名詞言い換えコーパスの作成環境

藤田篤 [*1]　乾健太郎 [*2] [*3]　乾裕子 [*1]

[*1] 九州工業大学大学院情報工学研究科
[*2] 九州工業大学情報工学部知能情報工学科
[*3] 科学技術振興事業団さきがけ研究 21「情報と知」領域
〒 820-8502 福岡県飯塚市川津 680-4
電話: 0948-29-7626　Fax: 0948-29-7601
{a_fujita,inui,h_inui}@pluto.ai.kyutech.ac.jp
http://pluto.ai.kyutech.ac.jp/plt/inui-lab/

あらまし : 名詞や動詞などの内容語の言い換えでは，言い換えの生成に必要な大規模な知識をどのように獲得するかが当面の課題になる．知識のソースには国語辞典の語釈文やシソーラスの同概念語が考えられるが，言い換えの対象語を語釈文や同概念語と置き換えるだけのナイーブな方法では構文・意味的制約を満たす言い換えは実現できない．これに対し，単語間の共起制約や語釈文と文脈の文法的整合性といった自明な制約を実装し，ある程度良質の言い換え事例を半自動的かつ大規模に生成する環境を構築した．また，この環境を用いて生成された事例を分析し，語彙的言い換えの技術的な問題点を整理した．

キーワード : 語彙的言い換え，同概念語，語釈文，自然言語処理

An Environment for Constructing Nominal-Paraphrase Corpora

FUJITA Atsushi[*1]　INUI Kentaro[*2] [*3]　INUI Hiroko[*1]

[*1] Graduate School of Computer Science and Systems Engineering,
Kyushu Institute of Technology
[*2] Department of Artificial Intelligence, Kyushu Institute of Technology
[*3] PRESTO, Japan Science and Technology Corporation
Iizuka, Fukuoka, 820-8502, JAPAN
Phone: +81-948-29-7626　Fax: +81-948-29-7601
{a_fujita,inui,h_inui}@pluto.ai.kyutech.ac.jp
http://pluto.ai.kyutech.ac.jp/plt/inui-lab/

Abstract :

In lexical paraphrasing, which is the task of replacing a content word with some semantically equivalent expression in a given context, one of the most critical issue is how to acquire the large-scale knowledge about the fine-grained semantic equivalence between near-synonyms. In this paper, we explore how one could use the existing thesauri and the descriptions appearing in machine readable dictionaries for humans for this purpose. The key idea is to extract the information about the fine-grained semantic difference between near-synonyms by matching their meaning descriptions. This paper presents a computational environment for developing nominal-paraphrase corpora, which would serve as the basis for our on-going research on paraphrasing, discussing several technical issues that have come up through our preliminary experiments.

key words : lexical paraphrasing, synonym, explication, Natural Language Processing

1 はじめに

ある言語表現をできるだけ意味を保持したまま別の言語表現に変換する「言い換え」の技術は，自然言語処理の諸分野において様々な応用が考えられる重要な要素技術である [19]．たとえば，翻訳や要約などの前処理として言い換えを行う試みはすでにいくつか報告されており [7]，日本語・手話翻訳や音声合成といった異なるタスクの前処理にも利用できる可能性がある．また，要約や推敲支援，文章読解支援のように言い換え技術が根幹をなす応用もあり，それぞれの文脈の中で言い換えの実現を目指す研究も見られるようになってきた [3, 12, 2, 8]．最近では，言い換え技術を応用からある程度独立した要素技術として捉え，その性質や実現方法を解明する試みもいくつか見られる [4, 15, 19, 16]．しかしながら，日英翻訳のような言語間翻訳が長年精力的に研究されてきた経緯と比べると,「単言語内翻訳」である言い換えの研究はまだ極めて萌芽的な段階にあると言わざるをえない．

このような背景から，我々は「要素技術としての言い換え」の事例研究として，連体修飾節に着目した構文的言い換え [18] や，文中の内容語を別の語句に言い換える語彙的言い換えの研究を進めている．本稿では，語彙的言い換えの一つである普通名詞の言い換えについて，問題の性質と取り組むべき課題を論じる．

言い換えは，ある言語表現を，同じ意味を持つ別の言語表現に変換する作業と言える．しかしながら，実際には完全に同義の言い換え対はほとんど存在せず，一般に，ある言語表現を言い換えると何らかの意味の変化が生じる．とくに，単体で指示的意味 (denotation) を持つ内容語を言い換える場合は，この「意味の差」を慎重に考慮する必要がある．たとえば,「格差」と「落差」は，EDR 日本語単語辞書 [6] によると同概念に属するが，厳密には異なる指示的意味 (denotation) を持つため，互いに言い換え可能かどうかは文脈に依存する．たとえば，次の文に現れる「格差」を「落差」に置き換えることはできない．

