日英専門用語言い換え規則のパラレル化

影浦　峡
国立情報学研究所　人間・社会情報研究系
kyo@nii.ac.jp

Abstract

日英の複合専門用語の言い換え（異形）規則[1]をパラレル化するための第一歩として、言い換え操作をその内部的な観点から文法的レベルでパラレル化した規則を定義したので報告する。この規則は、言い換え抽出システムで動くようになっているが、基本的には、「言語間でパラレルな言い換え」という未規定の概念とそれに対応する現象の領域を手続き的に確定するための第一歩となることを目的としたものである。現在、これをウェブ上にて公開し経験的なデータを得ながら「パラレルな言い換え」の概念規定を規則の拡充をはかるかたちで精緻化しようと試みている。

1　はじめに

本研究は、次の二つの目的を持っている。(a) そもそも「パラレルな言い換えとは何か」の概念／現象規定にアプローチするための枠組みを試行的に提案すること、(b) そのために現実に動く粒度の並列的な言い換え規則の集合を日英を対象に定義すること、である。

以下では、まず、「パラレルな言い換え」の概念を簡単に整理し、次いで、その中で最も単純に定義できるパラレルな言い換えを、メタ規則とユニフィケーションに基づく言い換え認定システム Fastr (Jacquemin, 2001) の上でインプリメントした規則を示す。

2　パラレルな言い換え

「パラレルな言い換え」は自明な概念／現象ではない。そもそも言語トークンのレベルでは、「言い換え」は基本的に記録対象外の現象であるし、タイプのレベルでの「言い換え」は概念的にも現象の範囲としても自明ではない。まして二言語にわたる「パラレルな言い換え」は定義のために理論的抽象及び／あるいは目的論的枠組みを要する。

本研究では、人間による翻訳の支援のためにパラレルな言い換え例の抽出システム構築を目的としている。技術翻訳から第二言語での執筆、言語学習にわたる広い範囲でこうした多言語にわたる言い換え例の抽出システムが求められている。

この目的から望ましい言い換えは、「対応する文脈」で「発動される」、「対応する」言い換えのパターンである。しかしながら「対応する文脈」とは何か、言い換えが「発動される」文脈的契機は何か、その中で言い換えパターンの「対応」とは何かのすべてにわたって、調べた範囲では、既往の研究はほとんどない。

従って、本研究では、全体としてこれを実現するための第一歩として、「対応する語内部の文法的構造において対応すると考えられる言い換え」をパラレルな言い換え規則として定義する。このために考慮すべき要素的対応の概念は、次のものである。

1. 言い換え操作の文法的対応性：例えば、日本語複合語の名詞修飾部を形容詞に変える操作と英語複合語の名詞修飾部を形容詞に変える操作は対応すると見なすこと。
2. 言い換え結果対応性：例えば information retrieval の異形としての retrieval of relevant information を、「情報検索」の異形としての「適合情報の検索」に対応すると見なすこと。
3. 対応する言い換え規則が発動される元の用語レベルの対応性（用語対応性）：対訳関係にある語彙レベルで、「対応する言い換え規則」と定義されたものがどの程度そもそも適用可能なのかの度合い。

1 と 2 は、規則の対応性に関わり、3 は語彙に依存した対応性の度合いである。それゆえ、以下で「パラレルな言い換え操作」を規則として定式化するときには 1 と 2 の要素的対応を指針とし、3 の語彙依存対応性はそのようにして概念化された「パラレルな言い換え操作」の現象範囲を近似的に見るために必要な概念と位置づける。

3　パラレルな言い換え操作

言い換え／異形研究は色々なされている (Carl et al., 2002; Yamamoto, 2002; 佐藤, 1999)。用語の言い換えに関しても、いくつかの言語を対象に研究が進められてきた (Daille et al., 1996; Jacquemin, 2001; Jacquemin,

[1]言い換えと異形とはほぼ同じ意味で用いる。

表 1: 用語品詞シフトの並列規則

言い換えタイプ	日本語例	英語例
項－動詞	X1 NS1 → X1 'を' NS1 VS	N1 N2 → V2 ART? N1
日：7 規則	(概念学習 → 概念を学習する)	(word category → categorise words)
英：3 規則	NS1 X1 → X1 'を' NS1 VS	X1 N2 → V1 N2
	(実装システム → システムを実装する)	(impl. system → implement system)
修飾－動詞	NA1 NS2 → NA1 S4 NS2 VS	A1 N2 → ADV1 V2
日：5 規則	(曖昧分類 → 曖昧に分類する)	(ambiguous classification →
英：2 新規則		ambiguously classify)

表 2: 主要部変換の並列規則

	日本語例	英語例
日：12 規則	NS1 NX2 → NX2 'の' NS1	V1 N2 → N2 V1
+ 2 新規則	(追加資料 → 資料の追加)	(added material → material addition)
英：7 規則	NX1 TPNS1 NX2 → NX2 NX1	V1 N2 → N2 V1
+ 2 新規則	(共有化メモリ → メモリ共有)	(shared memory → memory sharing)

