クラス指向事例収集手法による言い換えコーパスの構築

藤田　篤[†]　乾　健太郎[††]

語彙・構文的言い換えの中には，形態・構文的パターンに基づいて一括りにできるものの，表現を構成する語の統語・意味的な特性に依存して言い換えの可否や言い換え方が決まる現象が少なくない．本論文では，そのような言い換えを語彙構成的言い換えと呼ぶ．たとえば，複合語を構成語に分解するような言い換え，機能動詞構文の言い換え，態や格の交替，種々の動詞交替，語彙的派生などは語彙構成的言い換えの範疇に含まれる．我々は現在，これら語彙構成的言い換えに関わる語の統語・意味的な特性を明らかにするため，および言い換え生成技術の定量的評価のために，個々の言い換えクラスごとに言い換え事例集（言い換えコーパス）を構築している．本論文では，言い換え前後の表現の形態・構文的パターンと既存の言い換え生成システムを用いて言い換え事例を半自動的に収集する手法について述べる．また，日本語の機能動詞構文の言い換え，動詞の自他交替を対象とした予備試行の結果を報告する．

キーワード: 言い換え，言い換えコーパス，語彙構成的言い換え，言い換えクラス，機能動詞構文，自他交替

Building a Paraphrase Corpus Based on Class-oriented Candidate Generation

Atsushi Fujita[†] and Kentaro Inui[††]

Several classes of paraphrases have a potential to be compositionally explained by referring to syntactic and semantic properties of constituent words: e.g., composing/decomposing compounds, voice/case alternation, various verb alternation, and lexical derivation. Toward analyzing the compositionality underlying these paraphrase classes, we have examined a class-oriented framework for collecting paraphrase examples, in which sentential paraphrases are collected for each paraphrase class separately by means of automatic candidate generation based on morpho-syntactic paraphrasing patterns, followed by manual judgement. Our preliminary experiments on building two paraphrase sub-corpora have so far been producing promising results with regard to cost-efficiency, exhaustiveness, and reliability.

KeyWords: *paraphrasing, paraphrase corpus, lexically compositional paraphrase, paraphrase class, light-verb construction, transitivity alternation*

1　はじめに

意味が近似的に等価な言語表現の異形を**言い換え**と言う．言い換えの問題，すなわち同じ意味内容を伝達する言語表現がいくつも存在するという問題は，曖昧性の問題，すなわち同じ言

† 名古屋大学大学院工学研究科, Graduate School of Engineering, Nagoya University
†† 奈良先端科学技術大学院大学情報科学研究科, Graduate School of Information Science, Nara Institute of Science and Technology

語表現が文脈によって異なる意味を持つという問題と同様，自然言語処理における重要な問題である．

言い換えの自動生成に関する工学的研究には，言い換えを同一言語間の翻訳とみなし，異言語間機械翻訳（以下，単に機械翻訳）で培われてきた技術を応用する試みが多い．たとえば，構造変換方式による言い換え生成 (Lavoie et al. 2000; Takahashi et al. 2001)，コーパスからの同義表現対や変換パターン（以下，合わせて言い換え知識と呼ぶ）獲得 (Shinyama and Sekine 2003; Quirk et al. 2004; Bannard and Callison-Burch 2005) の諸手法は，機械翻訳向けの手法と本質的にはそれほど違わない．ただし，言い換えは入出力が同一言語であるため，機械翻訳とは異なる性質も備えている．たとえば，平易な文章に変換する，音声合成の前処理として聴き取りやすいように変換するなど，ミドルウェアとしての応用可能性が高いことがあげられる．すなわち，言い換えを生成する過程のどこかに，応用タスクに合わせた言い換え知識の使い分け，および目的適合性を評価する処理が必要になる (乾，藤田 2004)．

事例集の位置付けも異なる．翻訳文書は日々生産・蓄積されており，大規模な対訳コーパスが比較的容易に利用可能である．これらは主に，翻訳知識の収集源あるいは統計モデルの学習用データとして用いられている．一方，言い換え関係にある文または文書の対が明示的かつ大規模に蓄えられることはほとんどない．2 節で述べるように，言い換えの関係にある文の対を収集して**言い換えコーパス**を構築する試みはいくらか見られるが，我々が知る限り，現在無償で公開されている言い換えコーパスは Dolan ら (Dolan and Brockett 2005) が開発したものしかない[1]．さらに，言い換え知識の収集源として用いられるようなコーパスはあっても，言い換えと呼べる現象の類型化，個々の種類の言い換えの特性の分析，言い換え生成技術の開発段階における性能評価などの基礎研究への用途を意図して構築された言い換えコーパスはない．

我々は，言い換えの実現に必要な情報を実例に基づいて明らかにするため，また言い換え生成技術の定量的評価を主たる目的として言い換えコーパスを構築している．本論文では，このような用途を想定して，

- どのような種類の言い換えを集めるか
- どのようにしてコーパスのカバレージと質を保証するか
- どのようにしてコーパス構築にかかる人的コストを減らすか
- 言い換え事例をどのように注釈付けて蓄えるか

などの課題について議論する．そして，コーパス構築の方法論，およびこれまでの予備試行において経験的に得られた知見について述べる．

以下，2 節では言い換えコーパス構築の先行研究について述べる．次に，我々が構築している言い換えコーパスの仕様について 3 節で，事例収集手法の詳細を 4 節で述べる．予備試行の設定を 5 節で述べ，構築したコーパスの性質について 6 節で議論する．最後に 7 節でまとめる．

1 Web 上のニュース記事から抽出した 5,801 文対に対して 2 名の評価者が言い換えか否かのラベルを付与したコーパス．http://research.microsoft.com/research/nlp/msr_paraphrase.htm

2　先行研究

言い換えコーパスの構築に関する先行研究は，内省に基づく生成，コーパスからの自動獲得の 2 種類に大別できる．いずれにおいても，コーパスを構成する個々の事例は，言い換えの関係にある文対である．

2.1　内省に基づく言い換え生成

同じ原文に対して複数の翻訳がある場合，それらは言い換えとみなすことができる．機械翻訳では，システムの評価方法として 1 つの原文に対して複数の正解翻訳例を用意することが一般的になってきており，そうした複数の翻訳例を含む対訳コーパスもいくつか整備されつつある (白井，山本 2001a, 2001b; Zhang et al. 2001; 金城ら 2003; 下畑ら 2003)．