(1) 二人の賃金に 格差 をつける。

このことから言い換えは，

i) 言い換え前後の言語表現 (言い換え対) の間の意味の差 (意味差分) を計算し，
ii) その意味差分が所与の文脈に照らして無視できるかどうかを判断する

作業と考えることができる．したがって，少くとも語彙的言い換えに関する限り，次のような研究課題が挙げられる．

a) 言い換え対の意味差分の記述にはどのようなオントロジが適しているか．これについては，Edmonds が類義語 (near-synonym) の意味を記述するオントロジを開発し，自然言語生成の語選択に用いる方法を示している [5]．ただし，類義語の意味差分の知識を安価に獲得・記述する方法については白紙の状態である．そこで以下の研究課題が重要になる．
b) シソーラス，国語辞書，コーパスといった既存の言語資源から意味差分知識の一部を自動獲得できるとすれば，どのような方法が効果的か．カバレッジと精度はどの程度か．
c) 言い換え対の意味差分の重要性は文脈にどのように依存するか．文脈に照らして意味差分の重要性はどのような方法で計算できるか．
d) 既存の資源から獲得できない意味差分は最終的には人間 (レキシコグラファ) が手作業で記述するしかない．どのような環境を構築すれば，この作業を効率的に支援できるか．これについては，(b) と (c) に取り組む過程で必然的に明らかになると期待できる．

本研究では，上記の課題を明らかにし，大規模な語彙的言い換えを実現することを目的とする．次節では，とくに類義語間の言い換えを対象に，より具体的な研究方針について述べる．

2 言い換え対の意味差分と文脈における言い換えの可否

まず，1) 意味差分の獲得方法，2) 所与の文脈における言い換えの可否の判断方法，についてもう少し問題を整理する．

2.1 共起情報と語釈文を用いた意味差分の獲得

前節で研究課題に挙げたように，既存の言語資源から意味差分を取り出す方法としては，まず第一に語の共起情報を利用することが考えられる．共起情報を利用した統計的単語クラスタリングに関する先行研究が示すように，共起情報は語の意味的類似度の推定にある程度は有効に働く．同様に，意味差分の獲得においても共起情報の利用は欠かせないものと予想される．

ただし，4 節で述べるように共起情報だけでは類義語間の微妙な意味の差を計算することができない．そこで，二つの語の語釈文を比較することによってそれらの意味差分を計算する方法を考える．たとえば，岩波国語辞典 [10] は,「格差」と「落差」という類義語にそれぞれ次のような語釈文を与えている．

「格差」価格・資格・等級の差
「落差」落下または流下する水の，高低二か所における高さの差．転じて一般に，高低の差．

これを比較すると，語釈中の共通語である「差」の修飾語句「価格・資格・等級の」「落下または流下する水の，高低二か所における高さの」から，たとえば「差を評価する対象の違い」という知識を両者の意味差分として自動獲得できる可能性がある．

以上より，共起情報と語釈文を用いた意味差分抽出では，

- 類義語への言い換えが適格か否かを判断する知識として語の共起情報はどの程度有効に働くか．
- 国語辞書や類義語辞書の語釈文からどのような種類の意味差分情報を抽出することができるか．

ということが当面明らかにすべき課題と考えられる．

2.2 文脈における言い換えの可否

次に考えるべき課題は，意味の重なり方と言い換えの可否との関係である．

「格差」と「落差」の意味差分である「差を評価する対象の違い」が文脈上の言い換え可否にどのように影響しているか見てみよう．

(2) a. ところが，かつて学説の趨勢は「法律上の結婚に基づく家族関係を保護することに立法目的があり，格差があっても相続規定は尊重されなければならない」という法律婚重視の趣旨から，合憲としてきた．
b. 格差があっても非嫡出子に相続を認めた点に意義を見いだそうとする判断が果たして説得力を持つであろうか．
c. かつて，ヨーロッパ諸国でも嫡出子と非嫡出子との間で格差があった．

上記の例ではいずれも同一の文脈上で「格差」が現れているが，これらはすべて「落差」への言い換えが不可である．この文脈において言い換えに必要な情報として,「落差」には「資格」の意味が欠けているため不可となっている．同様に，この文脈において「格差」を言い換えるためには，この「格差」が前後の相続資格に関する文脈から「資格の差」であるという知識を得る必要がある．

「格差」のように言い換え前の語を S,「落差」のように言い換え後の語を T と呼ぶ (以下同様)．任意の S と T が与えられたとき，それらの意味の重なり方は必然的に図 1 の 5 種類のいずれかに分類できる.「格差」と「落差」の重なり方は (D) である．

上記の例からも推測できるように，(D) の関係の言い換えには様々な考慮すべき問題がある．これらの意味関係のうち，問題なく言い換えられると考えられるのは，S，T の意味が完全に同義であることを示す (A) の場合である．ただし，すでに多くの指摘があるように，二つの語の意味が完全に一致することは，表記のゆれなどを除けば，ほとんどない [5]．ただし,「コンピュータ」と「計算機」のように，形式性などの connotation の違いを無視すればほぼ同義と考えられる対は少くない．

本研究では，語彙的言い換えを目的としているため，このような connotation の違いには重点を置かず，ある一定のプロセスに従えば denotation が一致していると判定できる場合は，言い換え可能とみなす．

