情報処理学会論文誌 Vol.56 No.3 983–992 (Mar. 2015)

語釈文を用いた小学生のための語彙平易化

梶原 智之[1,a)] 山本 和英[1,b)]

受付日 2014年6月30日, 採録日 2014年11月10日

概要：小学生の文章読解支援に向けた語彙平易化を目的として，国語辞典の語釈文から平易な語彙的換言を獲得する手法を提案する．国語辞典の語釈文は，見出し語を平易な語を用いて説明しており，見出し語から語釈文中の語への換言によって語彙の平易化が見込まれる．従来は主要部終端型である日本語の特徴を利用した語釈文末の語への換言が行われてきたが，我々は語釈文全体から見出し語と換言可能性のある候補を広く収集して換言する手法を提案する．換言候補から最終的な換言を選択する際には，文脈を考慮するよりもシソーラスに基づく語の類似度を用いた選択の効果が高いことを実験的に示す．

キーワード：語彙平易化，言い換え，読解支援，語彙制限

Japanese Lexical Simplification for Children Using Definition Statements

Tomoyuki Kajiwara[1,a)] Kazuhide Yamamoto[1,b)]

Received: June 30, 2014, Accepted: November 10, 2014

Abstract: We propose a method for acquiring lexical paraphrase using dictionaries in order to achieve lexical simplification for children. The definition statement of the headword in a dictionary explains the headword in simple vocabulary. The paraphrasing by replacing the headword with the most similar word in the definition is expected to be an accurate means of lexical simplification. Conventionally in the Japanese simplification task, the last probable word of the statements in the definition is applied as a paraphrase, for Japanese is a head-final language. Our method collects candidates widely from the whole definition statements. For the paraphrase-selection from candidates, the experimental results show that the word-similarity method using a thesaurus outperforms the context-aware method.

Keywords: lexical simplification, paraphrase, reading assistance, lexical restriction

1. はじめに

容易に大量かつ多様なテキストデータに触れることができる現代であるが，読者が情報収集を効果的に行うためには，読者間の言語能力の差を埋める技術が必要である．たとえば，子ども，高齢者，外国人などの言語学習者，障がい者に対して，言語能力の差を埋める技術は有用である [1]．我々は，このような言語能力の差を埋める課題の事例として，新聞記事に出現する難解な語を小学生向けに平易化する課題を取り上げる．学習途上である小学生は，理解できる語彙が少ないという言語能力の課題を有している．本論文では，新聞記事に出現するすべての語を，小学生が理解できる平易な語へ制限することにより，小学生向けの語彙平易化を実現する．

我々は語彙平易化のために，平易な語彙を定義し，難解な語を対応する平易な語に換言する．本論文では小学生を対象とするため，平易な語彙として小学生が十分に理解できる学習基本語彙 [2] を用いる．学習基本語彙とは，小学国語教科書などの語彙分析に基づいて選定された，小学生が言語表現に駆使できる 5,404 語である．以下，学習基本語彙に含まれない語を「難解語」，難解語を換言して得られる学習基本語彙に含まれる語を「平易語」と呼ぶ．2 章以

[1] 長岡技術科学大学電気系
Department of Electrical Engineering, Nagaoka University of Technology, Nagaoka, Niigata 940–2188, Japan
[a)] kajiwara@jnlp.org
[b)] yamamoto@jnlp.org

降では，難解語を平易語に換言する手法について述べる．

2. 関連研究

Web テキスト中から換言可能な表現を自動的に獲得する手法がいくつか提案されているが [3]，これらの品質はまだ不十分である．パラレルコーパスから換言を獲得する研究もさかんに行われている．Barzilay ら [4] は，同じ文書から作られた複数の英訳を用いて換言を獲得している．また，Shinyama ら [5] は，同じ報道を行っている複数の新聞記事を用いて換言を獲得している．テキスト平易化タスクにおいては，Coster ら [6] が English Wikipedia と Simple English Wikipedia を対応付けたパラレルコーパスを作成し，統計的機械翻訳の枠組みで平易化を行っている．これらのパラレルコーパスを用いる手法では，対応する表現どうしのアライメントの精度や利用可能なコーパスの量に課題がある．本論文で扱うのは語彙レベルの換言であり，既存の語彙資源を用いることで，アライメントの精度の問題を度外視することができる．

既存の語彙資源から換言可能な語の対を獲得する手法には，シソーラスを用いるものや国語辞典を用いるものがある．シソーラスを用いる手法は，概念間の階層の距離を計算して意味的な近さを測ることができる利点がある．平原ら [7] は，語彙的換言を用いてテキスト要約の評価を行っており，日本語 WordNet [8] などのシソーラスを用いて換言知識を収集している．一方，国語辞典は見出し語を語釈文で説明する語彙資源であり，一般的に語釈文は見出し語よりも平易な語を用いて説明されている．そのため，国語辞典を用いる手法には，平易な語を獲得できるという利点がある．本研究では平易な語への換言を目的とするため，国語辞典を用いた手法により平易語を獲得する．

藤田ら [9] や美野ら [10] は，名詞の見出し語に関する語釈文間の一致度や類似度を用いて，見出し語どうしの換言を行っている．ただし，美野らの報告にもあるように，これらの手法で得られる換言先の語は，換言元の語より平易なわけではない．本論文では語彙の平易化を目的としているため，見出し語よりも語釈文が平易と期待される国語辞典の特性を活かし，見出し語から語釈文への換言を行う．

鍛治ら [11] は「見出し語が用言であれば，その語釈文は用言を主辞とする形で記述されており，なおかつ主辞は語釈文の末尾に位置する」と仮定し，国語辞典を使った用言の換言手法を提案した．美野らはまた「語釈文における主要文の最終文節が，見出し語の意味を表している」と考え，国語辞典を使った体言の換言手法も提案している．梶原ら [12] は語釈文の最も後ろに現れる見出し語と同じ品詞の 1 語を換言先とし，国語辞典を用いて換言元の品詞を限定せずに換言を行った．これらの研究は，いずれも主要部終端型である日本語の特徴を利用して，語釈文の末尾から換言先の語を抽出している．