人間が内省に基づいて言い換えを記述するアプローチは大きな人的コストを要する．それにも関わらず，上述の先行研究では，どのような種類の言い換えを集めるのか，その範囲の言い換えをどのようにして網羅するのかという課題に対する解は示されていない．先行研究の多くが，言い換えそのものへの関心よりもむしろ，機械翻訳の被覆率・訳質の改善を主たる目的としているためであろう．たとえば，(白井，山本 2001a, 2001b) は機械翻訳の被覆率向上を目的として低頻度語や語のあらゆる語義を網羅するため例文収集方法を提案している．しかしながら，語や語義ごとの例文を得るための手段として言い換えを用いているに過ぎず，様々な種類の言い換えを網羅する，あるいは所与の例文に対して十分多様な言い換えの例を収集することについては焦点を当てていない．

2.2　コーパスからの自動獲得

近年，同義表現対や変換パターンなどの言い換え知識を獲得するために，言い換え関係にある文対を自動的に収集する試みが報告されるようになってきた．とくにここ数年は，同じ出来事を報道している複数の新聞社の記事を対応付ける試みが多い (Barzilay and Lee 2003; Shinyama and Sekine 2003; Quirk et al. 2004; Dolan et al. 2004; Dolan and Brockett 2005; Brockett and Dolan 2005)．このアプローチでは，異なるコーパス中の文と文を，内容語や固有表現の重なり具合，構文構造の類似度，文の抽出元の記事の日付や記事中の文の位置などのメタ情報に基づいて照合し，言い換えらしい文対を得る．

言い換え文対の自動獲得手法には人的コストを必要としないという利点がある．収集された個々の言い換え文対には多くのノイズが含まれるが，これを人手で除外するにしても，内省に基づいて事例を記述する手法に比べてコストは低い．また，未知の種類の言い換えを発見できる可能性も秘めている．しかしながら，収集可能な言い換え事例の種類は文の照合における制約によって擬似的に限定されるため，コーパス中に出現している言い換えを網羅的には収集で

きない．また，制約を特に設けずに言い換えらしい文対を集めるとしても，類似する文対を漠然と集めているに過ぎず，複数の言い換えが組み合わさった複雑な言い換え事例が含まれてしまう可能性がある．このような事例を現象解明に向けた分析に利用するには，人手による言い換えの分解・分類を要する．

3　対象とねらい

言い換えと呼べる現象は多岐にわたる．その中には談話の状況に関する高度な推論を要するものもあり (乾，藤田 2004)，現在の技術ですべてを実現することは難しい．そこで，まずはどのような種類の言い換えの事例を集めるかについて議論する．

言い換えに関する工学的研究のほとんどが，語あるいは言語表現の内包的意味が等価であるような現象を対象としている．そのような現象は，主として語と語の意味の同一性や自他の構文交替，態交替などの構文的な変形に基づいて実現されるため，**語彙・構文的言い換え**と呼ばれる．本論文で扱う対象もこの例に洩れない．語彙・構文的言い換えに限っても，純粋に統語論で扱えそうな言い換えから語の詳細な意味に立ち入る必要のある言い換えまで多岐にわたるが，実現に必要な知識の観点から以下のように 4 種類に分けて考えることができる．

統語的言い換え：　個別の語の意味に立ち入らなくても，統語論の記述レベルで概ね説明できる言い換え

(1)　最初に合格した**のは**高橋さん**だ** ⇔ 高橋さん**が**最初に合格した

語彙的言い換え：　語の同義性だけで概ね説明できる，統語操作を伴わない局所的言い換え[2]

(2)　一層の**苦境**に陥る**恐れ**がある ⇔ 一層の**窮地**に陥る**可能性**がある

語彙構成的言い換え：　言語の統語的特性と意味的特性に基づいて構成的に説明できると考えられる規則性の高い言い換え

(3)　2 位**が**先頭**との**距離**を縮めた**⇔ 2 位**と**先頭**の**距離**が縮まった**

推論的言い換え：　世界知識や社会慣習に根ざし，統語論，語彙意味論のような言語に関する知識だけでは説明が難しい言い換え[3]

(4)　財政再建**が急務の課題だ** ⇔ **緊急に**財政再建**する必要がある**

言い換えの計算モデルが実用規模で機能するためには，大規模な言い換え知識が必要となるので，その開発および保守を効率化するための方法論が重要な研究課題になる．知識開発に関しては，人手で作成された既存の語彙資源を利用するアプローチと 2.2 項で述べたような手法で得た言い換えコーパスから言い換え知識を自動獲得するアプローチがある．言い換え知識の

2 言い換えの実現に必要な知識という観点では，慣用表現から文字通りの意味を持つ表現への言い換えもこの分類に入る．たとえば「手を上げる」という表現を言い換える場合，表現全体を「降参する」または「殴る」に言い換えるべきか，「手」や「上げる」と言う構成語のみを言い換えるべきかという曖昧性がある．ただし，言い換え前の表現が構成的か非構成的かを見分けることも広く語義曖昧性解消の課題と位置付ければ，言い換えそのものは，語を別の語に置き換える場合と同様，局所的に同義の表現対の知識を用いて実現できる．

3 比喩表現や間接発語行為から文字通りの意味を持つ表現への言い換えなども含む．

自動獲得に関する研究動向についての詳細は乾，藤田 (2004) の解説に譲るが，既存の語彙資源から抽出できるのは限定的な種類の言い換え知識だけであり，またコーパスから力任せに自動獲得する方法もこれまでのところ実用に耐える成果を挙げられていないのが現状である．