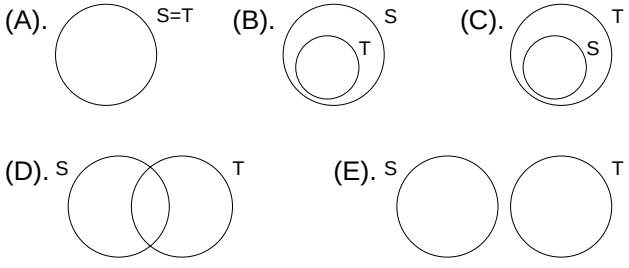

図 1: 言い換え対の意味の重なり方

(B) で言い換えが可能なのは，文脈中，S が T の意味で用いられている場合である．たとえば,「立法府」を S,「国会」を T とすると，意味の重なり方は (B) であるが，次の文脈では「立法府」が「(日本の) 国会」を指しているので，言い換え可能である．

言い換え事例

(3) 多数意見は立法府の裁量判断を尊重する立場を明示した．
→ 多数意見は国会の裁量判断を尊重する立場を明示した．

言い換え対と語釈文

「立法府」立法に参与する機関．立法機関．

「国会」国家の議会．日本国憲法では，衆議院と参議院から成り，国権の最高機関で，国の唯一の立法機関．

「格差」と「落差」の例で示したとおり，(D) の関係にある二語が言い換え可能であるのは，所与の文脈で S が S と T の意味の共通部分を指しており，かつ同じ文脈で T も S と T の意味の共通部分を指しうる場合と考えられる．したがって，言い換え可否の制約は，(B) や (C) の場合よりも厳しい.「格差」「落差」の例では言い換え不可であったが，次のように言い換え可能の場合もある．

言い換え事例

(4) 地区によって反対論に強弱はあるが，批判は「住民不在の決定方法」に集中した．
→ 区域によって反対論に強弱はあるが，批判は「住民不在の決定方法」に集中した．

言い換え対と語釈文

「地区」一区画の地域．一定の地域．

「区域」あるくぎりをつけた地域．くぎった範囲．

また，同概念語の意味が (E) のように重ならないと考えられるものもある．この場合，原則として言い換えは難しいと予想される．

言い換え事例

(5) 寄付行為の適用が社交の範囲にまで拡大されて以来，同容疑での逮捕者は初めて．
→ *寄付行為の適用が外交の範囲にまで拡大されて以来，同容疑での逮捕者は初めて．

言い換え対と語釈文

「社交」人人が集まって交際すること．また，社会上の交際．

「外交」[1] 外国との交際・交渉．[2] 外部との交際・交渉．特に，銀行・保険会社・商社などでする勧誘・交渉・注文取り．

以上を整理すると，同概念語間の言い換えの可否については下記のように仮定できる．

- (A) の関係にあれば言い換え可能である
- (E) の関係にあれば言い換え不可である
- (B)，(C)，(D) の関係にある場合は，言い換え可能と不可の両方があり得る
  - 可能な場合は，文脈上 S の意味が S，T の重なり方のうち重なる部分を指している
  - 不可の場合は，文脈上 S の意味が S，T の重なり方のうち重ならない部分を指している

上記の仮説のように文脈における言い換えの可否にもとづいて，その言い換え対の意味の重なり方との対応を考察することは以下の課題を明らかにすることである．

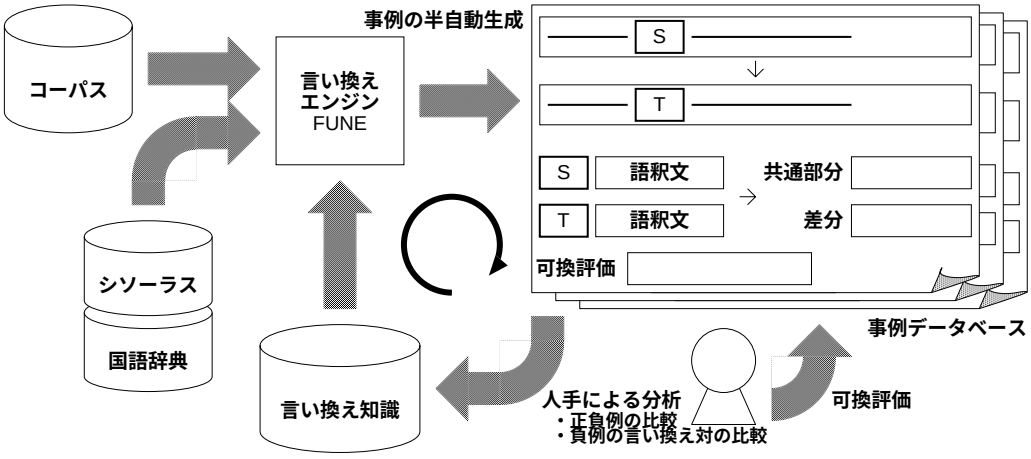

図 2: 名詞言い換えコーパスの作成環境

- 語釈文から得られる意味差分情報は実際の意味差分の重要性をどの程度近似できるか.

以下本稿では，上記で示した三点の具体的課題に取り組む準備として我々が構築した名詞言い換えコーパス作成環境について述べるとともに，予備調査で収集した事例の分析から得られた知見について報告する.

