複数の翻訳に見られる差異の分析

本田友乃 [1]　山本真佑花 [2]　影浦峡 [3]
[1,3] 東京大学　[2] 立教大学
tomono20@g.ecc.u-tokyo.ac.jp　yamamoto.mayuka@rikkyo.ac.jp
kyo@p.u-tokyo.ac.jp

概要

本稿では，同一の原文に対する複数の翻訳の差異を明らかにするために，2 種類の翻訳間の差異を記述するための概念と概念付与の手続きを枠組みとして定めた．この枠組みを技術文書に適用したところ，文法的な差異は広く出現しており，意味的な差異や語用論的な差異は限定的であることが明らかになった．また，構築した枠組みをもとに差異の記述を自動化するメカニズムを一部実装した．

1　はじめに

翻訳の質を論ずる際，評価者は，ありうる翻訳を想定し，それとの差異を暗黙のうちに考慮しているはずである．産業翻訳の品質については，Multidimensional Quality Metrics (MQM) [1] のような指標が複数定められているものの，その前提として翻訳間にはどのような差異があり，それは何を対象としてどのような観点から観察されるかといった点を認識する重要性は意識されてこなかった．

翻訳間の差異を考慮することは，機械翻訳の自動評価にも知見を与えうる．機械翻訳の自動評価では，BLEU [2] をはじめとして参照訳との類似度をもとにした評価が主に行われているが，そこで用いられる参照訳は検討の対象とされない．不適切な評価を行わないためにも，複数の翻訳の現実的な存在可能性の範囲を明らかにすることは重要である．

複数訳の検討は，文芸翻訳において行われているが，言語表現に着目して体系的に差異を記述しようとするものではなく，作品の背景や翻訳者の個性に着目した複数訳の検討が中心となっている．翻訳指南書でも，学習の観点から複数訳を並置しその差異を示そうとする場面があるが，目的を先取りした限られた範囲の差異が指摘されるにとどまる．

以上を背景に，本稿では複数訳の差異を記述するための枠組みを構築し，それを用いて産業翻訳の文書を対象に複数訳の差異を体系的に記述する．さらに，差異の自動同定・記述メカニズムの一部実装についても述べる．

2　研究対象

この課題に取り組むために，産業翻訳で扱われる文書を対象とし，目標言語文書（TD）が独立に翻訳者の個性に依存しないかたちで生成されたデータである，ParaNatCom [3] を利用する．ParaNatCom は Nature Communications の抄録をもとに作成された英日の対訳コーパスであり，同一の起点言語文書（SD）に対し 3 種類の TD が付された 148 セットの文書が含まれる．表 1 に基本統計を示す．

表 1　ParaNatCom の基本統計（文数，延べ語数，異なり語数は各文書あたりの平均）

文書	文書数	文数	延べ語数	異なり語数
SD	211	7.55	182.50	107.94
TD1	211	7.70	265.16	122.39
TD2	148	7.76	270.99	124.28
TD3	148	7.68	277.51	125.01

3　枠組みの構築

差異を記述する概念は，翻訳方略体系 [4]，品質評価，言語学，翻訳指南書を参照しつつ，ParaNatCom に含まれる 32 文書（96 対）の分析をもとに構築した．表 2 に示すように，文法上の差異を表す Syntactic，意味上の差異を表す Semantic，語用論的な差異を示す Pragmatic の 3 カテゴリからなる．

これらの概念を TD に付与するために，以下のように記述の手続きを定める．

step 0 では [5] を参考に，以下の 4 つの規則に従って SD を分割することで，文と語・句の中間的な言語単位としてチャンクを設定した．

1. 節を導く接続詞，関係詞（主語を修飾する場合は除く）の前後

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 2　差異を記述するための概念

Syntactic		Semantic	Pragmatic
g1 Not applicable	g9 Cohesion difference	s1 Abstraction difference	p1 Not applicable
g2 Sentence structure difference	g10 Unit difference	s2 Perspective difference	p2 Domain adaptation difference
g3 Punctuation difference	g11 Part of speech difference	s3 Emphasis difference	p3 Information difference
g4 Chunk structure difference	g12 Word form difference	s4 Distribution difference	p4 Cultural filtering difference
g5 Phrase structure difference	g13 Functional word difference	s5 Synonym	
g6 Person difference		s6 Paraphrase	
g7 Spelling difference		s7 Semantic equivalence	
g8 Loan difference		s8 Semantic difference	