さて，語彙・構文的言い換えの中には，次に示す一連の例のように，構成的に計算できる可能性が高い，上で語彙構成的言い換えと呼んだ現象も少なくない．

(5) **動詞交替**

a. 洗濯物**が**風**に揺れる** ⇔ 風**が**洗濯物**を揺らす**　（自他交替）

b. 円**の**レート**が下がった** ⇔ 円**が**レート**を下げた**　（自他交替・再帰）

c. 先輩**が**後輩**に**合格の秘訣を**教える** ⇔ 後輩**が**先輩**から**合格の秘訣を**教わる**　（授受の動詞交替）

d. 多くの地域**が**暴風雨**に見舞われた** ⇔ 暴風雨**が**多くの地域**を見舞った**　（直接受身）

e. 翔一**が**誰か**に**自転車を**盗まれた** ⇔ 誰か**が**翔一**の**自転車を**盗んだ**　（間接受身）

f. 通り**が**群衆**で**あふれた ⇔ 群衆**が**通り**に**あふれた　（場所格交替）

g. 柳**が**芽**を**ふく ⇔ 柳**に**芽**が**ふく　（湧出動詞の交替）

h. 太郎**が**犯人**であると**認める ⇔ 太郎**を**犯人**と**認める　（補文構文）

(6) **範疇交替（品詞交替）**

a. 息子が友人の活躍に**刺激を受ける** ⇔ 息子が友人の活躍に**刺激される**　（機能動詞構文（格+動詞） ⇔ 動詞）

b. 部屋は十分**暖まっている** ⇔ 部屋は十分**暖かい**　（動詞 ⇔ 形容詞）

c. 彼女は頬を**赤らめて**うなずいた ⇔ 彼女は頬を**赤くして**うなずいた　（動詞 ⇔ 形容詞+する/なる）

d. 身体**のだるさを**感じる ⇔ 身体**がだるいと**感じる　（名詞（句） ⇔ 形容詞（節））

e. 水がとても**清らかだ** ⇔ 水がとても**清い**　（ナ形容詞+だ ⇔ イ形容詞）

(7) **その他の構文的な交替**

a. 彼の言葉**に温かみ**を感じた ⇔ 彼の言葉**の温かさ**を感じた　（係り先の交替（格））

b. **厳密に**審査基準を定める ⇔ **厳密な**審査基準を定める　（係り先の交替（修飾語））

c. 彼**の**顔が真っ赤だ ⇔ 彼**は**顔が真っ赤だ　（係り先の交替（主題化））

d. 目的地は**赤い屋根の**建物だ ⇔ 目的地は**屋根が赤い**建物だ（主辞交替（名詞句 ⇔ 節））

e. **リサイクルの効率化**が求められる ⇔ **効率的なリサイクル**が求められる　（主辞交替（名詞句 ⇔ 名詞句））

f. **財政再建**が課題だ ⇔ **財政を再建すること**が課題だ　（複合名詞の分解・構成）

g. 夕飯を**食べ過ぎた** ⇔ 夕飯を**必要以上に食べた**　（複合動詞の分解・構成）

h. 新しい機材**の必要性**を議論する ⇔ 新しい機材**が必要かどうか**を議論する　（名詞接尾辞の着脱）

これらの例はそれぞれ異なる形態・構文的パターンによって特徴付けられる．このパターンに基づいて一括りにできる言い換え現象を，本論文では**言い換えクラス**と呼ぶ．言い換えクラスの実在性は言語学的な分析 (Mel'čuk and Polguère 1987; Jackendoff 1990; Levin 1993; 影山 2001) においても示されており，場所格交替や自他交替の構成性を言語学的に説明する試みもある．これをふまえると，語彙構成的言い換えについては，個別の語の統語・意味的特性に関する知識と一般性の高い原理的な変換規則によって実現することが望ましい．語彙構成的言い換えが構成要素の語彙的知識から組み合わせ的に計算できるとすれば，少なくともそのクラスの言い換えについては，人手で開発・保守できる規模の語彙資源で実現することができる．

我々の言い換えコーパス構築の動機は，これら語彙構成的言い換えに関わる語の統語・意味的な特性を明らかにすること，その過程で言い換え生成に関する仮説を定量的に評価することにある．そこで，次に示すような要求仕様を念頭におき，個々の言い換えクラスごとに言い換えコーパスを構成する．

- 言い換えコーパスは言い換えクラスごとのサブコーパス群からなる．
- 各サブコーパスは所与の言い換えクラスに属する言い換え関係にある文対の集合からなる．
- 各サブコーパス中の言い換え事例は実世界における表現の分布（密度，多様性）を反映している．

4　形態・構文パターンを用いた言い換え事例の半自動収集

3 節の議論をふまえ，所与の言い換えクラス C に属する言い換え事例を，文集合 S から **(i) できるだけ網羅的に**，**(ii) できるだけ少ない人的コストで**収集するという目標を設定する．当然，各事例の言い換えとしての適否の判定の **(iii) 信頼性をできるだけ高く**保たねばならない．

まずは，どのような方法論でどのような言い換えクラスの言い換え事例を収集できるかを経験的に調査する必要がある．その試みの一つとして，本論文では，次の 3 ステップからなる半自動的な事例収集手法について検討する．

ステップ 1.　所与の言い換えクラス C について，形態・構文的変換パターン集合と辞書的な知識を記述する．

ステップ 2.　既存の言い換え生成システムを用いて，所与の文集合 S に変換パターンを適用し，言い換え事例の候補集合を生成する．

ステップ 3.　言い換えクラスごとに**適否判定ガイドライン**を用意し，それに基づいて個々の言い換え候補を適格，不適格に分類する．

この手法は 2 節で述べた 2 種類のアプローチの中間に位置付けられる．すなわち，言い換え生成システムの利用により，(i) 言い換え事例収集における人的コストを低減すると同時に，(ii) 所与の言い換えクラスに対するカバレージ，および (iii) 適否判定の質を保証することをねらい

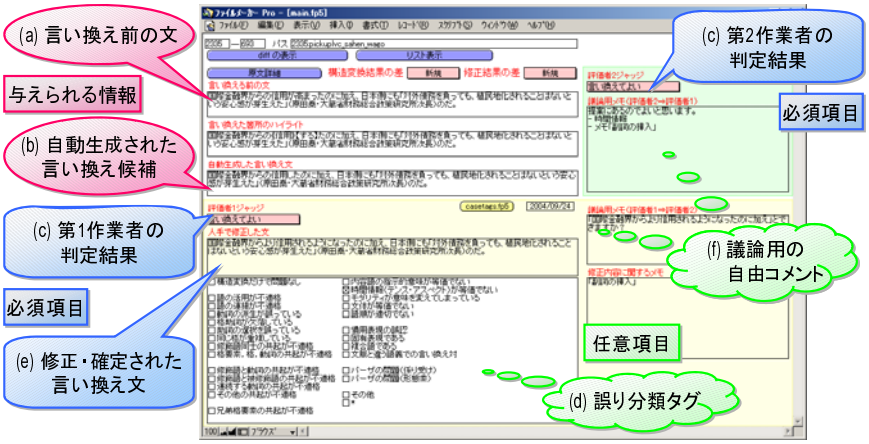

図 1　自動生成した事例の適否を判定するための作業環境

としている．

この手法では，ステップ 1 と 3 に人手を要する．まず，ステップ 1 において，所与の言い換えクラス $\mathcal{C}$ を定義するための形態・構文的パターンを記述する必要がある．これは，文献 (Dras 1999) における文法開発と同様に，少数の典型的な言い換え事例に基づいて帰納的に作成する．たとえば，例 (6a) から，機能動詞構文の言い換えに関する (8) のようなパターン[4]を記述する．