3 言い換え事例の作成環境

1 節で述べたように言い換えの研究はまだ蓄積が乏しく，同概念語間の語彙的言い換えについても問題の性質自体が明らかでない．そこで，図 2 に示すように次の 3 つの作業を順に繰り返すことによって，問題の性質を明らかにしながら，それぞれをスパイラル状に大規模化・高度化するというアプローチをとる.

(a) **言い換え事例の収集・蓄積：** 言語間翻訳の場合と異なり，言い換えは大規模な事例集合が存在しないことが研究の障害の一つになっている．本研究では，開発中の言い換えエンジンを利用して，言い換え知識の獲得に必要な事例を選択的かつ半自動的に生成する環境を構築し，言い換え事例の収集・蓄積を行う．今回のタスクでは，各言い換え対の意味差分，文脈を考慮した意味差分の重要性の評価モデルといった知識が言い換え知識に当たる．各言い換え事例には，統語的・意味的適格性の評価を人手で与えておく.

(b) **言い換え知識の獲得方法の実装：** (a) で蓄積した言い換え事例を基に，同概念語への言い換え，語釈文への言い換えのそれぞれについて，言い換えの可否を判断する言い換え知識の獲得方法を検討する．得られた獲得方法は仮説として言い換えエンジン上に実装し，(a) の作業に利用するとともに，(c) でその性能を評価する.

(c) **獲得した言い換え知識の評価：** (a) で作成した人手による適格性評価データを用いて，(b) で獲得した言い換え知識の性能を評価し，知識獲得方法の性能と限界を明らかにする.

上述の過程で用いる言い換えエンジン FUNE は，実装された言い換え知識を解釈し，入力に対応する可能な言い換えを複数生成する，言い換え実験のプラットフォームである．言い換え知識は，図 3 のような形式の言い換え規則として実装するか，または複数生成される言い換え候補の適格性を判定する基準として実装する．入力には統語・意味タグつきの XML テキストを仮定し，エンジンは同じ形式の言い換えテキストを出力する．個々の言い換え規則は，選択点，選択肢，適用条件，処理本体からなる．処理本体は，子の選択点あるいはプリミティブな変換処理の列で定義する．プリミティブな変換処理とは，タグを修正する操作や文の一部 (連体節，名詞句，単語など様々な単位で) を挿入・削除する操作などで，すでに基本的なものはライブラリとして用意してある．また，事例分析の過程で明らかになった変換処理については適宜ライブラリとして追加している.

選択点:	名詞の言い換え (S, P, M)
選択肢:	同概念語への言い換え
適用条件:	名詞である (S, P, M, Word)
	辞書引きできる (Word, Synonym, _)
	文脈中の他の語との共起事例がある (S, P, M, Synonym)
処理本体:	同概念語での置換 (S, P, M)

図 3: 同概念語との置換を表す言い換え規則

4 予備調査 (1)：同概念語への言い換え

前述の環境における名詞言い換えコーパス作成の見通しを得るために，予備調査として実際に京大コーパス [17] テキスト中の名詞を EDR 日本語単語辞書中の同概念語で言い換える実験を行った.

4.1 言い換え事例の収集と評価

京大コーパスの全文 (38,383 文) から日本語基本語彙 6000 語 [13, 14] 以外の名詞[1]を網羅的に収集したところ 121,034 箇所 (17,977 語) 集まった．このうち EDR 日本語単語辞書中で語義が一意に決まるものを取り出す[2]と 40,472 箇所 (6,951 語) あった．さらに，EDR 日本語単語辞書中で日本語概念説明が与えられている語のみを取り出すと 40,207 箇所 (6,931 語) あり，そこから接辞

[1]基本語を言い換える場合，不要な意味が付与されることが多いと考えられるため対象外とした.

[2]語の多義性解消は扱わない.

または他の名詞を伴い複合語を形成する語を除くと[3]，19,768 箇所 (5,245 語) の名詞が抽出された．以降これを対象名詞と呼ぶ．

対象名詞を入力とし図 3 の規則に従って言い換えを生成した．同概念語をもつものは 9,857 箇所 (2,670 語) あり，のべ 57,441 語の同概念語が得られたが，このうち，統語的制約の初歩として実装した EDR 日本語共起辞書 [6][4]中に共起事例がないものをフィルタアウトした結果，421 箇所 (178 語) に対する 587 文が生成された．以降これを言い換え事例と呼ぶ．

同概念語と共起制約を用いることによって生成された言い換え事例が統語的・意味的制約をある程度満たしているという仮説に対し，人手で統語的・意味的適格性の評価を行った結果，正例 342 文，負例 230 文が得られた．これ以外の 15 文は，発音によって正負が変わるような文などであり，今回は評価しなかった．

4.2 語釈文の照合パタン

共起情報だけでは微妙な意味の差を計算できないため，次に語釈文に着目する．言い換え前後の語 S，T の語釈文はそれぞれ一般に一つ以上の要素の列として記述されている．たとえば,「若者」の語釈文が次のように与えられているとすると，

「若者」年が若い者．青年.