1998; Yoshikane et al., 2003; Schmidt-Wigger, 1999)。以下では複合専門用語言い換えの枠組みで定式化された日英の言い換え規則 (Jacquemin, 2001; Yoshikane et al., 2003) を出発点として[2]、日英言い換え規則の第一次的並列化をはかる。形態素解析等の前処理の前提に応じて日英の記述粒度は異なるが、ここでは各粒度を尊重した。

3.1 並列化の基準

2 節で示した並列の枠組みを、次のように具体化する。

- 機能要素・内容要素の区別は日英で対応するとする。さらに「の」と of のように語レベルでの対応も必要に応じて認める。
- 用語全体の大まかな品詞レベルは日英で対応するとする。
- 語構成要素毎の大まかな品詞レベルは日英で対応するとする。

常識的な対応ではあるが、個別に対応関係にある語対では、そもそもこうした対応が維持されない場合は少なくない。

これを用いて、言い換え操作のカテゴリーを以下のように定める[3]。

1. 用語品詞シフト：元の複合語の品詞を変更する操作。例えば「概念分類 → 概念を分類する」等。要素の統語形態的変更を伴いうる。
2. 主要部変換：「共有メモリ → メモリ共有」など。要素の統語形態的変更を伴いうる。
3. 内部異形：全体・主要部の中核を保持する異形：
   a. 機能操作：機能要素の挿入／削除等による異形。「概念分類 → 概念の分類」など。
   b. 内容語操作：内容語の挿入／削除等による異形で、大きく、修飾と並列がある。「開発環境 → 開発支援環境」、"language processing→ language and speech processing"等。
   c. 形態／形態統語操作：構成要素の形態／統語的変換。"categorial grammar→category grammar", 「概念階層 → 概念階層化」、"syntactic structure→syntactical structure"等。

3.2 個別規則の並列化

個別規則の並列化においては、(1) 日英ともに元の複合語単位の外部からの修飾を禁止するとして作られていた部分を対応関係を導入するために拡張すること、(2) 前処理の前提を維持しながら対応関係を強化するための規則記述の変更を加えつつ、Jacquemin (2001) の規則と Yoshikane (2003) の規則とを対応づけた。表 1 から 3 に、並列規則のカテゴリー、規則数と例を示す。

4 規則の語彙覆い率

表 1 から 3 に定義した規則は、2 節と 3.1 節で定義した並列性を維持している[4]ため、「パラレルな言い換え」

[2]英語規則は合計で 51 規則、日本語規則は 64 規則。

[3]Jacquemin (2001) も Yoshikane et. al. (2003) も言語依存の言い換えカテゴリーを用いているため、ここでのカテゴリーとは大きく異なっている。

[4]記述粒度の差により 1 対 1 の並列性は存在しないが、1 対 1 の「言い換え規則の並列性」とは何かはわかっていない。

表 3: 内部異形の並列規則

言い換えタイプ	日本語例	英語例
機能操作		
日：8 規則	NX1 NX2 → NX1 ‘の’ NX2	N1 N2 → N2 PREP N1
英：5 規則	(関数計算 → 関数の計算)	(job amount → amount of jobs)
	NX1 ‘の’ NX2 → NX1 NX2	N1 PREP N2 → N2 N1
	(関数の計算 → 関数計算)	(amount of jobs → job amount)
内容語操作		
- 修飾		
日：4 規則	NX1 NX2 → NX1 {NX TPX?}+ NX2	X1 N2 → X1 {A|N|V} N2
+ 4 新規則	(開発環境 → 開発支援環境)	(word type → word class type)
英：8 規則	NX1 NX2 NX3 →	X1 A2 N3 → X1 {A|N|V} A2 N2
+ 2 新規則	NX1 {NX TPX?}+ NX2 NX3	
	(概念分類問題 → 概念範疇分類問題)	(big fat cat → big fat stinky cat)
- 並列		
日：10 規則	NX1 NX2 → NX1 C NX S NX2	X1 N2 → N1 C N N2
+ 1 新規則	(学習制御 → 学習と対話の制御)	(word class → word and concept class)
英：11 規則	NX1 NX2 → NX1 S NX C NX2	X1 N2 → X1 N C N2
+ 2 新規則	(知識獲得 → 知識の生成と獲得)	(word type → word type and class)
形態操作		
- 名詞－名詞		
日：1 規則	NX1 TPNS NX2 → NX1 NX2	X1 N2 → X1 N2’
+ 5 新規則	(共有化メモリ → 共有メモリ)	(word classification → word class)
英：5 規則	NX1 NX2 → NX1 ‘の’ NX2 TPNS	N1 N2 → N2’ PREP N1
	(概念階層 → 概念の階層化)	(word class → classification of words)
- 名詞－動詞		
日：1 規則	NS1 N2 → NS1 VS N2	N1 N2 → V1 N2
英：2 規則	(暴走車両 → 暴走する車両)	(index grammar → indexed grammar)
- 動詞－名詞		
日：3 規則	NX1 VS NX2 → NX1 NX2	V1 N2 → N1 N2
英：1 規則	(分類機械 → 分類する機械)	(indexed grammar → index grammar)
- 名詞－形容詞		
日：4 規則	NA1 NX2 → NA1 MD NX2	N1 N2 → A1 N2
英：2 規則	(曖昧情報 → 曖昧な情報)	(category grammar → categorial grammar)
- 形容詞－名詞		
日：3+1 規則	NX1 TPNA NX2 → NX1 NX2	A1 N2 → N1 N2
英：1+1 規則	(幾何的モデル → 幾何モデル)	(categorial g. → category g.)
- 形容詞－形容詞		
日：2 規則	NX1 TPNA2 MD NX3 →	A1 N2 → A1’ N2
英：1 規則	NX1 TPNA2 NX3	(syntactic information →
	(幾何的な表現 → 幾何的表現)	syntactical information)