(8) s.　N を $_{(\Rightarrow V)}$　V

t.　$v(N)$

ここで，N，V は各々名詞，動詞を表す変数，$v(N)$ は N の動詞形を表す．係り受け関係に関する条件を右下に添えた矢印で表す．上の例では「N を」が「V」に係ることを条件としている．

所与の文集合 $\mathcal{S}$ に含まれる言い換え可能な文を網羅的に収集するために，形態・構文的パターンには過剰な制約を記述しないように配慮する．逆に，不適格な例を大量に生成してステップ 3 のコストを増やしてしまわないように，言い換えのクラスごとに語彙的な資源を用意する．たとえば，パターン (8) においては，N と $v(N)$ を動作性名詞とその動詞形に限定する．形態素解析や係り受け解析の精度が十分実用的な精度であるため，形態・構文パターンと語彙的制約に基づいて言い換えクラスを定義するアプローチは現実的であると考える．

ステップ 3 では自動生成された言い換え候補の適否判定に人手を要する．ここでは言い換えクラスごとにどのような種類の誤りが生じるかをある程度予測できるという仮説に基づき，言い換えクラスごとに適否判定ガイドラインを作成しておく．このガイドラインは，文献 (藤田，乾 2003) が示す言い換え誤りの分類に基づき，あらかじめいくらかの作例に基づいて予測でき

4 s.，t. は各々言い換え前後の文あるいはパターンを表す．言い換え後の文が文法的・意味的に不適格な場合は記号 “$*$” を，文法的・意味的に適格でも言い換えとして適切でない場合は記号 “$\neq$” を記す．また，言い換えの適否に関する作業者の判定結果が分かれた事例については記号 “?” を記す．

る範囲で誤りの種類を列挙したものである．言い換え候補の適否を判定する作業の過程で未知の誤りが出現した場合や作業者間で判定結果が分かれた場合は，いくらか事例が溜まった時点で議論し，このガイドラインを更新する．

作業者は図 1 に示す作業環境で個々の言い換え候補の適否を判定する．(a) 言い換え前の文と (b) 自動生成した言い換え候補が与えられたときに，作業者は，(c) その言い換え候補の適否，(d) もしも不適格であればその原因の分類情報，(e) 修正することで言い換え可能，あるいは複数の言い換えが可能ならばそれらを記述する．判定に迷った候補については，(f) 議論用に自由形式でコメントを記述する．

5　言い換えコーパス構築の予備試行

言い換えコーパス構築における種々の課題に対し，前節で述べた言い換え事例の収集方法がどれだけ有効であるかを検証するため，事例分析および言い換え生成実験の評価に利用できる規模のサブコーパスを構築した．今回は，機能動詞構文の言い換え，および動詞の自他交替の 2 つの言い換えクラスを取り上げた．この節では，共通の設定について述べた後，各言い換えクラスを対象としたサブコーパス構築の詳細について述べる．

5.1　共通の設定

我々の手法では，形態・構文的パターンと対象文との照合のためにいくつかのソフトウェアを必要とする．今回は，形態素解析器『茶筌』[5]，係り受け解析器『南瓜』[6]，言い換え生成システム『KURA』[7] を用いた．言い換え候補を収集する文のドメインは新聞記事中の文とした．具体的には日本経済新聞[8]（2000 年，一文あたり平均 25.3 形態素）を用いた．茶筌，南瓜が新聞記事中の文を学習に用いているため，また，非常に稀なクラスの言い換え事例を集める場合でも十分大規模な文集合を用意できるためである．

言い換え候補の適否判定は，日本語母語話者であり大学卒業程度の教養を備えている 2 名の作業者が実施した．今回は作業コストをできるだけ削減するというねらいから，2 名が完全に独立に言い換え候補の適否を判定するのではなく，図 2 に示す 3 ステップの手順で判定した．以下，各ステップについて述べる．

ステップ 1.　第 1 作業者は自動生成した言い換え候補の各々の適否を判定する．

ステップ 2.　第 2 作業者は，第 1 作業者が『適格』とした言い換え候補をすべて判定する．また，第 1 作業者の判定が過度に『適格』,『不適格』に偏っていないかを確認するため，第 1 作業者が『不適格』とした言い換え候補をサンプリングして判定する．

5 http://chasen.naist.jp/
6 http://chasen.org/~taku/software/cabocha/
7 http://cl.naist.jp/kura/doc/
8 http://sub.nikkeish.co.jp/gengo/zenbun.htm

ステップ 3.　2 名の判定結果が分かれた候補について，数日に一度作業者間で議論する．また，適否判定のガイドラインを更新し，一貫性を保つためにステップ 1 に戻って再度判定する．議論を経ても適否判定が一致しなかった場合は『保留』とする．

5.2　機能動詞構文の言い換え（LVC）

機能動詞構文とは，動詞が動作性名詞を格要素に持つときに，実質的な意味を失い，単に動詞としてのみ機能するような構文である．ここで例 (6a) を再掲して説明しよう．

(9) s.　息子が友人の活躍に**刺激を受ける**．
　　t.　息子が友人の活躍に**刺激される**．

例文 (9s) では，動作性名詞「刺激」が実質的な動作内容を表しており，動詞「受ける」は動作の方向を表しているに過ぎない．今回は，(9s) を (9t) に言い換えるように，機能動詞構文の動詞を取り除き，動作性名詞の動詞形を主辞に据えるような言い換えを扱う．