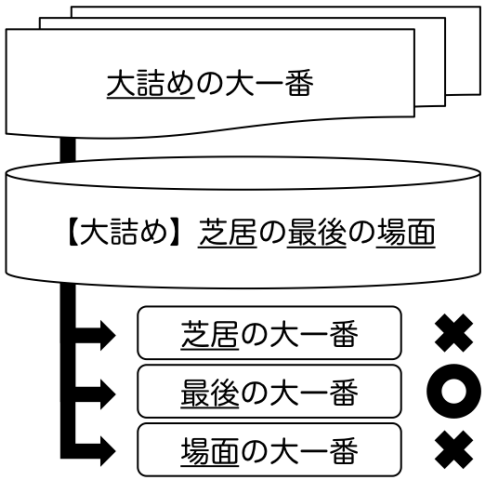

図 1 語釈文の末尾で換言できない例
Fig. 1 Example of a word that cannot be paraphrased as the end portion of the definition statement.

しかし，**図 1** に示すように，見出し語と換言可能な語は必ずしも語釈文の末尾に存在するわけではない．図 1 の例では，入力文「大詰めの大一番」に含まれる名詞の難解語「大詰め」を，語釈文「芝居の最後の場面」中の名詞「芝居」「最後」「場面」にそれぞれ換言している．この例では，語釈文の文末に出現する「場面」には換言することができない．しかし，語釈文の途中に出現する「最後」に換言することで「最後の大一番」という文が得られ，換言の前後で意味を保持することができる．そこで本論文では，語釈文全体を用いて換言可能性のある語を広く収集する手法を提案する．語釈文のすべてを扱うことによって，複数の換言先候補を得ることができる．これにともない，複数の選択肢の中から最適な換言を選択する処理が必要となる．適切な語の選択については，シソーラスから得られる意味の類似度を用いる Ma らの手法 [13] や，分布仮説 [14] に基づいて大規模コーパスから得られる統計的な情報を用いる Lapata ら [15] や Keller ら [16] の手法がある．シソーラスは意味のまとまりで単語を階層的に分類した語彙資源であり，シソーラス中での単語間の距離を測ることで，単語間の意味の近さを測ることができる．また，分布仮説によると，意味の似た語は似た文脈で用いられる．この仮説に基づき，Lapata らや Keller らは共起頻度や n-gram を用いて，表現の適格さを判定することができると報告している．本論文では，シソーラスに基づく語の類似度を用いる指標と，共起頻度や n-gram 頻度などの文脈を考慮する指標を比較し，換言の選択手法を検討する．

3. 換言候補の収集

この章では，国語辞典の語釈文の中から，見出し語と換言できる可能性のある語を収集する方法を説明する．国語辞典の語釈文を用いることで，平易な語彙的換言を網羅的に収集することができる．

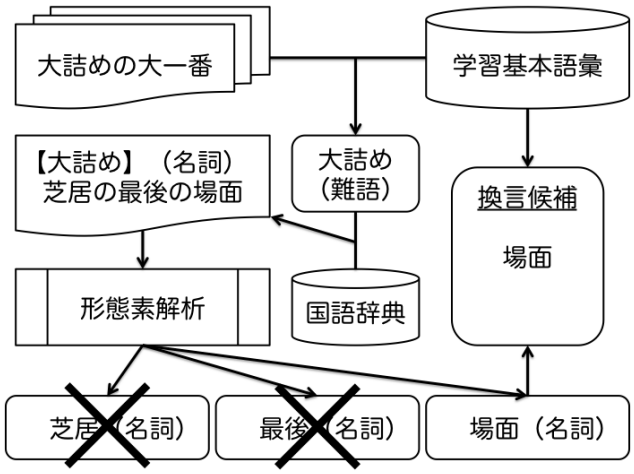

図 2 従来手法における換言候補の収集
Fig. 2 Collection of paraphrase candidates in the conventional method.

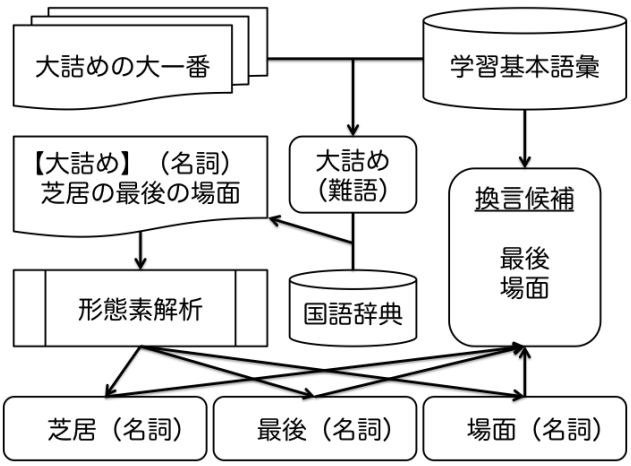

図 3 提案手法における換言候補の収集
Fig. 3 Collection of paraphrase candidates in the proposed method.

3.1 従来手法

2 章でも述べたように，国語辞典の語釈文から見出し語の換言を獲得する従来の研究 [10], [11], [12] では，主要部終端型である日本語の特徴を利用して，語釈文の末尾から換言先の語を抽出している．たとえば梶原ら [12] は，次の手順で換言候補の語を抽出する（**図 2**）．

(1) 入力文から難解語を検出．
(2) 難解語を見出し語として国語辞典から語釈文を抽出．
(3) 語釈文を形態素解析．
(4) 難解語と同じ品詞の語のうち最も文末の近くに出現する語を収集．
(5) 収集した語から難解語を取り除き平易語のみを残す．

こうして得られた平易語が，従来手法における換言候補となる．図 2 において，「芝居」「最後」の 2 語は，難解語「大詰め」と同じ品詞ではあるが，最も文末の近くに出現する語ではないので，換言候補にはならない．