表 4: 言い換え規則タイプ毎の用語対応性

言い換えタイプ	用語対応対数	用語対応対率
用語品詞シフト	6312	32.3 %
項－動詞	1316	6.7 %
修飾－動詞	6279	32.1 %
主要部変換	6141	31.4 %
内部異形	12405	63.5 %
機能操作	10792	55.3 %
内容語操作	10433	53.4 %
- 修飾	10361	53.0 %
- 並列	10433	53.4 %
形態操作	8870	45.4 %
- 名詞－名詞	33	0.17 %
- 名詞－動詞	1816	9.3 %
- 動詞－名詞	2	0.01 %
- 名詞－形容詞	5301	27.1 %
- 形容詞－名詞	3560	18.2 %
- 形容－形容	33	0.17 %

の第一歩として捉えられる現象の範囲を近似的に記述したものである。ここで、2 節の 3 の考えに従って、ここで定式化したパラレルな規則が、対応する語彙に対してどの程度適用できるか（規則の用語対応性）を見ておくことは有用であろう。表 4 は、技術用語の辞典 (小谷ら, 1990) の日英語見出しリストに対して、規則がどのくらい理論的に適用可能かを調べた数値である。

小谷ら (1990) の見出しは 30,024 語対であり、そのうち 19,532 対が複合語である（日英語ともに単純語の対は 4170 語対、英語が複合語の対は 1543 語対、日本語が複合語の対は 4779 語対であった）。表 4 では、そのうち日英語ともに複合語である 19,532 語対を対象としている。用語品詞シフトと主要部変換は約 3 分の 1 の複合語対に、内部異形は約 3 分の 2 の複合語対に、原則的には適用可能であることがわかる。

一方、いかなる言い換え規則も適用できない複合語対は 7090 (36.3%) あった。このうち日英いずれにも規則が適用できない対は 558 対、日本語に適用規則がないものは 1473、英語に適用規則がないものは 1179 であった。残りの 3880 対については、個別に規則は適用できるが、パラレルな規則は存在していない。

5 終わりに

本稿では、日英間の複合専門用語を対象とした「パラレルな言い換え規則」の予備的な概念整理を行うとともに、第一次的な「パラレル」概念とそれが対象とする現象を規則のかたちで記述した。これに対しては、Web 上で動くシステムを現在日英を対象に試験的に作成公開しているところである。そこから得たデータを元に、「パラレルな言い換え規則」という概念はどのようなものか、それを定義しうる視点にはどのようなものがあり、それが覆う言語現象はどのような範囲となるのかを整理しながら、同時に応用システムとしての並列言い換え検索システムを構築していくという作業を進めていく予定である。

謝辞

本研究を進めるにあたり、Fastr システムを提供してくださった CNRS-LIMSI の Christian Jacquemin 先生に感謝します。また、言い換えや異形を巡る様々な議論につき合って下さった京都大学の佐藤理史先生、Web 上の異形認識システムの作成公開を行なっている大学評価・学位授与機構の芳鐘冬樹・野澤孝之先生に感謝します。

References

Michael Carl and Philippe Langlais. 2002. An intelligent terminology database as a pre-processor for statistical machine translation. *Computerm 2002*, 15–21.

Béatrice Daille, Benoît Habert, Christian Jacquemin and Jean Royauté. 1996. Empirical observation of term variations and principles for their description. *Terminology*, 3(2):197–257.

Helmut Felber. 1984. *Terminology Manual.* Paris: Unesco & Inforterm.

Christian Jacquemin. 1998. Analyse et inférence de terminologie. *Revue d'Intelligence Artificielle*, 12(2):163–205.

Christian Jacquemin. 2001. *Spotting and Discovering Terms through Natural Language Processing.* Cambridge, Mass: MIT Press.

小谷卓也, 郡亜都彦 1990. **日・英・西技術用語辞典**. 東京: 研究社.

佐藤理史. 1999. 論文表題を言い換える. **情報処理学会論文誌**, 40(7):2937–2945.

Antje Schmidt-Wigger. 1999. Term checking through term variation. *TKE'99*, 570–581.

Kazuhide Yamamoto 2002. Machine translation by interaction between paraphraser and transfer. *Coling 2002*, 1107–1113.

Fuyuki Yoshikane, Keita Tsuji, Kyo Kageura and Christian Jacquemin. 2003. Morpho-syntactic rules for detecting Japanese term variation: establishment and evaluation. *Journal of Natural Language Processing*, 10(4):3–32.