2. to 不定詞，前置詞，動名詞の後に 3 語以上続く場合，それらの前（動詞が後続する場合を除く）
3. コンマ，セミコロン，ハイフン等のパンクチュエーションの後
4. 主語が 3 語以上続く場合，その後

step 1-3 の手続きは，それぞれ文，チャンク，語・句を対象としており，概念を付与する順序と判断基準を含めて決定木の形で整理した[1)]．step 1・2 は言語単位が広いため，構造を捉えるという観点から Syntactic に属する概念の一部のみを付与し，step 3 では，Syntactic の一部の概念及び Semantic，Pragmatic からそれぞれ一つずつ概念を付与する形で設定した．

また，枠組みを適用するにあたって，一貫して概念が付与されるよう，SD と TD の対応づけや TD の単位の認定方法について，以下のように整理した．

1. step 3 における SD の分割方法
   (a) 原則として SD は 1 単語ずつ分割し，各単語に対応する TD 間の差異を記述する
   (b) 受動態や否定文，完了形などの用言や，SD のイディオム，その他訳対応をとることが難しい語や句については，SD を 1 単語で分割しなくてよい
   (c) 記号の訳出に文字化が関わる場合や SD が記号で連結された複合語である場合は，単語中の記号の後で分割する
2. 対応づけに伴う TD の扱い
   (a) 用言を伴わない機能語や同格表現で SD と独立に対応づけられない語や句は，左の語とセットにする
   (b) 用言の使用に伴う機能語で SD と独立に対応づけられない語や句は，用言（右の語）とセットにする
   (c) パンクチュエーションは左の語とセットにして step 2 で観察するため，SD との直接の対応づけは行わない
3. TD の単位
   (a) SD の 1 単語と対応する場合，名詞の連接はまとめて 1 単語としてカウントする
   (b) サ変動詞や形容動詞の語幹に助詞や動詞が付加された形の動詞は 1 単語としてカウントする
   (c) （）内の語や句は（）も含め 1 つの単位とみなして記述する

1) 詳細は付録に記載した．

4 ParaNatCom の記述

ParaNatCom に含まれる文書のうち，概念の構築で用いなかった 50 種類（150 対）の TD に対して枠組みを適用した．付与した概念の出現頻度について，step 及びカテゴリごとの分布を図 1–5 に示す．