例文 (9s) では動作性名詞は対格に現れているが，例 (10)，(11) のように，動作性名詞が主格，与格になる場合でも機能動詞構文が形成されうる．

(10) s.　彼女に対する気持ちに**変化が起こった**．
　　t.　彼女に対する気持ちが**変化した**．

(11) s.　日本の住宅事情を**考慮に入れる**．
　　t.　日本の住宅事情を**考慮する**．

また，和語動詞，サ変名詞は形態的には異なるが同じように動作性名詞として機能動詞構文を形成する．これらを考慮し，(12) のようなパターンを 4 種類記述した．

(12) s.　$N\{\text{が}, \text{を}, \text{に}\}_{(\Rightarrow V)}\quad V$
　　t.　$v(N)$

ここで，N，V，$v(N)$ は (8) と同様，各々名詞，動詞を表す変数，N の動詞形を表す関数である．格助詞の部分の “{”，“}” は選言を表す．

次に，⟨ 間違い, 間違う ⟩，⟨ 考慮, 考慮する ⟩ のような動作性名詞と動作性名詞の動詞形の組を用意した．具体的には，茶筌が用いる日本語形態素解析用辞書『IPADIC』からサ変名詞，和語動詞の連用形とそれらの動詞形の組を 20,155 組抽出した．この集合を N と $v(N)$ に関する語彙的制約とする．他方，機能動詞構文を形成しうる動詞については，文献 (村木 1991) に約 60 語例示されているものの網羅的とはいえない．このため，V についてはとくに制約を設けなかった．

形態・構文的パターンは Kura によって自動的に係り受け構造の対に変換され，所与の文集合に網羅的に適用される．10,000 文を入力したときに自動生成された言い換え候補は 2,566 件であった．

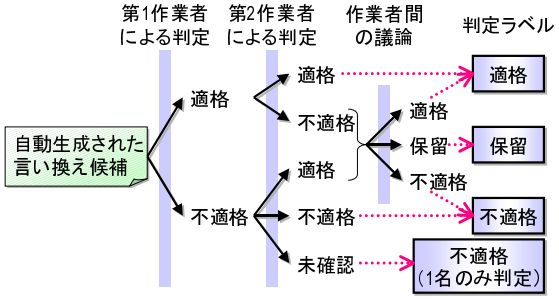

図 2　事例ごとの適否判定の確定までの流れ

個々の言い換え候補の適否判定にあたり，可能ならば作業者が言い換え候補を修正する．形態・構文的な情報のみでは言い換え先の表現を一意に決められない場合や KURA に実装されている誤り修正機構が不適切な修正を施してしまう場合があるためである．機能動詞構文の言い換えについては，(i) 活用形の変更，(ii) 格助詞の変更，(iii) 副詞の挿入，(iv) ヴォイス表現，アスペクト表現，ムード表現などの動詞性接尾辞の追加，の 4 種類の修正処理を許可した．たとえば，(12) に示したパターンを例文 (9s) に適用した場合は，正しい言い換え文 (9t) を得るために受動態を表すヴォイス表現「られる」を後続する．一方，例文 (13s) に同じパターンを適用した場合は，始動相を表すアスペクト表現「しだす」を後続するとともに格助詞「の」を「を」に置き換える．

(13)　s.　コンサートのチケット**の販売を始めた**.
　　t.　コンサートのチケット**を販売しだした**.

現在までに，最初の 4,500 文に対する言い換え候補 983 件の判定を終えている．また，残りの 5,500 文に対する言い換え候補から無作為に選んだ 131 件のみ判定を終えている．内訳は，判定結果が『適格』であった候補が 547 件,『不適格』であった候補が 520 件,『保留』であった候補が 47 件であった．ある文が異なる形に言い換えられる場合は作業者の内省に基づいて思い付くだけ例を記述しており，547 件の言い換え候補に対して 591 件の言い換え事例を得ている．

参考までに,『不適格』,『保留』とされた言い換え候補の例を各々例 (14)，(15) に示す．

(14)　s.　憲政擁護をさけぶ民衆の**デモに包囲された**.
　　t. ≠憲政擁護をさけぶ民衆に**デモされた**.

(15)　s.　「存続は不可能」と**区切りをつけ**たがっている感じもしないではなかった.
　　t. ?「存続は不可能」と**区切り**たがっている感じもしないではなかった.

例文 (14s) における「包囲する」はヴォイスあるいはアスペクトなどの機能を持っておらず「対象を取り囲む」という意味を持っている．これを取り除くように言い換えると意味が変化してしまうため，不適格とした．一方，例 (15) では,「区切りをつける」を一つの慣用句とみなし「終わらせる」というべきか,「(仕事を) 区切る」とすべきかで意見が分かれた．また，自動生

成した言い換え候補の中には，例 (16)，(17) のように，収集しようとしたのとは異なるクラスの言い換えもあった．このような候補についても可能ならば言い換え例を記述したが，適否については『不適格』とした．

(16) s. **帰りに立ち寄る**温泉も大きな楽しみだ．
　　t. **帰りの**温泉も大きな楽しみだ．　（動詞省略による換喩化）

(17) s. **調査によると**，仕事でのパソコン利用率は八六・一%．
　　t. **調査の結果**，仕事でのパソコン利用率は八六・一%．　（複合辞の言い換え）

5.3 動詞の自他交替（TransAlt）

例 (5a) のような動詞の自他交替を実現するためには ⟨ 揺れる, 揺らす ⟩ のような自動詞と他動詞の組に関する知識が必要である．しかし，語彙調査の過程で作られた辞書や自他交替を扱う言語学の文献に断片的には記述されているものの，網羅性の高い資源はない．そこで，少なくとも収集源からは言い換え候補をもれなく収集できるように，自動詞と他動詞の組を人手で記述する．まず，次の (18) のような動詞の抽出パターンを記述した．このパターンは，形式的には言い換え候補の自動生成のための形態・構文的パターンと等しいが，言い換え後の表現を与えていない点のみ異なる．

(18) s. N_1 が $_{(\Rightarrow V)}$　$N_2\{$ に, から, で $\}_{(\Rightarrow V)}$　V
　　t. 変形なし．

ここで，N_1，N_2 は名詞を表す変数，V は動詞を表す変数である．なお，2 つの格要素が動詞に係ることを条件としているが，これらの順序は問わない．

言い換え候補を 1,000 件程度生成することにし，LVC とのおおまかな頻度の比較から言い換え候補の収集源として 25,000 文を用いることにした．この文集合に (18) などのパターン群を適用したところ，V に対応する動詞 800 語が取り出された．そして，各動詞に対して人手で自動詞，他動詞を付与を記述したところ，自動詞と他動詞の組を 212 組収集できた．