「年が若い者」,「青年」がそれぞれ語釈文の要素である．予備実験では，S と T の語釈文の照合は，S の語釈文の各々の要素と T の語釈文の各々の要素を組み合せ的に照合し，最も類似性の高いペアの照合をもって語釈文全体の照合と見なすことにした．以下，とくに断らない限り,「語釈文」と「語釈文の要素」を区別せずに用いる．

S，T の語釈文間の照合パタンは，少なくとも図 4 のように分類することができる．

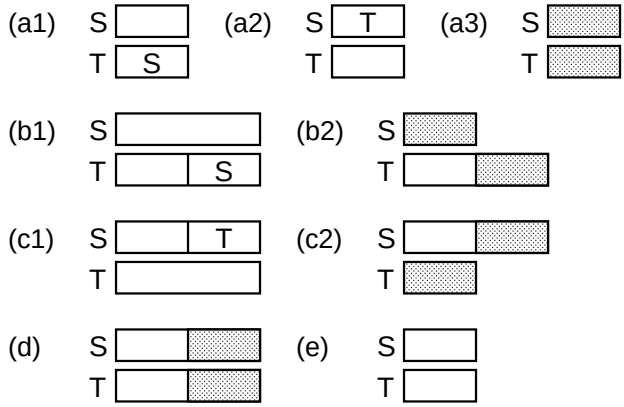

図 4: 語釈文の照合パタン

図 4 中の各図は，以下のように表現されるものであり，今回は，これらの分類を，先に述べた意味の重なり (図 1) に対応付ける．

(a) 語釈文が一致する
- **(a1)** T の語釈文が S と一致する
- **(a2)** S の語釈文が T と一致する
- **(a3)** S の語釈文と T の語釈文が一致する

(b) S が T を包含している
- **(b1)** T の語釈文が S に何らかの修飾要素を加えたものに等しい
- **(b2)** T の語釈文が S の語釈文と一致する部分を持ち，それに何らかの修飾要素を加えた形になっている

(c) T が S を包含している
- **(c1)** S の語釈文が T に何らかの修飾要素を加えたものに等しい
- **(c2)** S の語釈文が T の語釈文と一致する部分を持ち，それに何らかの修飾要素を加えた形になっている

(d) S と T が共通する部分を持ち，各々何らかの異なる修飾要素を持っている

(e) S と T が共通する部分を持たない

以下，これらの照合パタンと言い換えの可否の関係について，今回の予備調査のデータに基づき技術的な問題点と見通しを整理する．

4.3 語釈文の照合パタンごとの問題点

岩波国語辞典 [10] の語釈文を用い，前項で述べた照合パタンにより事例を分類すると，表 1 のようになった．以降，表 1 中の照合パタンごとに問題点を述べる．

表 1: 語釈文の照合パタンと言い換え事例の分布

照合パタン		正例	負例	未評価	合計
(a)	(a1)	19	3	0	22
	(a2)	29	6	0	35
	(a3)	49	9	1	59
	total	97	18	1	116
(b)	(b1)	2	2	0	4
	(b2)	16	1	2	19
	total	18	3	2	23
(c)	(c1)	25	8	0	33
	(c2)	13	6	0	19
	total	38	14	0	52
(d)		9	9	0	18
(e)		33	67	1	101
関連語・派生語・表記のゆれ		52	11	0	63
語釈文が得られなかった		95	108	11	214
Total		342	230	15	587

4.3.1 語釈文が一致する (a)

前述のとおり，S と T の語釈文が一致することは，意味の重なり方も一致することであると考えている．したがって，語釈文が (a) である関係は，意味の重なりでは (A) の関係と対応しており，仮説のとおり，言い換え可能な関係であるとみなせる．

完全に一致する場合でなくても，(6) のように語釈文の主辞「提出する」やその対象である「考え」のように語釈文の構成要素が一致する場合も言い換え可能である．

(6) **言い換え対・語釈文・言い換え事例** (a3)
「提言」考え・意見を提出すること．その内容.
「提案」議案・考えを提出すること．提出されたその考え.

[3]複合語の意味解析は扱わない.
[4]修飾語，非修飾語，修飾関係の 3 つ組からなる事例 935,860 組

経済団体の加盟企業・トップは，提言 を現実のものにしてほしい．
→ 経済団体の加盟企業・トップは，提案 を現実のものにしてほしい．

しかし，語釈文の一致の度合が高くても，(7) のように言い換え不可の場合がある．

(7) **言い換え対・語釈文・言い換え事例** (a3)
「欠陥」欠けて足りないもの．不備．欠点．
「デメリット」短所．欠点．
機体の設計に 欠陥 があったと指摘した．
→ *機体の設計に デメリット があったと指摘した．

「デメリット」が用いられる際には「メリット，デメリット」のように対応関係の一方として捉えることが自然であるように，暗黙のうちに「短所，欠点」に対して「損得，長所短所，利点欠点のうちの一方である」という意味的な限定が行なわれている．実際，英語辞典を参照すると demerit の訳語も「欠点」であるように，英語辞典であっても国語辞典であっても辞書の語釈からは，この限定的意味を得ることはできない．しかし，この例に見られるように，S，T のいずれかに「一対で用いられる概念のうちの一方」という意味差分があると，単純に言い換えることができないと考えられる．また，英単語が外来語として日本語化する際に，限定的意味が加えられたとも考えられる．

