子情報通信学会「言語理解とコミュニケーション」研究会主催)，pp.96-103，2003.

[8] 酒井浩之，増山繁，"コーパスからの名詞と略語の対応関係の自動獲得"，言語処理学会第 9 回年次大会 発表論文集，pp.226-229，2003.

[9] G. Salton, "Automatic Text Processing," Addison-Wesley, 1988.

[10] 佐藤理史：論文表題を言い換える，情処学論，Vol.40, No.7, pp. 2937-2945 (1999)

[11] 山本和英：換言処理の現状と課題，言語処理学会第 7 回年次大会ワークショップ論文集，pp. 93-96 (2001)

[12] Yan Qu, Gregory Grefenstette, Daviv A. Evans, "Automatic Transliteration for Japanese-to-English Text Retrieval," SIGER 2003, pp.353-360.