3.2 提案手法

図 1 に示したように，見出し語と換言可能な語は必ずしも語釈文の末尾に存在するわけではない．そこで我々は，語釈文全体を用いて換言可能性のある語を広く収集する手法を提案する．提案手法では，次の手順で換言候補の語を抽出する（**図 3**）．

(1) 入力文から難解語を検出．
(2) 難解語を見出し語として国語辞典から語釈文を抽出．
(3) 語釈文を形態素解析．
(4) 難解語と同じ品詞の語をすべて収集．
(5) 収集した語から難解語を取り除き平易語のみを残す．

こうして得られた平易語集合が，提案手法における換言候補となる．提案手法の従来手法との違いは，(4) の処理において文末の 1 語に限定せず，難解語と同じ品詞の語をすべて収集することである．これにより，図 3 の例では，文末ではない「最後」という語が換言候補に含まれるようになり，従来手法では扱えなかった図 1 のような換言ができるようになる．また，図 3 において「芝居」は，学習基本語彙に含まれないため換言候補にはならない．このような語は，再び語釈文を参照することによって平易に置き換えられる可能性がある．しかし，梶原ら [12] の報告にもあるように，完全に同義な換言対の存在は稀で，2 度以上の再帰的な処理によって換言可能な語が得られることは少ない．そのため，本論文では語釈文を 1 度だけ参照する．

3.3 難解語と同じ品詞の平易語

換言候補の収集に際して，初めに所与の入力文から難解語を検出するが，ここでは難解語を内容語に限定する．内容語とは，名詞，動詞，形容詞および副詞である．また，得られる換言候補語は，難解語と同じ品詞の平易語である．ただし，動詞にはサ変動詞（サ変名詞 + する）を含む．そのほか，形容詞の連用形と副詞，形容詞の連体形と「名詞 + の」なども文法上同じ働きをすると考えられるが，本論文ではこれらには対応していない．

また，同じ品詞の語の中には，反義語や否定語「ない」をともなって反対の意味を表す語も含まれる．反義語や否定語が用いられる文脈は，元の語が用いられる文脈と似ていることが実験的に知られている [17]．ただし，否定語「ない」を含む語釈文は約 4%と少ない（EDR 日本語単語辞書 [18]）．そのため，本論文では反義語や否定語に対して特別な処理は行っていない．

3.4 複数の語釈文

見出し語に対応する語釈文は語義の数だけ存在するため，ある見出し語が多義である場合，その見出し語には複数の語釈文が対応する．また，1 つの語義の中にも複数の語釈文が含まれている場合もある．このように，ある見出

し語に複数の語釈文が対応している場合，それらすべての語釈文から換言候補の収集を行う．

4. 換言候補の選択

この章では，3 章で収集した換言候補の中から，適切な換言を選択する方法を説明する．本論文では，シソーラスに基づく語の類似度を用いる指標と，共起頻度などの文脈を考慮する指標を比較し，換言の選択手法を検討する．

4.1 シソーラスに基づく語の類似度を用いる指標

換言の前後で意味をできるだけ保つために，収集した換言候補の中から，難解語との意味の類似度が最も高い平易語を選択する．類似度の計算には，日本語 WordNet を用いる．WordNet は，同義語の集合が階層的に分類されている言語資源であり，式 (1) に示すように，2 つの単語が属する同義語集合の深さおよび共通上位概念の深さを用いて意味の類似度を計算することができる．類似度は，単語 w の属する同義語集合の深さを $d(w)$，単語 w と難解語 n の共通の上位概念の深さを $d_c(w,n)$ として，式 (1) で定義される．類似度が最も高い平易語が複数存在する場合は，類似度が最も高い平易語の中からランダムに 1 つを選択することとする．

$$score_{sim}(w,n) = \frac{2d_c(w,n)}{d(w)+d(n)} \tag{1}$$

4.2 語釈文中での出現頻度を用いる指標

より多くの種類の国語辞典から同じ平易語が換言候補として収集されるほど，その平易語が難解語の換言として適切であると考えることができる．そこで，単語 w の換言候補中での出現頻度 $freq\_candidate(w)$ を用いた式 (2) のスコアを定義する．

$$score_{freq}(w) = freq\_candidate(w) \tag{2}$$

4.3 共起頻度を用いる指標

入力文中の内容語とよく共起する平易語は，難解語の換言として適切であると考えることができる．そこで，換言候補中の単語 w と入力文中の内容語 $c \in C$ のコーパス中での共起頻度 $co\_freq(w,c)$ を用いた式 (3) のスコアを定義する．ここで，C は入力文に含まれる内容語の集合である．

$$score_{co\_freq}(w,C) = \sum_{c \in C} co\_freq(w,c) \tag{3}$$

4.4 自己相互情報量を用いる指標

式 (3) では共起頻度を用いた指標を定義したが，共起頻度は単語単体での出現頻度から大きく影響を受け，出現頻度の高い単語ほど共起頻度も高くなる場合が多い．これに対して，自己相互情報量は単語単体での出現頻度の影響を差し引いて共起性を測ることができる．そこで，換言候補中の単語 w と入力文中の内容語 $c \in C$ のコーパス中での自己相互情報量を用いた式 (4) のスコアを定義する．ここで，C は入力文に含まれる内容語の集合である．また，$freq(w)$ は単語 w のコーパス中での出現頻度，$co\_freq(w,c)$ はコーパス中での単語 w と単語 c の共起頻度を表す．

$$score_{pmi}(w,C) = \sum_{c \in C} \log \frac{co\_freq(w,c)}{freq(w)freq(c)} \tag{4}$$