次に，言い換え候補の自動生成のために，(19) のようなパターンを記述した．

(19) s. N_1 が $_{(\Rightarrow V_i)}$　N_2 に $_{(\Rightarrow V_i)}$　V_i
　　t. N_2 が $_{(\Rightarrow v_t(V_i))}$　N_1 を $_{(\Rightarrow v_t(V_i))}$　$v_t(V_i)$

N_1，N_2 はここでも名詞を表す変数である．一方，V_i，$v_t(V_i)$ は自動詞とそれに対応する他動詞を表しており，上の 212 組を用いて実現する．動詞の自他交替には例 (20)，(21) のように様々な助詞が関わるが，どの要素を主格に据えるべきかは文脈に依存するため，すべての候補を別々に生成する．また，例 (22) のように他動詞文を自動詞文に言い換える例も同時に収集するため，合計 8 種類のパターンを記述した．

(20) s. 与党の法案**に**野党**から**反対意見**が出る**．
　　t. 与党の法案**に**野党**が**反対意見**を出す**．

(21) s. 戦火や迫害**で**難民**が生まれる**.
　　t. 戦火や迫害**が**難民**を生む**.
(22) s. 2 位**が**先頭**との**距離**を縮めた**.
　　t. 2 位**と**先頭**の**距離**が縮まった**.

動詞の自他交替についても適否を判定するためのガイドラインを作成し，修正の例を掲載した．具体的には，(i) 活用形の変更，(ii) 格助詞の変更，(iii) ヴォイス表現の変更，の 3 種類の修正処理を許可した．たとえば，例文 (22s) のように他動詞を自動詞に置き換える場合,「2 位が」や「先頭との」をどのように残すべきかは形態・構文的な情報のみでは特定できない．ゆえに，非決定のまま生成した候補を人手で修正する．

自動詞と他動詞の組を得る際に用いた 25,000 文に上述のパターン群を適用した結果，985 件の言い換え候補が生成された．これまでにこれらすべての判定を終えており，その内訳は『適格』が 461 件,『不適格』が 503 件,『保留』が 21 件であった．LVC の場合と同様，ある文が異なる形に言い換えられる場合があったため，461 件の言い換え候補に対して 484 件の言い換え事例を得ている．参考までに,『不適格』,『保留』とされた言い換え候補の例を各々例 (23)，(24) に示す．

(23) s. 議会の多数党**が**政権の座**についた**.
　　t. ≠議会の多数党**を**政権の座**につけた**.
(24) s. ビスマルクの左ＣＫ**を**熊谷**が**頭**で決めた**.
　　t. ?ビスマルクの左ＣＫ**が**熊谷**の**頭**で決まった**.

例 (23) は 2 名の作業者が同じ理由で『不適格』とした．言い換え前の文が自然発生的な出来事を指すにも関わらず，言い換えた後の文においては，それが何らかの主体の行為によってなされたという含みを持ってしまうためである．一方，例 (24) は，言い換えることによって「(ゴールを) 決める」が表していた行為の動作主性が損なわれると考えるか否かで意見が分かれたため『保留』とした．また，LVC の場合と同様に，収集しようとしたのとは異なるクラスの言い換えもいくらか出現したが『不適格』とした．例を (25) に示す．

(25) s. 北朝鮮側の提案**が**米側の希望**を**十分に**満たし**ていなかった.
　　t. 北朝鮮側の提案**で**米側の希望**が**十分に**満たされ**ていなかった.　（直接受身）

6　議論

前節で述べた 2 つの言い換えサブコーパスの仕様を表 1 に示す．また，図 3，4 に適否の判定結果が確定した言い換え候補の数を示す．図中の横軸は 2 名の作業時間の合計であり，言い換え候補の判定時間，作業者間の議論の時間，適否判定ガイドラインの更新後に各候補の適否を再度判定する時間を含む．以下，(i) 事例収集効率，(ii) 収集した事例の網羅性，(iii) 判定結果の信頼性について述べ，(iv) 言い換えクラスの定義について議論する．

表 1　構築した言い換えサブコーパスの仕様

言い換えクラス	LVC	TransAlt
言い換え候補の収集源の文数	10,000	25,000
言い換えパターンの数	4	8
語彙知識の種類	$\langle n,\, v_n \rangle$	$\langle v_i,\, v_t \rangle$
語彙知識の規模	20,155	212
言い換え候補の数	2,566	985
作業者が適否を判定した言い換え候補の数（Judged）	1,114	985
判定結果：『適格』（Correct）	547	461
判定結果：『不適格』（Incorrect）	520	503
判定結果：『保留』（Deferred）	47	21
収集した言い換え事例の数	591	484
作業時間（人時間）	118	169.5

6.1　事例収集効率

現在までに 2,031 件の言い換え候補の判定結果が確定（5.1 項で述べた通り不適格な候補の大半は 1 名のみの判定結果）しており，1,075 件の言い換え事例が収集できた．図 3，4 が示すように判定の速度は比較的安定していた．一人時間あたりでは，7.1 件の言い換え候補の適否を確定，3.7 件の言い換え事例を収集できている．先行研究では事例収集効率を定量的に評価していないため，我々の手法がどれほど効率的であるかを比較によって示すことはできない．ただし，同じ作業者が判定結果を見直すための時間，作業者間の議論の時間も計上していることを考慮すれば，妥当な速度といえよう[9]．

さらなる事例収集効率の向上のためには，どの作業に最も時間を要しているかの分析が必要である．今回は各作業の時間を計測していなかったため，作業者のヒアリングに基づいて次の 2 つの原因を取り上げる．第一に，言い換え候補を不適格とした場合にどのような誤りが原因で不適格としたかの記述（図 1 の (d)）に時間を要していた．誤り分類の体系は形態素情報や品詞体系，係り受け構造の情報に基づいているため，馴染みのない作業者には分類が難しかったようである．第二に，言語テストの難しさが作業効率を低下させる原因となっていた．これは，TransAlt において LVC よりも顕著に（1.75 倍）作業効率が悪かったことにも現れている．本論文で用いたクラス指向の言い換え事例収集手法の効率は，用いている言語テストにも影響される．これについては 6.4 項で詳述する．

ある程度の時間をかけても適否が判定できなかった場合に判定を保留することにすれば，さらなる効率化は実現できる．ただしこれは，次に述べる言い換え事例の網羅性という指標とのトレードオフになる．