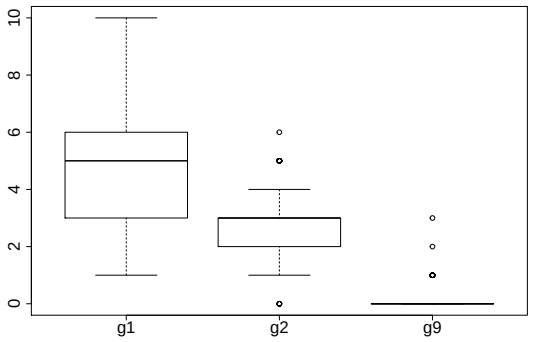

図 1　step 1 における差異の分布

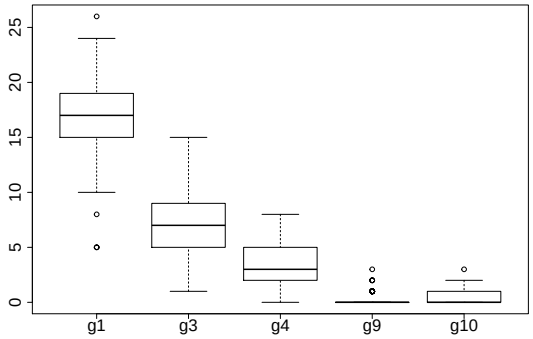

図 2　step 2 における差異の分布

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

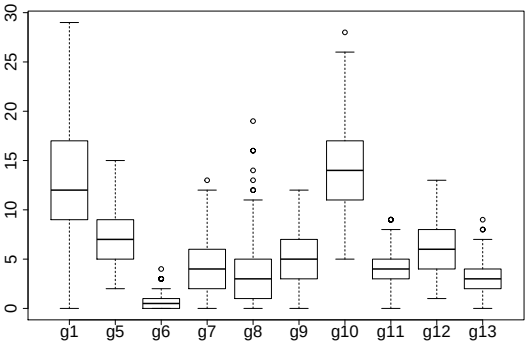

図 3　step 3 Syntactic における差異の分布

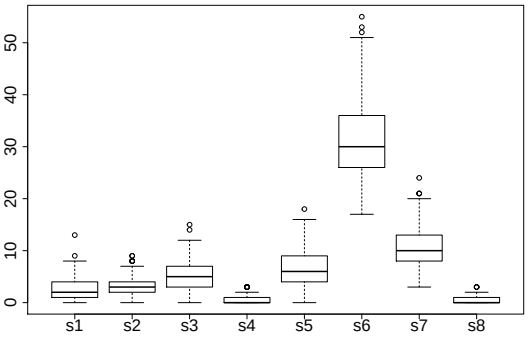

図 4　step 3 Semantic における差異の分布

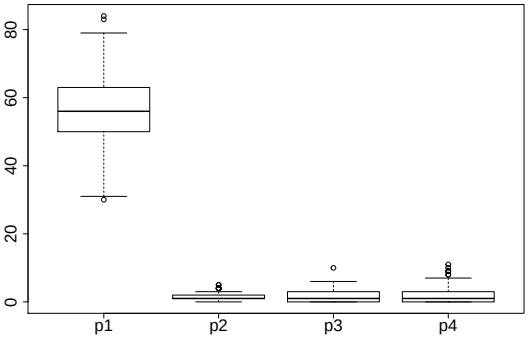

図 5　step 3 Pragmatic における差異の分布

図 1 では，g2 Sentence structure difference の分布から，文レベルでの構造上の差異はほとんどの文書で一定程度存在することがわかる．また，g9 Cohesion difference の分布より，SD に明示的に存在しない接続詞や主語を文頭で補うといった結束性の有無による差異は，多くの文書では現れていないものの，TD 間の差異としては存在することが確認される．

図 2 では，g3 Punctuation difference の分布から，句読点やコロンなどのパンクチュエーションの打ち方に関する差異はどの文書対でも必ず生じており，出現頻度も高いことがわかる．また，g1 Not applicable と g4 Chunk structure difference の分布より，チャンクを対象とした構造上の差異は存在するものの，相対的には少ないことがわかる．g9 Cohesion difference, g10 Unit difference の分布より，主語を明示化するといった結束性に関する差異や，節と句で訳出単位が異なるという差異が現れている文書数はわずかである．

図 3 より，step 3 の中では，Syntactic に属する概念の分布は比較的差が小さく，語や句のレベルで見た場合には文法的な差異が広く出現していることがわかる．g10 Unit difference が最も多く見られ，語や句のレベルでは訳出単位の違いがどの文書でも一定数以上存在し，比較的多く見られることがわかる．g5 Phrase structure difference や g12 Word form difference の分布より，句として訳出されていた場合の構造上の差異や用言の語形に関する差異も，全ての文書中に現れていることがわかる．また，step 1・2 とは異なり，言語単位が小さい場合には，語の訳出の仕方や省略の有無といった結束性による差異は珍しい現象ではないことが g9 Cohesion difference の分布からわかる．表記の差異を表す g7 Spelling difference と音形借用の有無による違いを表す g8 Loan difference も一定程度見られるが，この 2 つの概念は特に，抄録中における用語の訳出の差異を反映していると考えられる．また，人称や人称表現の違いを表す g6 Person difference は，Syntactic の中では最も観測されなかった．これは，ParaNatCom のみを記述の対象としたため，論文中で用いられる人称表現のバリエーションが限定的であり，人称表現自体の出現数が少ないことからも説明される．

図 4 より，Semantic ではどの文書にも一定以上 s6 Paraphrase が付与されており，意味的なカテゴリの中では言い換えとして説明される差異が多いことがわかる．また，s7 Semantic difference についても，s6 ほど多くは見られないものの全ての文書対で付与されており，意味的には等価であると判断される現象が比較的多く観察されたことがわかる．その他の概念については，文書によって出現するものと出現しないものとがあった．その中で比較的多く観察された差異は，s5 Synonym や s3 Emphasis difference であり，同義関係にある語として説明される差異や，意味の焦点，強調の仕方に関する差異が現れていることがわかる．s1 Abstraction difference で表される具体・抽象に関する差異と s2 Perspective difference で表される視点に関する差異は，多くの文書で生じていることが確認できるものの，頻度としては少ない．同じ意味を表しており，かつ語の長さが異なるという現象を表す s4 Distribution difference と，同音異義