他にも，(8) のように S が慣用句の一部である場合には単純に言い換えることができない．

(8) **言い換え対・語釈文・言い換え事例** (a3)
「途方」[1] 手段．てだて．[2] 条理．すじみち．
「方法」てだて．特に，ある目的を達するための，計画的な操作．
デルタ農民が抱える課題は 途方 もなく大きい．
→ *デルタ農民が抱える課題は 方法 もなく大きい．

また，次の例では，今回収集した事例のうち少数ではあるが見つかった 4 例すべての「けじめ」が「区別」への言い換えが不可であった．これは，下記のとおり「けじめ」の語釈に「区別」そのものがあることによる問題であると考えられる．

(9) **言い換え対・語釈文・言い換え事例** (a2)
大蔵省は幹部の接待問題に けじめ をつける処分を発表した．
→ *大蔵省は幹部の接待問題に 区別 をつける処分を発表した．

「けじめ」[1]区別．[2] 隔て．
「区別」それとこれとの間に認める違い．また，それをこれと違うものとして扱うこと．

(10) は (7) の二点めの問題として挙げたように，英単語が外来語として定着する際に意味的限定が加わる例と考えられる．

(10) **言い換え対・語釈文・言い換え事例** (a2)
「制服」ある集団に属する人が着る，色や型の定められた服装．ユニホーム．
「ユニホーム」制服．特に，そろいの運動服．ユニフォーム．
登校時は 制服 を着る普通の子がいる．
→ *登校時は ユニホーム を着る普通の子がいる．

ユニホームも語釈に「制服」とあることから，意味の重なり方が一致している例とみなせたが，上記の例で言い換え不可となるのは「特に，そろいの運動服」としてのユニホームをイメージすることが自然だからである．ここで，注意しなければならないのは，この例における言い換えの不可が，外来語であることだけによるものではなく，前述の「語釈文」と「語釈文の要素」を区別しなかった本稿での立場にも起因することである．つまり，この語釈では「ユニホーム」の「制服」と「特に，そろいの運動服」の denotation の差が大きく，意味差分として利用していかなければいけないことを意味している．

また，次のように語用論的な問題を提示する事例も得られた．

(11) **言い換え対・語釈文・言い換え事例** (a2)
「地べた」〔俗〕地面．
「地面」土地の表面．転じて，土地．地所．
東京というと高層ビルばかりが思い浮かぶが，どっこい人々は 地べた の上でたくましく生きている．
→ *東京というと高層ビルばかりが思い浮かぶが，どっこい人々は 地面 の上でたくましく生きている．

俗語は，connotation の問題としてできるだけ差異を無視したいと考えていたが，denotation の違いとして意味差分を取り出さなければいけない問題も含んでいることを示した例である．

2 節で述べた仮説では問題が少ないと思われた意味の重なり方 (A) の関係は，数量的には言い換え可能な正例の方が多く，仮説を裏付ける結果となった．しかし，前述のとおり (A) の関係に対応する語釈の照合パタン (a) にも様々な問題が含まれることが明らかになった．

4.3.2 一方が他方の語・語釈文を包含する (b,c)

ここでは，意味の重なり方 (B)，(C) の関係に対応する語釈の照合パタン (b)，(c) の特徴と問題点について，言い換え可否の観点から考察する．

(12) **言い換え対・語釈文・言い換え事例** (c2)
「威力」他を圧倒するような強い 勢い．
「力」[1] ちから．...[2] ものを動かす作用．(1) 腕力・暴力・勢力．...
(i) W杯で 威力 を発揮したカウンター戦法も，相手に警戒されて不発に終わった．
→ W杯で 力 を発揮したカウンター戦法も，相手に警戒されて不発に終わった．
(ii) 出足を止められると足がそろいがちで，突き押しも腰が入らず 威力 を失う．
→ *出足を止められると足がそろいがちで，突き押しも腰が入らず 力 を失う．

(12)(i) は，S が S と T の意味の重なり部分を指しているため，言い換え可能な正例となっている．

しかし，(12)(ii) では，言い換えることによる意味の変化を無視することができない．この例では，「突き押しの程度」すなわち動作の効果としての程度を示す「突き

押しの威力」が「力」に言い換えることによって動作主である力士が「力を失う」意味に変わってしまう.「力」の多義性が，この文脈においては「突き押しの力」よりも「力士の力」に結び付きやすいため負例になった例と考えられる.

4.3.3 S と T が異なる修飾要素を持つ (d)

表 1 から分かるように，S と T が語釈文中に共通する部分を持ち，各々何らかの異なる修飾要素を持っている (d) の場合，言い換え事例の正負例のばらつきが多い．この部分は，S が T を包含する (b) の問題と T が S を包含する (c) の問題の両方を同時に持つ形であり，2 節でも述べたとおり意味の重なり方 (D) に対応する複雑な関係であるため，このように言い換え不可となる可能性が高くなるのは必然である．(b) と (c) で考えられることを解消しなくては (d) の問題解消は困難であると考えられる.