4.5 単語 3-gram 頻度を用いる指標

入力文中の難解語を，換言候補中の単語 w と置き換える．このときの単語 w の周囲の語を用いて，コーパス中での単語 3-gram 出現頻度を求める．この単語 3-gram 出現頻度が高い単語ほど，難解語の換言として適切であると考えることができる．そこで，単語 w で終わる 3-gram，単語 w が中央にくる 3-gram，単語 w から始まる 3-gram の 3 つの単語 3-gram 出現頻度の総和を用いた式 (5) のスコアを定義する．ここで，入力文中の各語を難解語からの相対位置 i を用いて $x_i \in X$ で表す．また，$freq_{3\text{-}gram}(x_{-1},w,x_{+1})$ は x_{-1}，w，x_{+1} という単語 3-gram のコーパス中での出現頻度を表す．

$$\begin{aligned} score_{3\text{-}gram}(w,X) = {} & freq_{3\text{-}gram}(x_{-2},x_{-1},w) \\ & + freq_{3\text{-}gram}(x_{-1},w,x_{+1}) \\ & + freq_{3\text{-}gram}(w,x_{+1},x_{+2}) \end{aligned} \tag{5}$$

4.6 共起頻度ベクトルの類似度を用いる指標

難解語 n と入力文中の他の内容語 $c \in C$ の共起頻度ベクトルと換言候補中の単語 w と入力文中の難解語を除く内容語の共起頻度ベクトルの類似度を計算する．この類似度が高い単語ほど，難解語と似た文脈で用いられることを表しており，難解語の換言として適切であると考えることができる．ここでは，ベクトルの類似度として余弦類似度を用いた式 (6) のスコアを定義する．ここで，$vector(w,C)$ は単語 w と入力文中の内容語集合 C に含まれる各内容語 c のコーパス中での共起頻度ベクトルである．

$$score_{vec}(w,n,C) = \frac{vector(w,C) \cdot vector(n,C)}{|vector(w,C)||vector(n,C)|} \tag{6}$$

5. 実験

5.1 使用したツールおよび語彙資源

3.2 節でも述べたが，完全に同義な換言対の存在は稀である．そのため，ある文に含まれる複数の難解語を同時に平易化しようとしたとき，個々の平易化が妥当な変換であっても文全体ではニュアンスが異なるという場合も考えられる．そこで本論文では，難解語が 1 語だけ含まれる文を新聞記事から抽出して実験を行った．このような文は，毎日新聞 2000 年度版 [19] に 14,344 文（全 232,038 文）含まれる．毎日新聞 2000 年度版には全 26,709 種類の難解語が含

まれるが，実験に用いるのは 50 回以上出現する 221 語とする．この 221 語のうち，語釈文中に平易語が含まれない 56 語は換言対象から除いた．さらにすべての換言候補が換言可能であり，選択の必要がない 13 語も換言対象から除き，最終的に 152 種類の難解語を実験で使用した．なお，この 152 種類の難解語は毎日新聞 2000 年度版に延べ 72,153 回出現している．

本実験では 1 種類の難解語につき 1 文ずつ実験対象の文を用意した．これらの文は，毎日新聞 2000 年度版の難解語を 1 語だけ含む 14,344 文から無作為に抽出した．我々は，152 語の難解語につき対象とした各 1 文が無作為である限り，難解語 1 語で複数文を対象とする場合の実験結果と同様の傾向が得られると考え，このような実験設定とした．

換言の網羅性を高めるため，換言知識として複数の国語辞典を併用した．本論文では，EDR 日本語単語辞書 [18]，チャレンジ小学国語辞典 [20]，三省堂国語辞典 [21] の 3 種類の国語辞典を用いた．

3 章で説明した換言候補の収集では，形態素解析に IPADIC(2.7.0) および MeCab(0.993) [22] を用いた．

4 章で説明した換言候補の選択では，単語の出現頻度や共起頻度などを計算するコーパスとして Web 日本語 N グラム [23] を用いた．Web 日本語 N グラムとは，一般に公開されている日本語の Web ページ約 200 億文から抽出された出現頻度 20 回以上の単語 1-gram から単語 7-gram である．本論文では，最も長い単語 7-gram データから単語出現頻度や共起頻度を求めた．出現頻度については 1-gram データをそのまま使うこともできるが，式 (4) の計算時に分母と分子の母集団を揃えるために 7-gram データを用いて求めた．ただし，単語 3-gram 出現頻度については 3-gram データから求めた．なお，3-gram データの異なり数は 394,482,216，7-gram データの異なり数は 570,204,252 である．

5.2 実験方法

まず，3 章に示した提案手法および従来手法で換言候補の収集を行った．そして，提案手法で収集した換言候補に対して，4 章に示した 6 つの指標で換言候補の選択を行った．

評価は，日本語を母語とする工学系大学院生 3 人が行った．3 人中 2 人以上が，選択した平易語が原文において難解語と換言可能であると判断した場合を正解とした．評価者 2 人ずつの kappa 係数を**表 1** に示す．kappa 係数はすべて 0.6 以上であり，評価者間の評価の一致度は十分高い．

5.3 実験結果

5.3.1 換言候補の収集

表 2 に換言候補の収集の実験結果を示す．「得られた換言候補の数」とは，提案手法または従来手法を用いて 152 語の難解語から収集した換言候補の延べ数である．「換言可能な候補の数」とは，収集された換言候補の中で難解語と換言可能な語の延べ数である．たとえば，難解語「サポート」は候補「支える」および候補「助ける」の両方と換言可能である．「換言可能な難解語の数」とは，換言可能な候補によって換言できる難解語の延べ数である．提案手法では従来の文末から換言候補を収集する手法よりも多くの換言候補を収集することができ，従来手法では 64%（難解語 152 語のうち 98 語）のところ，提案手法では 84%（難解語 152 語のうち 128 語）の難解語の平易な換言を収集することができる．

表 1 評価者 2 人ずつの kappa 係数
Table 1 Kappa coefficient of each two evaluators.

評価者	A and B	B and C	C and A
kappa 係数	0.617	0.600	0.662

表 2 換言候補の収集
Table 2 Collection of candidates.

	従来手法	提案手法
得られた換言候補の数	320	1076
換言可能な候補の数	141	239
換言可能な難解語の数	98 (64%)	128 (84%)

表 3 換言候補の選択
Table 3 Selection of candidates.