9 文献 (Brockett and Dolan 2005; Dolan and Brockett 2005) では，Web 上のニュース記事から抽出した 10,000 文対を 2 名の作業者が独立に言い換えか否かに分類している．Chris Brockett 氏とのパーソナルコミュニケーションによると 2〜3 日（4〜6 人日）で作業を終えたとのことであるが，この試みでは，(i) 言い換えのクラスを限定せず，(ii) 適否に関する厳密なガイドラインなしに節の重複の度合いと作業者の直感に基づいて判定し，(iii) 判定結果が分かれた場合は議論なしに不適格としているためである．すなわち，本論文のような言い換えの適否に関する議論はない．

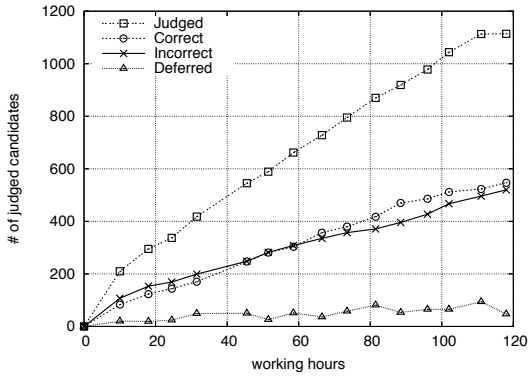

図 3　適否を判定した言い換え候補の数およびその判定結果の内訳（LVC）
各線の意味は表 1 を参照されたい.

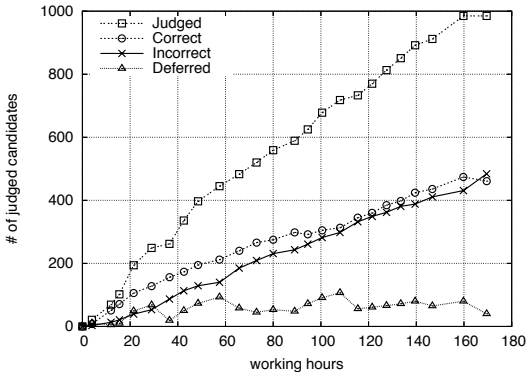

図 4　適否を判定した言い換え候補の数およびその判定結果の内訳（TransAlt）
各線の意味は表 1 を参照されたい.

6.2　網羅性

どれだけ網羅的に言い換え事例を収集できているかを見積もるために，LVC で用いた文集合から無作為に 750 文取り出し，人手で同じクラスの言い換えを試行した．作成された 206 事例のうち獲得済みの事例は 158 事例であり，カバレージは約 77%（158 / 206）となった[10]．形態・構文的パターンでは収集できなかった 48 事例のうち，解析誤りによるものは 1 件のみであった．ゆえに，形態・構文的パターンを用いた候補生成は現実的なアプローチであると言える．34 件は形態・構文的パターンをいくつか追加することで自動的に収集できる．たとえば，(26) のようなパターンを追加すれば，(27) のような事例も収集できるようになる．

(26)　s.　N 化{ が, を, に }$_{(\Rightarrow V)}$　V
　　t.　v(N 化)

(27)　s.　これは市場**の活性化にむけた**規制緩和策だ.
　　t.　これは市場**を活性化する**規制緩和策だ.

残りの 13 件の取りこぼしは，⟨ ズレ, ズレる ⟩，⟨ 伸び, 伸びる ⟩ のような動詞形を持つ語の品詞が IPADIC において一般名詞となっていたことに起因する．これらをあらかじめ辞書に記述しておくことはパターンの記述に比べると難しいが，形態素辞書の整備が進めばカバレージを上げられると期待できる．

手持ちのパターンおよび語彙資源がどれだけのカバレージを持っているか，制約としてどれだけ適切であるかを，言い換え生成および人手による適否判定の前に知ることはできない．ゆえに，上のような人手による分析は，我々がある言い換えクラスに対して持っている直感的な定義と自動的に収集できる範囲との違いを見極めるために欠かせない作業である．

10 TransAlt の場合は格が省略されている文を抽出していないため，LVC よりもカバレージが低いと予想される．

6.3　判定結果の信頼性

判定結果の信頼性を保証するには，より多くの作業者を用いる必要がある．ただしそれは人的コストとのトレードオフになる．そこで我々は，作業者間の判定結果に揺れが生じないように言い換えクラスごとに適否判定ガイドラインを設け，適格な言い換え候補についてのみ多重判定を施した（図 2）．また，判定に悩んだ場合は何日か後に見直す，作業者間で判定結果が分かれた場合は議論を通じて適否判定ガイドラインを更新するなどの工夫を施した．

適格と判定された言い換え候補に関する作業者間の一致率は，作業への習熟，および適否判定ガイドラインの更新に伴って上昇した．たとえば，LVC の場合の作業者間一致率は，74%（3 日目），77%（6 日目），88%（9 日目），93%（11 日目）であった．このことは，作業者間の議論によって判断に悩むような言い換え候補や作業者間で判定結果が分かれるような言い換え候補に関する情報が整理され，ガイドラインが洗練されてきていることを示唆している．

図 2 の判定手順がもたらす判定結果の信頼性をより正確に見積もるため，今後は第 1，第 2 作業者とは独立に言い換え候補の適否を判定する第 3 作業者を立てる予定である．

6.4　言い換えクラスの定義に関する議論

特定の言い換えクラスのみを考えるならば言い換えの適否の判定基準を明確に定義できると期待していた．しかし，LVC と TransAlt の作業効率の比較から，必ずしもその期待は満たされないことが明らかになった．

TransAlt では他動詞を自動詞に言い換える際に格要素が欠落することをどこまで認めるかが議論になり，我々は，言い換えによって生成された自動詞文の主格要素が意志性（あるいは内在的コントロール (影山 1996)）を持つか否かに着目した．すなわち，自動詞文に「自ら」,「勝手に」などの副詞を挿入した場合に文として成り立つ場合には，言い換え前の他動詞文の主格が自動詞文では含意されないため不適格とした．この言語テストに照らすと，例 (28) は適格，(29) は不適格と判定される．

(28)　s.　彼がスープ**を温めた**.
　　t.　スープ**が温まった**（* 勝手に）.

(29)　s.　彼が氷**を溶かした**.
　　t. ≠氷**が溶けた**（勝手に）.