4.3.4 S と T が共通部分を持たない (e)

同概念語かつ共起制約を満たし，かつ意味の重なり方の度合が高ければ言い換え可能であるという仮説に対して，実際の評価結果，負例が多く生成された．各々が負例となった原因を明らかにするために言い換え対 S と T の語釈文の照合を図ったが，S と T の語釈文が共通部分を持たない (e) の場合，言い換えの可否の判断基準が得られない．表 1 から分かるようにこの共通部分を持たない場合は負例の割合が非常に多い．たとえば，(13) が挙げられる.

(13) **言い換え対・語釈文・言い換え事例** (e)
「懸念」[1] 気にかかって不安がること．心配．[2] 執念．執着.
「責任」人や団体が，なすべき務めとして，みずから引き受けなければならないもの.
この背景には，新民連が新進党と連携するのではないか，という 懸念 がある.
→ *この背景には，新民連が新進党と連携するのではないか，という 責任 がある.

しかしながら (14) のような正例も生成されている.

(14) **言い換え対・語釈文・言い換え事例** (e)
「随所」いたる所.
「あちこち」あちらやこちら．[1] ほうぼう．[2] くいちがうさま.
随所 でがれきの山が生まれ，火災も発生し，死傷者も多数，確認されている.
→ あちこち でがれきの山が生まれ，火災も発生し，死傷者も多数，確認されている.

T が同概念語であることを考慮するとこれは必然である．したがって，語釈文から共通部分が得られないことが，負例たる条件であるとはいえない．正例を獲得するためには，たとえば，岩波国語辞典から適切な語釈文を得られなかったとしても，類語辞典や使い分け辞典など，語の共通部分・差分の記述に重点をおいている別のリソースを用いることが考えられる．たとえば，角川類語新辞典 [11] によれば,
「随所」限定されないどの場所にも．方々
「あちこち」方々.「あちらこちら」よりぞんざいな言い方
という語釈文が得られ，S と T が同じ語で説明されている (a3) の事例であるとみなすことができる.

4.4 課題

前項で負例の原因，正負例の差を見ることによって明らかになった問題点を整理し，具体的な技術的課題について述べる.

本研究の課題として，(i) 言い換え対間の意味差分の獲得，(ii) 所与文脈に照らして差分を無視できるか否かの判定，を挙げたが，まず，言い換えの可否を判定するための入力となる意味差分を語釈文の照合パタンで近似した意味の包含関係から抽出することが課題となる．前項で述べたように，照合パタン (d) については，(b), (c) の問題を複合的に持っているため，まずは (b), (c) を対象とし，意味差分の抽出と，文脈における S の指示推定を行う.

次に，意味の重なりを近似するための語釈文の照合に関する課題が二つ挙げられる．一つ目は，語釈文中の複数の要素を考慮することである．語釈文の照合パタンとしては最も類似性の高い要素のペアを用いたが，実際は語釈文中の各要素もまた，語の多義性を表し得ることを考慮するべきである．二つ目は,「転じて〜」,「〜の略」,「〜の一種」,「〜という」などのメタな表現を利用することである．これらの表現に着目することで，語彙の指示的意味を抽出するためのヒューリスティックを設けることができると考えられる.

最後に，外来語，慣用句などを当該処理の対象外と同定することも重要である．外来語を用いて日本語の意味を捉えようとする場合は，語釈文との照合とは別に，意味にずれが生じないように前処理を行う必要がある．また，慣用句は，句全体を言い換えるべきなので，選択的事例収集の段階でケアすべき課題である.

5 予備調査 (2)：語釈文への言い換え

また以下に示すような，対象語を国語辞典の語釈文に言い換えるタスクも並行して進めている.

EDR 日本語単語辞書の語釈文 (日本語概念説明) によると,「核軍縮」は「核兵器の軍事的規模を縮小すること」,「容疑者」は「罪を犯した疑いをかけられている人」であるが，これを用いて (15), (16) の言い換えを実現するには，語釈文の構文的カテゴリを変換したり，冗長な部分を削除するといった処理が必要になることは自明である.

(15) 核保有国は 核軍縮 への努力を誓ったことになる.
→ 核保有国は 核兵器の規模の縮小 への努力を誓ったことになる.

(16) 役員は 容疑者 の業者からうまい話を持ち掛けられた.
→ 役員は 罪を犯した疑いをかけられている 業者からうまい話を持ち掛けられた.

このタスクについては，図 5 に示す仮説の規則を用いて言い換えの生成を行った.

選択点:	名詞の言い換え (S, P, M)
選択肢:	語釈文への言い換え
適用条件:	名詞である (S, P, M, Word)
	辞書引きできる (Word, _, Explication)
	語釈文!=見出し語 (Word, Explication)
処理本体:	語釈文の構文的カテゴリ変換 (S, P, M, Explication)
	語釈文の冗長部分削除 (S, P, M, Explication)
	語釈文の原文への埋め込み (S, P, M, Explication)

図 5: 言い換え規則 (語釈文の埋め込み)