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

語が用いられているなど明らかに意味が異なる現象を表す s8 Semantic difference については，多くの文書で見られなかった．

図 5 より，Pragmatic の分布は偏りが大きく，p1 Not applicable で記述される差異が最も多かった．そして，他の概念で説明される差異，すなわち文書タイプへの適応による差異や情報量の違い，異化・同化に関する差異はあまり観察されなかった．文書外の情報を参照することで生じたと判断される差異は限定的であることがわかる．

5 記述の自動化

最後に，TD 間の差異を自動的に出力するための見通しとして，枠組みの一部を実装した方法と結果について簡単に述べる．

3 で構築した枠組みでは，同一の SD を起点として 2 種類の TD 間の差異を記述したため，SD と TD の対応づけが必要である．対応づけには，単語アライメントツールの SimAlign [6] を使用し，SD の一単語に対して MeCab [7] で分かち書きされた形態素を対応づける．ただし，アライメントの精度の向上のため，TD については，3 で述べた TD の対応づけや TD の単位に関するルールをもとに分かち書きを修正した．概念の付与は 3 で定めた手続きに従って条件分岐の形で実装し，実装に必要な文法情報は，NLTK [8] と MeCab による分析結果を利用した．

以上の手順に従って，4 で記述の対象とした ParaNatCom の TD のうち，32 種類（96 対）の TD について自動的に概念の付与を行った．4 の記述で付与した概念と，自動的に付与された概念とが一致した割合を，正解率として計算した．正解率の結果を表 3 に示す．概念の数や偏りの問題もあるが，特に Syntactic に関して改善が必要であることがわかる．

表 3 自動化の正解率 (%)

	Syntactic	Semantic	Pragmatic
平均	59.2	65.1	74.4
最大値	79.5	84.6	91.4
最小値	43.7	43.8	52.2
中央値	60.4	63.6	74.0

6 おわりに

本稿では，複数訳としてありうる範囲を明らかにする試みとして，技術文書を対象として TD 間の差異を記述する枠組みを構築し，枠組みを適用して実際に TD 間に生じている差異を記述した．また，構築した枠組みを足場として，記述の一部について自動化を行った．枠組みの構築では，差異を記述するための概念を Syntactic，Semantic，Pragmatic の 3 つのカテゴリに類型化し，概念を付与する手続きを整理した．また，差異の記述では枠組みに従って概念を付与し，対象データについて，意味的なカテゴリや語用論的なカテゴリを用いて説明できる差異は限定的であるが，文法や構造についての差異は広く生じていることを，分布に基づいて具体的に示した．自動化では，枠組みをもとに手続きの一部を実装した．手法の見直しや，文，チャンクを対象とした記述の自動化は今後の課題である．

参考文献

[1] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH and QTLaunchPad. Multidimensional Quality Metrics, 2005.

[2] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[3] Masao Utiyama. ParaNatCom — Parallel English–Japanese abstract corpus made from Nature Communications articles, 2019. https://www2.nict.go.jp/astrec-att/member/mutiyama/paranatcom/.

[4] 山本真佑花, 山田優, 藤田篤, 宮田玲, 影浦峡. メタ言語としての翻訳方略体系の構築と検証. 言語処理学会第 27 回年次大会発表論文集, pp. 1111–1116, 2021.

[5] 岡村ゆうき, 山田優. 「順送り訳」の規範と模範：同時通訳を模範とした教育論への試論. **MITIS Journal**, Vol. 1, No. 2, pp. 25–48, 2020.

[6] Masoud Jalili Sabet, François Yvon Philipp Dufter, and Hinrich Schütze. SimAlign: High Quality Word Alignments without Parallel Training Data Using Static and Contextualized Embeddings. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings**, pp. 1627–1643, 2020.

[7] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In **Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)**, pp. 230–237, 2004.

[8] Steven Bird, Ewan Klein, and Edward Loper. **Natural Language Processing with Python**. O'Reilly Media, 2009.