選択の指標	選択の正解数	選択の正解率
WordNet 類似度	86	67%
出現頻度	54	42%
共起頻度	61	48%
自己相互情報量	79	62%
3-gram 頻度	74	58%
共起頻度ベクトル	74	58%

本論文では，国語辞典の見出し語を語釈文中の語に換言することで語彙の平易化ができるという考えのもと，図 1 のような従来の語釈文末の語に換言するだけでは扱えない換言について，語釈文中の他の語を用いて換言することを提案した．この実験結果は，語釈文が文末以外にも見出し語と換言可能な語を多く含んでいることを示すものであり，提案手法がより網羅性の高い換言対の収集のために効果的であることを示している．

5.3.2 換言候補の選択

表 3 に換言候補の選択の実験結果を示す．ここでは，収集した換言候補中に換言可能な平易語が含まれていないものを除き，表 2 で示した換言可能な 128 の難解語についてのみ候補の選択を行った．なお，128 の難解語に対して，合計 784 語の換言候補が収集されており，ランダムに選択を行った場合の正解率は 30%である．

表 3 より，文脈を考慮する他の指標に比べて，シソーラスに基づく語の類似度を用いる指標で選択した場合の正解率が高いことが分かる．これは，比較的単純な方法で文脈を考慮しても効果は限定的であり，難解語との意味の類似

度が高い語を選択することが有効であることを示している．

6. 考察

6.1 換言知識の特性

表 4 に，難解語 1 語あたりの平均語釈文数（語義数），語釈文 1 文あたりの平均文長（内容語数），語釈文 1 文あたりの平均平易語数を，辞書ごとに示す．一般向けの EDR 日本語単語辞書や三省堂国語辞典は見出し語 1 語あたりの語義数が多く，小学生向けのチャレンジ小学国語辞典は 1 語ごとの語義数が少ないことが分かる．一方でチャレンジ小学国語辞典は，語釈文が長く，平易語を多く含んでいる．これは，語釈文でより平易な説明を行うために，平易な語を積極的に使用して詳細な説明をしているためである．平易な語を用いて詳細に説明するというのは，Simple English Wikipedia[*1]の書き方のガイドラインにも記載されている内容であり，平易な文章の書き方の 1 つの特徴である．

次に，表 2 の結果を辞書ごとに見る（**表 5**）．EDR 日本語単語辞書と三省堂国語辞典では，従来手法に比べて提案手法で 3.1 倍多くの換言候補を収集することができている．また，換言可能な候補も従来手法より 1.6 倍多く収集することができている．これに対して，チャレンジ小学国語辞典は換言候補を従来手法の 4.2 倍，換言可能な候補を従来手法の 1.9 倍多く収集することができており，提案手法でより効果的に換言候補を収集することができている．これは表 4 にも示したとおり，小学生向けの国語辞典が語釈文中により多くの平易語を含むためであり，小学生を対象とする本論文の語彙平易化のためには小学国語辞典を用いるのが有効であるといえる．ただし，表 4 にも示したとおり，収録されている見出し語の総数が少ないのは小学生向けの国語辞典の弱みであり，小学国語辞典単体での換言では網羅性は低い．EDR 日本語単語辞書など一般向けの国語辞典は，見出し語数も語義数も多いので，網羅性の高い換言候補の収集のためには，やはり換言知識を組み合わせることは重要である．

続いて，表 2 で換言可能な候補を収集できなかった難解語について，学習基本語彙のいずれの語とも換言ができないのか，換言知識に用いた辞書のために候補が列挙できなかったのかについて調査する．語釈文中に平易な語が含まれないとして表 2 の実験に用いなかった 56 語および提案手法で換言可能な候補の収集に失敗した 24 語の計 80 の難解語について調査を行う．5.2 節の評価方法に従い，学習基本語彙中に換言可能な語が含まれているか否かについて評価を行ったところ，80 語のうち 50 語は換言可能な平易語が存在し，30 語は学習基本語彙のいずれの語を用いても換言できないことが分かった．平易語が存在する 50 語には，難解語「廃棄する」に対しての平易語「捨てる」など，語釈文に含まれていてもおかしくない例があり，換言知識をいっそう充実させる必要がある．平易語が存在しない 30 語には，難解語「後続」など「後に続く」のように句単位で換言すべき例がある．句単位での換言は今後の課題である．

6.2 換言候補の収集

表 6 に，語釈文の文末以外の語を用いた換言の例を示す．これらの例では，語釈文の末尾に出現する難解語と同じ品詞の語は難解語と換言不可能であり，文中の他の難解語と同じ品詞の語が難解語と換言可能である．「継続」と「冒頭」の例に注目されたい．難解語「継続」はサ変名詞であり，語釈文末は「行う」という動詞であるので，これは換言候補である．しかし，この例では「行う」の手前に出現している「続ける」が難解語の換言として適切である．また，難解語「冒頭」は名詞であり，語釈文末は「部分」という名詞であるので，これは換言候補である．しかし，この例では「部分」の手前に出現している「初め」が難解語の換言として適切である．これらの「行うこと」や「～の部分」という表現は，語釈文の中で意味の中心となるような表現ではなく，換言処理においては不要な表現である．このような不要な表現は，「行うこと」や「～の部分」のほかにも「～すること」「～の一種」「～の一つ」など多数存在し，語釈文の末尾からのみ換言を抽出する従来手法では，これらの不要な表現を獲得して換言として適切な他の

表 4 辞書ごとの基礎データ
Table 4 Basic data for each dictionary.

	チャレンジ	EDR	三省堂
見出し語総数	33,700	270,000	73,000
平均語釈文数（語義数）	1.79	4.74	2.28
平均文長（内容語数）	4.10	3.83	3.12
平均平易語数	3.39	2.55	2.43

表 5 辞書ごとの換言候補の収集
Table 5 Collection of candidates for each dictionary.

	従来手法			提案手法		
	チャレンジ	EDR	三省堂	チャレンジ	EDR	三省堂
得られた換言候補の数	79	232	117	334	712	358
換言可能な候補の数	40	106	55	77	165	90
換言可能な難解語の数	34	76	41	49	102	60

[*1] http://simple.wikipedia.org/wiki/Main_Page

表 6 語釈文の文末以外の語を用いた換言例
Table 6 Example of paraphrase using not involving end of definition statement.