処理本体中，語釈文の構文的カテゴリ変換の際に語釈文と文脈の文法的整合性を考慮するための情報としては，次の二種類の情報を用いた．一つは，文脈情報として，文献 [9] に挙げられた名詞の構文上の働きの分類であり，もう一つは語釈文の情報として，語釈文の主辞による分類である．後者は，語釈文末尾の形式名詞「こと」を除いたものを対象としている．予備調査では前述の言い換えエンジンを用いて，名詞の構文上の働きのタグのうち数種類を対象名詞に自動的に付与した．そしてまずは，頻度が多く文脈の他の内容語との結びつきが比較的少ない，

- 修飾を受けない述語の項 (6,676 箇所 (2,767 語))
- 修飾を受けない修飾語 (3,344 箇所 (1,503 語))

として出現しているものに対象を絞り，振る舞いがある程度記述できる構文的なカテゴリ変換に対するライブラリの実装と，言い換えの生成を行った．これらの作業については，実装に対する動作確認を終えている．

6 おわりに

本稿では，大規模な語彙的言い換えの実現に向け，語彙間の意味差分の解明のために大規模な事例集合を半自動的に作る環境の構築と，予備調査，今後取り組むべき課題について報告した．今回の調査を通して，語彙の指示的意味を知識として獲得する際の問題の性質を整理することができた．

今後は，言い換え知識の獲得に必要な事例を選択的に生成し，人手による統語的・意味的適格性について評価して事例を蓄積することにより，言い換えの知識の獲得方法，言い換え知識，言い換えコーパスをスパイラル状に大規模化・高度化していく方針である．

具体的にはまず，以下のことに取り組む．

(i) 基本語の基準として「日本語の語彙特性：単語親密度」[1] を用いる．これを用いると任意の語彙レベルを指定することができるため，テキスト簡単化という目的を考える上で，語彙的言い換えの簡単化の評価を比較的容易に行える可能性がある．また，その分類の判断基準を何らかの知識としてを使うことも考えられる．

(ii) 高精度の統計的部分構文解析器 [8] を利用し，新聞記事データから単語の共起事例を大量に収集する．

謝辞

本研究を進めるにあたり，東京工業大学の奥村学氏には大変貴重なコメントをいただきました．ここに厚く御礼申し上げます．また，日頃より活発にコメントを下さった九州工業大学乾研究室の皆様に深く感謝します．

参 考 文 献

[1] 天野成昭, 近藤公久. 日本語の語彙特性 1: 単語親密度. 三省堂, 1999.

[2] Carroll, J., et al. Simplifying text for language-impaired readers. *Proc. of EACL*, pp. 271–272, 1999.

[3] Dras, M. Reluctant paraphrase: Textual Restructuring under an Optimization Model. *Proc. of PACLING'97*, pp. 98–104, 1997.

[4] Dras, M. Representing Paraphrases Using S-TAGs. *Proc. of ACL-EACL'97*, pp. 516–518, 1997.

[5] Edmonds, P. Semantic Representations of Near-Synonyms for Automatic Lexical Choice. Ph.D. thesis, published as technical report CSRI-399, Department of Computer Science, University of Toronto, 1999.

[6] EDR. 電子化辞書仕様説明書 第 2 版. Technical Report, 日本電子化辞書研究所, 1995.

[7] 江原暉将, 福島孝博, 和田裕二, 白井克彦 聴覚障害者向け字幕放送のためのニュース文自動短文分割. 情報処理学会自然言語処理研究会予稿集, NL-138-3, pp. 17–22, 2000.

[8] 乾健太郎. テキスト簡単化による聾者向け読解支援 －現状と展望－. 電子情報通信学会福祉情報工学研究会予稿集, WIT2000-34, 2000.

[9] 情報処理振興事業協会技術センター. 計算機用日本語基本名詞辞書 IPAL(Basic Nouns) －解説編－. 1996.

[10] RWC. RWC テキストデータベース第 2 版, 岩波国語辞典タグ付き/形態素解析データ第 5 版. RWC, 1998.

[11] 角川書店. 角川類語新辞典. 1981.

[12] 片岡明, 増山繁, 山本和英. 要約のための連体修飾節の“ A の B ”への言い換え. 情報処理学会自然言語処理研究会予稿集, NL-133-7, pp. 37–44, 1999.

[13] 国立国語研究所. 分類語彙表. 大日本図書, 1964.

[14] 国立国語研究所. 日本語教育のための基本語彙調査. 秀英出版, 1984.

[15] 近藤恵子, 佐藤理史, 奥村学. 「サ変名詞＋する」から動詞相当句への言い換え. 情報処理学会論文誌 Vol. 40, No. 11, pp. 4064–4074, 1999.

[16] 近藤恵子, 佐藤理史, 奥村学. 格変換による単文の言い換え. 情報処理学会自然言語処理研究会予稿集, NL-135-16, pp. 119–126, 2000.

[17] 黒橋禎夫, 長尾 眞. 京都大学テキストコーパス・プロジェクト. 言語処理学会第 3 回年次大会発表論文集, pp. 115–118, 1997.

[18] 野上優, 藤田篤, 乾健太郎. 文分割による連体修飾節の言い換え. 言語処理学会第 6 回年次大会発表論文集, pp. 215–218, 2000.

[19] 佐藤理史. 論文表題を言い換える. 情報処理学会論文誌, Vol. 40, No. 7, pp. 2937–2945, 1999.