[9] Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. Japanese SemCor: A Sense-tagged Corpus of Japanese. In **The 6th International Conference of the Global WordNet Association (GWC-2012)**, pp. 56–63, 2012.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

付録　差異を記述する概念を付与する手続き

step	ID	質問	付与する概念または次の質問	
			True	False
	Q0a	2 種類の TD が文として同定される	Q1a	Q0b
	Q0b	2 種類の TD がチャンクとして同定される	Q2a	Q3a
1	Q1a	同じ文字列である	n	Q1b
	Q1b	文頭の接続詞や主語の訳出の有無が異なる	g9・Q1c	Q1c
	Q1c	文内のチャンクの順序が異なる／文分割の有無が異なる	g2	g1
2	Q2a	同じ文字列である	n	Q2b
	Q2b	パンクチュエーションの有無や位置が異なる	g3・Q2c	Q2c
	Q2c	チャンクの訳出の有無が異なる／ SD に明示的に存在しない主語や接続詞などの訳出の有無が異なる	g9・Q2d	Q2d
	Q2d	訳出されたユニットが節と句で異なる	g10	Q2e
	Q2e	SD に対応するチャンク内の語もしくは句の訳出順序が異なる／複数の語に係る修飾語の訳出語数が異なる／複数の語が係る被修飾語の訳出語数が異なる	g4	g1
3	Q3a	同じ文字列である	n	Q3G-a/Q3S-a/Q3P-a
	Q3G-a	語を対象として同定される	Q3G1-a	Q3G-b
	Q3G-b	句を対象として同定される	Q3G2-a	Q3G3-a
	Q3G1-a	SD の代名詞や関係代名詞に対する訳出表現が異なる	g9	Q3G1-b
	Q3G1-b	SD の人称代名詞に対する訳出表現が異なる	g6	Q3G1-c
	Q3G1-c	文字種が異なる／英語表記が異なる（大文字と小文字の使い分けなど）／日本語表記が異なる（漢字と仮名の使い分けなど）／記号の種類，有無が異なる	g7	Q3G1-d
	Q3G1-d	音形借用の有無が異なる	g8	Q3G1-e
	Q3G1-e	品詞が異なる	g11	Q3G1-f
	Q3G1-f	用言であり，時制，アスペクト，モダリティ，活用形，態，動詞の種類（自動詞と他動詞）のいずれかが異なる	g12	Q3G1-g
	Q3G1-g	異なる助詞が用いられている	g13	g1
	Q3G2-a	句の構造が異なる（語数と品詞が同じであり，文字列のみが異なる）	Q3G2-b	Q3G1-a
	Q3G2-b	文脈からの内容語の置き換えの有無が異なる／代名詞化の有無が異なる	g9	g5
	Q3G3-a	指示語や冠詞等の訳出の有無が結束性により異なる	g9	Q3G3-b
	Q3G3-b	SD の人称代名詞に対する訳出表現が異なる	g6	g10
	Q3S-a	同音異義語が用いられている，肯定と否定が逆であるなど，意味が明らかに異なる	s8	Q3S-b
	Q3S-b	文脈から具体的な言語表現に置換したり付加したりする操作の有無が異なる／（）等を用いた注釈の有無が異なる	s1	Q3S-c
	Q3S-c	動詞の種類（自動詞，他動詞）が異なる／受動態もしくは使役態の使用の有無が異なる／人称が異なる	s2	Q3S-d
	Q3S-d	修飾語の有無が異なる／取り立て助詞の有無が異なる／冠詞や代名詞の訳出の有無が異なる	s3	Q3S-e
	Q3S-e	内容語が共通している，もしくは同義関係の語が用いられており，かつ，品詞やユニットが異なる	s4	Q3S-f
	Q3S-f	同義関係である（日本語 Wordnet [9] を参照）	s5	Q3S-g
	Q3S-g	同じ語が用いられており，時制，活用形，アスペクトのいずれかが異なる／同じ語が用いられており，表記のみが異なる／記号の種類のみが異なる／付属語の有無のみが異なる	s7	s6
	Q3P-a	「著者」「研究」「本稿」「本論文」といった論文中で用いられると想定される語の使用の有無が異なる	p2	Q3P-b
	Q3P-b	修飾語の有無が異なる／（）等を用いた注釈の有無が異なる	p3	Q3P-c
	Q3P-c	SD と表記が同じ訳語の使用の有無が異なる	p4	p1

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).