難解語	語釈文	平易語	換言例
継続	前からやっていることを，続けて行うこと	続ける	話し合いを { 継続し/続け } ていく
出演	映画・放送・舞台に出て，演じること	出る	日本の映画に { 出演する/出る }
警戒	よくないことが起きないように注意し，用心すること	注意	{ 警戒/注意 } は厳重
確信	かたく信じて動かない心	信じる	「自分たちが進む道はこれだ」と { 確信し/信じ } た
講師	大学教師の職の名の一つ	教師	{ 講師/教師 } と朗読作品は次の通り
収益	利益を手に入れること	利益	{ 収益/利益 } は寄付
悲劇	悲しい出来事や不幸な人生を題材にした劇や映画	不幸	私にとっては大きな { 悲劇/不幸 } だ
冒頭	文章や談話の初めの部分	初め	{ 冒頭/初め } いきなり次のような場面からはじまる
国籍	その国の国民であるという資格や身分	国	{ 国籍/国 } や民族の違いは関係ない

語を収集できない．

6.3 各選択指標の特徴

換言候補の選択に関する表 3 の実験結果を考察する前に，各指標の特徴について考える．語彙的換言を選択するこのタスクでは，単語間の意味の近さを適切に測定することが重要である．4 章に示した各指標は，いずれもこの単語間の意味の近さを測定するものである．

語釈文中での出現頻度を用いる指標では，難解語の説明に欠かせない語は多くの語釈文の中に出現するであろうというヒューリスティクスに基づいて換言可能性を計算している．しかし，6.1 節で述べたように，実際には見出し語の平易な換言が語釈文中に含まれるという保証はない．語釈文中に換言が含まれていたとしても，その語が平易語だとは限らない．たとえば，6.1 節であげた難解語「廃棄する」の例では，換言可能な「捨て去る」という語は語釈文中に出現するが，この語は平易語ではない．また，語釈の方法はひととおりではないため，「捨てる」や「捨て去る」という語以外にも「取去る」という語で難解語「廃棄する」の語釈をしている国語辞典もある．このような理由で，見出し語の換言が語釈文中で必ずしも高頻度で出現するという保証はない．

共起頻度や自己相互情報量，単語 3-gram 頻度を用いる指標では，入力文と換言候補の関係を扱っている．これは，難解語と換言候補の関係を間接的に扱っていると考えることができる．難解語「支援」を含む入力文「政党の支援なんて必要ない」を考える．難解語「支援」と換言可能な候補「助け」は，他の候補「貸し」や「はげまし」と比べて入力文中の語との共起頻度が高く，これらの指標では適切に「助け」が換言として選択される．難解語と換言可能な候補が，入力文との馴染みが良いのは，たしかであろう．一方で，難解語「騒ぎ」を含む入力文「それが騒ぎを大きくした」を考える．これらの指標で選択される候補は「問題」である．この候補で難解語を置換した文「それが問題を大きくした」は，この置換後の文だけを見ると自然な日本語の文である．文中の語と換言候補「問題」の共起頻度も高い．しかし，この候補「問題」は難解語「騒ぎ」とは厳密には換言の関係にはない．難解語「騒ぎ」は候補「問題」を含意するが，「問題」には「騒ぐ」「騒がしい」といった意味は含まれていないからである．また，難解語「騒ぎ」の換言候補には「秩序」も含まれる．換言不可能と評価されたこの候補は，入力文中の語との共起頻度は低い．入力文にそぐわない換言候補は難解語の換言になりえないが，入力文との馴染みが良い語が必ずしも難解語と換言可能なわけではない．つまり，共起頻度などで計算される入力文と換言候補の関係は，換言の必要条件にはなりうるが十分条件ではない．

共起頻度ベクトルの類似度を用いる指標は，難解語の文脈と換言候補の文脈を比較することで，難解語と換言候補の関係を間接的に扱うことができる．分布仮説によると意味の似た語は似た文脈で用いられるので，文脈の類似度が高い語を選択するこの指標には意味の近い語を選択することができるという期待がある．しかし，この指標も共起頻度に基づく指標であり，換言候補自身の出現頻度に影響を受ける傾向がある．先の難解語「騒ぎ」の例では，この指標は換言不可能な候補「面倒」を選択する．候補「面倒」の出現頻度は，換言可能な候補「騒動」の出現頻度の約 4 倍高い．また，難解語「概要」の例を考える．入力文は「インタビューの概要は次の通り」であり，換言可能な候補は「あらまし」である．この指標が選択する換言不可能な候補「物語」の出現頻度は，「あらまし」の出現頻度の約 30 倍高い．この指標では，出現頻度に大きな差がある場合，出現頻度の高い候補を選択してしまう傾向がある．

シソーラスに基づく語の類似度を用いる指標では，難解語と換言候補の関係を直接的に扱うことができる．これは，シソーラスが意味のまとまりで単語を分類した語彙資源だからである．難解語と換言候補の意味の近さを直接計算することができるのがこの指標の利点であるが，6 種類の指標の中で唯一換言候補の出現頻度に影響を受けないという点も特徴である．この指標では，難解語「騒ぎ」の例も難解語「概要」の例も，換言候補の出現頻度に影響されずに，難解語と意味の近い換言可能な候補を選択できる．

表 7 多義語の換言候補の選択
Table 7 Selection of candidate of the multisense word.