ただし，言い換え前の文の主格が言い換えによって欠けるため，両例とも不適格だとする考え方もある．今回の試みによって蓄えられた多くの言い換え事例と適否判定ガイドラインには，今後このような問題を議論するための素材としての用途もある．

7　おわりに

言い換えという現象を工学的・言語学的側面の両方から解明するためには，様々な言い換えを漠然と扱うだけでなく，特定の言い換えクラスに焦点をしぼった事例研究が欠かせない．本論文では，このような基礎研究の基盤となる言い換えコーパスを構築するため，言い換え前後の表現の形態・構文的パターンと既存の言い換え生成システムを用いる半自動的な事例収集手法について検討した．また，2 つの言い換えクラスを取り上げた予備試行を通じ，この手法が比較的頑健に作用することを示した．

言い換えコーパスに求められる仕様はその用途によって異なると予測される．たとえば，言い換え技術の性能評価用のコーパスは実際に用いられる表現の分布を反映する必要があるが，言い換えの構成性を裏付ける語の統語・意味的な特性を特定するためには，特定の構成要素ごとに偏りのないコーパスが求められる．ゆえに今後は，実際の言い換えコーパスの構築を通じてこれらの仕様の整理とそれを実現する技術の開発に取り組みたい．そして，事例収集効率と適否判定の信頼性の改善をはかりながら，3 節で示したような語彙構成的言い換えのそれぞれについてコーパスを構築していきたい．

参考文献

Bannard, C. and Callison-Burch, C. (2005). “Paraphrasing with bilingual parallel corpora.” In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp. 597–604.

Barzilay, R. and Lee, L. (2003). “Learning to paraphrase: an unsupervised approach using multiple-sequence alignment.” In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pp. 16–23.

Brockett, C. and Dolan, W. B. (2005). “Support Vector Machines for paraphrase identification and corpus construction.” In *Proceedings of the 3rd International Workshop on Paraphrasing* (*IWP*), pp. 1–8.

Dolan, B., Quirk, C., and Brockett, C. (2004). “Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources.” In *Proceedings of the 20th International Conference on Computational Linguistics* (*COLING*), pp. 350–356.

Dolan, W. B. and Brockett, C. (2005). “Automatically constructing a corpus of sentential paraphrases.” In *Proceedings of the 3rd International Workshop on Paraphrasing* (*IWP*), pp. 9–16.

Dras, M. (1999). *Tree adjoining grammar and the reluctant paraphrasing of text*. Ph.D. thesis, Division of Information and Communication Science, Macquarie University.

藤田篤, 乾健太郎 (2003). “語彙・構文的言い換えにおける変換誤りの分析.” 情報処理学会論文誌, **44** (11), 2826–2838.

乾健太郎, 藤田篤 (2004). “言い換え技術に関する研究動向.” 自然言語処理, **11** (5), 151–198.

Jackendoff, R. (1990). *Semantic structures*. The MIT Press.

影山太郎 (1996). 動詞意味論—言語と認知の接点. くろしお出版.

影山太郎（編） (2001). 日英対照 動詞の意味と構文. 大修館書店.

金城由美子, 青野邦夫, 安田圭志, 竹澤寿幸, 菊井玄一郎 (2003). “旅行会話基本表現に対する日本語パラフレーズデータの収集.” 言語処理学会第 9 回年次大会発表論文集, pp. 101–104.

Lavoie, B., Kittredge, R., Korelsky, T., and Rambow, O. (2000). “A framework for MT and multilingual NLG systems based on uniform lexico-structural processing.” In *Proceedings of the 6th Applied Natural Language Processing Conference and the 1st Meeting of the North American Chapter of the Association for Computational Linguistics* (*ANLP-NAACL*), pp. 60–67.

Levin, B. (1993). *English verb classes and alternations: a preliminary investigation*. Chicago Press.

Mel’čuk, I. and Polguère, A. (1987). “A formal lexicon in meaning-text theory (or how to do lexica with words).” *Computational Linguistics*, **13** (3-4), pp. 261–275.

村木新次郎 (1991). 日本語動詞の諸相. ひつじ書房.

Quirk, C., Brockett, C., and Dolan, W. (2004). “Monolingual machine translation for paraphrase generation.” In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp. 142–149.

下畑光夫, 竹澤寿幸, 菊井玄一郎 (2003). “旅行会話における英語の同義表現コーパスの作成と分析.” 情報科学技術レターズ, pp. 83–85.

Shinyama, Y. and Sekine, S. (2003). “Paraphrase acquisition for information extraction.” In *Proceedings of the 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications* (*IWP*), pp. 65–71.

白井諭, 山本和英 (2001a). “換言事例の収集—日英基本構文を対象として.” 言語処理学会第 7 回年次大会発表論文集, pp. 401–404.

白井諭, 山本和英 (2001b). “換言事例の収集—機械翻訳における多様性確保の観点から.” 言語処理学会第 7 回年次大会ワークショップ論文集, pp. 3–8.

Takahashi, T., Iwakura, T., Iida, R., Fujita, A., and Inui, K. (2001). “KURA: a transfer-based lexico-structural paraphrasing engine.” In *Proceedings of the 6th Natural Language Pro-*

cessing Pacific Rim Symposium (*NLPRS*) *Workshop on Automatic Paraphrasing: Theories and Applications*, pp. 37–46.

Zhang, Y., Yamamoto, K., and Sakamoto, M. (2001). "Paraphrasing utterances by reordering words using semi-automatically acquired patterns." In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium* (*NLPRS*), pp. 195–202.

略歴

藤田 篤（正会員）： 2005 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．京都大学情報学研究科産学官連携研究員を経て，2006 年より名古屋大学大学院工学研究科助手．現在に至る．博士（工学）．自然言語処理，特にテキストの自動言い換えの研究に従事．情報処理学会，ACL 各会員．

乾 健太郎（正会員）： 1995 年東京工業大学大学院情報理工学研究科博士課程修了．同年より同研究科助手．1998 年より九州工業大学情報工学部助教授．1998 年〜2001 年科学技術振興事業団さきがけ研究 21 研究員を兼任．2001 年より奈良先端科学技術大学院大学情報科学研究科助教授．2004 年文部科学省長期在外研究員として英国サセックス大学に滞在．現在に至る．博士（工学）．自然言語処理の研究に従事．情報処理学会，人工知能学会，ACL 各会員．

(2005 年 12 月 26 日 受付)
(2006 年 3 月 9 日 再受付)
(2006 年 3 月 20 日 採録)