	選択の正解数	選択の正解率
WordNet 類似度	55	71%
出現頻度	34	44%
共起頻度	37	47%
自己相互情報量	50	64%
3-gram 頻度	47	60%
共起頻度ベクトル	43	55%

6.4 換言候補の選択

表 3 の実験結果では，文脈を考慮する他の指標に比べて，シソーラスに基づく語の類似度を用いる指標で選択した場合の正解率が高かった．WordNet の同義語（同じ synset に定義されている語）や上位語を換言として用いることは，平原ら [7] の研究でも見られ，この指標を用いて一定の精度で適切な換言が選択できるのは妥当な結果である．しかし，WordNet での単語間の距離が小さくても，必ず換言可能であるというわけではない．たとえば，日本語 WordNet 同義語データベース[*2]は WordNet の同じ synset に属する語対から同義関係にある対を人手で選別しており，単語間の距離が最短であってもすべての対が換言可能なわけではないということを示している．本論文でも，距離が同一の語が複数ある場合には，その中からランダムに 1 語選択しており，この部分の選択方法の改良により，正解率のさらなる向上が期待される．

次に，シソーラスに基づく語の類似度を用いる指標で選択に失敗した例を考える．選択に失敗した 42（128 − 86）語のうち，最も多い 67%にあたる 28 語は難解語が多義語であり，多義性が解消できなかったことによる失敗であった．たとえば，難解語「ネット」は多義語であり，「網」「ネットワーク」「インターネット」「正味の量」などの語義がある．自己相互情報量を用いる指標などでは適切な候補「インターネット」を選択することができており，「今のところ，ネットでできるのは申し込みだけ.」という入力文を用いると比較的簡単に多義性の解消が可能な例である．しかし，WordNet では「網」と「インターネット」の両方が難解語「ネット」と同じ synset に定義されており，多義性を解消せずに選択することは難しい．

しかし，入力文の文脈を考慮する他の指標でも，多義性を解消する効果は限定的であった．**表 7** で，いずれの国語辞典でも語義が複数存在する語を多義語と定義して，多義語の選択の正解率を比較する．このような多義の難解語は，78 語（128 語中の 61%）存在した．表 7 より，入力文の文脈を考慮する指標よりも，むしろ文脈を考慮しないシソーラスに基づく語の類似度を用いる指標の方が多義語の換言の正解率が高いことが分かる．文脈を考慮した指標では，共起頻度などの統計量を近似的に語の類似度と見ているが，これらの指標ではシソーラスの距離を用いる指標ほど正確には語の類似度を測定できていないと考える．表 3 で選択に失敗した難解語と平易語の対を見ると，シソーラスに基づく指標では 55%が「少なくともある文脈では換言可能」な語であるが，他の指標では「少なくともある文脈では換言可能」な語が 22%から 33%と少ない（**表 8**）．

[*2] http://nlpwww.nict.go.jp/wn-ja/

表 8 少なくともある文脈では換言可能な難解語平易語対
Table 8 The number of word pair that can paraphrase in a certain context at least.

	入力文脈で換言不可能	ある文脈では換言可能
WordNet 類似度	42	23 (55%)
出現頻度	74	19 (26%)
共起頻度	67	15 (22%)
自己相互情報量	49	11 (22%)
3-gram 頻度	54	13 (24%)
共起頻度ベクトル	54	18 (33%)

6.3 節で述べたように，シソーラスに基づく指標は，難解語と換言候補の意味の近さを直接計算することができる点と換言候補の出現頻度に影響を受けない点で，他の指標よりも適切に換言を選択することができる．シソーラスに基づく指標では入力文の文脈によって多義性を解消して換言を選択することはできない．しかし，難解語と換言候補の意味の近さを直接計算できるこの指標では，少なくともある文脈では換言可能な語を選択しやすいため，多義性の解消はできなくても結果としては他の指標よりも適切な換言を選択できる場合が多い．一方，入力文の文脈を考慮する自己相互情報量などに基づく指標では，難解語と換言候補の意味の近さを間接的にしか計算できない．そのため，「少なくともある文脈では」という条件を付けても換言できない候補を選択してしまう場合が，シソーラスに基づく指標に比べて多くなってしまっている．

多義語の換言の選択においては，まず多義性の解消を行い語釈文を選択したうえで，語釈文の中から見出し語の換言を選択するという 2 段階の処理が必要である．選択の実験を行った 128 語中の 61%が多義語であり，最も選択の正解率が高いシソーラスに基づく指標で選択に失敗した 42 語中の 67%も多義性の解消が課題であるため，適切な換言の選択のためには多義性の影響は十分に大きい．今後は多義性を解消したうえで高精度に換言を選択できる方法を検討する必要がある．

6.5 選択指標の組合せ

試みに，本論文で扱った 6 つの指標の投票により 128 語の難解語の換言候補を選択した（**表 9**）．ここで，選択方法 1 は，表 3 に示した 6 つの選択指標で 1 語ずつ換言候補を選択し，最も多くの指標が選択した候補を選択する方法である．1 つの候補に絞ることができない場合には，最も多くの指標が選択した候補の中からランダムに 1 語を選択

表 9 6 指標を組み合わせた選択
Table 9 Selection combined six indicators.

		選択の正解数
1	6 つの指標で投票	85 (66%)
2	票が割れた場合に WordNet 指標を優先	88 (69%)
3	過半数を超えない場合に WordNet 指標を優先	90 (70%)

する．選択方法 2 は，同様に 6 つの指標で 1 語ずつ換言候補を選択するが，1 つの候補に絞ることができない場合にシソーラスに基づく指標の選択を優先する方法である．これは，表 3 においてシソーラスに基づく指標の正解率が最も高かったためである．選択方法 3 も同様に換言候補を選択するが，過半数である 4 つ以上の指標の選択が一致しなかった場合にシソーラスに基づく指標の選択を優先する方法である．この方法が最も正解率が高く，70%の正解率で適切な語彙的換言を選択することができた．最も正解率の高いシソーラスに基づく指標を優先しつつ，文脈を考慮する指標を組み合わせて選択の正解率を向上させた．

7. おわりに

本論文では，小学生のための文章読解支援に向けた語彙平易化を目的として，国語辞典を用いた語彙的換言を行い，新聞記事の語彙を小学生が十分理解可能な学習基本語彙に制限した．本論文の主たる主張は，次の 2 つである．

(1) 換言候補は，語釈文の末尾に限定せず，語釈文の全体から広く収集すべきである．
(2) 換言候補から適切な換言を選択する際には，比較的簡単な方法で文脈を考慮しても効果は限定的であり，シソーラスに基づく類似度を優先して選択するのが良い．

従来は語釈文の末尾の語へ換言が行われてきたが，表 2 に示したように，語釈文には末尾以外にも見出し語の換言が多く含まれている．そのため，語釈文全体から文法的に同じ働きをする換言候補を広く収集することで，従来手法よりも 20 ポイント多い 84%の難解語の平易な換言を収集することができた．

換言の選択は，シソーラスに基づく語の類似度を用いる指標の正解率が 67%で最も高く，文脈を考慮する指標の中では入力文中の内容語と平易語の自己相互情報量を用いる指標の 62%の正解率が最も高かった．そして，シソーラスに基づく指標を優先しつつ文脈を考慮する指標を組み合わせることで，選択の正解率は 70%まで向上した．

国語辞典の語釈文を最大限に活かして語彙的換言を獲得するためには，本論文で提案したように換言候補を広く収集するとともに，より高精度で換言候補から選択を行うことが望まれる．本論文の主たる目的が多義性の解消ではなかったため，今回の実験では難解語 1 語に対して 1 文の換言しか行っていない．多くの難解語は多義性を持つが，シソーラスに基づく指標のみでは多義性の問題を解決することは難しいため，今後は入力文の文脈を有効に活用したより高精度の選択手法を確立し，多様な文脈中での語彙の平易化を実現したい．

参考文献

[1] 乾健太郎，藤田 篤：言い換え技術に関する研究動向，自然言語処理，Vol.11, No.5, pp.151–198 (2004).
[2] 甲斐睦朗，松川利広：語彙指導の方法：語彙表編，光村図書出版株式会社 (2002).
[3] Yamamoto, K.: Acquisition of Lexical Paraphrases from Texts, *Proc. 2nd International Workshop on Computational Terminology* (*Computerm*) (2002).
[4] Barzilay, R. and McKeown, K.R.: Extracting Paraphrases from a Parallel Corpus, *Proc. 39th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.50–57 (2001).
[5] Shinyama, Y. and Sekine, S.: Paraphrase Acquisition for Information Extraction, *Proc. 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications* (*IWP*), pp.65–71 (2003).
[6] Coster, W. and Kauchak, D.: Simple Wikipedia: A New Simplification Task, *Proc. 49th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.665–669 (2011).
[7] 平原一帆，難波英嗣，竹澤寿幸，奥村 学：言い換えを用いたテキスト要約の自動評価，情報処理学会論文誌データベース（TOD），Vol.3, No.2, pp.91–101 (2010).
[8] Isahara, H., Bond, F., Uchimoto, K., Utiyama, M. and Kanzaki, K.: Development of the Japanese WordNet, *Proc. 6th International Conference on Language Resources and Evaluation* (*LREC*) (2008).
[9] 藤田 篤，乾健太郎，乾 裕子：名詞言い換えコーパスの作成環境，電子情報通信学会思考と言語研究会予稿集，pp.53–60 (2000).
[10] 美野秀弥，田中英輝：国語辞典を使った放送ニュースの名詞の平易化，言語処理学会第 16 回年次大会，pp.760–763 (2010).
[11] 鍛治伸裕，河原大輔，黒橋禎夫，佐藤理史：格フレームの対応付けに基づく用言の言い換え，自然言語処理，Vol.10, No.4, pp.65–81 (2003).
[12] 梶原智之，山本和英：小学生の読解支援に向けた複数の換言知識を併用した語彙平易化と評価，言語処理学会第 19 回年次大会，pp.272–275 (2013).
[13] Ma, X., Fellbaum, C. and Cook, P.R.: A Multimodal Vocabulary for Augmentative and Alternative Communication from Sound/Image Label Datasets, *Proc. NAACL HLT 2010 Workshop on Speech and Language Processing for Assistive Technologies*, pp.62–70 (2010).
[14] Harris, Z.: Distributional Structure, *Word*, Vol.10, No.23, pp.146–162 (1954).
[15] Lapata, M., Keller, F. and McDonald, S.: Evaluating Smoothing Algorithms against Plausibility Judgements, *Proc. 39th Annual Meeting of the Association for Computational Linguistics* (*ACL*), pp.346–353 (2001).
[16] Keller, F., Lapata, M. and Ourioupina, O.: Using the Web to Overcome Data Sparseness, *Proc. Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp.230–237 (2002).
[17] Lin, D., Zhao, S., Qin, L. and Zhou, M.: Identifying Synonyms among Distributionally Similar Words, *Proc.*

18th International Joint Conference on Artificial Intelligence (*IJCAI*), pp.1492–1493 (2003).
[18] EDR 日本語単語辞書：日本電子化辞書研究所 (1995).
[19] CD-毎日新聞 2000 年度版：毎日新聞社 (2000).
[20] 湊　吉正：チャレンジ小学国語辞典第五版，株式会社ベネッセコーポレーション (2011).
[21] 見坊豪紀，金田一京助，金田一春彦，柴田　武，飛田良文：三省堂国語辞典第 4 版，三省堂 (1994).
[22] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying conditional random fields to Japanese morphological analysis, *Proc. Conference on Empirical Methods in Natural Language Processing* (*EMNLP*) (2004).
[23] 工藤　拓，賀沢秀人：Web 日本語 N グラム第 1 版，言語資源協会 (2007).

梶原 智之　（学生会員）

2013 年 3 月長岡技術科学大学工学部電気電子情報工学課程卒業．同年 4 月同大学大学院工学研究科修士課程電気電子情報工学専攻に進学，現在に至る．自然言語処理の研究に従事．言語処理学会，人工知能学会各学生会員．

山本 和英　（正会員）

1996 年 3 月豊橋技術科学大学大学院工学研究科博士後期課程システム情報工学専攻修了．博士（工学）．1996〜2005 年（株）国際電気通信基礎技術研究所（ATR）研究員．1998 年中国科学院自動化研究所国外訪問学者．2002 年から長岡技術科学大学，現在，准教授．自然言語処理，テキストマイニングの研究開発に従事．言語処理学会，人工知能学会，電子情報通信学会，日本語教育学会各会員．