様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　３０　日現在

機関番号：　１７１０２
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ０２７２０

研究課題名（和文）学術分野共通性を優先した修訂情報付き英語科学論文コーパスの構築

研究課題名（英文）Compiling a Corpus of English Language Academic Papers with Information on Revision Considering Commonality of Academic Genres for Expressions

研究代表者

徳見　道夫（Tokumi, Michio）

九州大学・言語文化研究院・名誉教授

研究者番号：９００９９７５５

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：日本人が執筆した論文に修訂情報が付与されたコーパスが構築されれば，日本人学術英語解明において新たなコーパス研究が展開されることが期待される．論文の規範性は学術分野にも強く依存する．そこで本研究では「表現に対する学術分野共通性」という尺度を導入した．この尺度によって，分野に依存しにくい論文内の断片に優先的に修訂情報を付与することができる．本研究では，日本人による論文の英文概要を中心に，延べ語数約2万語規模の修訂情報が付与し，整備した．

研究成果の概要（英文）：It is necessary to compile a corpus of academic papers by Japanese researchers with revision information in order to promote research on their academic English. However, a constructive or expressional validity for academic papers depends partially on their genre. The research project described here developed an index to measure the degree of commonality of academic genres for an expression. It can give priority to expressions that require proofreading and revision. The project compiled a 20,000-word corpus based on English-language abstracts Japanese researchers.

研究分野：　英語教育

キーワード：　英語教育　学術英語　コーパス　修訂情報　学術分野

１．研究開始当初の背景

多くの学術分野で，最新の成果発信は英語で行われており，科学者にとって英語科学論文（適宜，「論文」と略記する）の執筆は必須である．十分に学術英語に通じていない日本語のみを母語とする日本人（以後，単に「日本人」という場合は，このような日本人を指すこととする）の論文には，たとえば「受け身が多い」「時制の混乱」といった日本人に共通の言語的特徴が認められる．このような日本人の英語科学論文の問題点を，網羅的に明らかとすることは，日本人の学術英語研究における重要な課題の一つである．

このような問題に対して有効な手段の一つとして注目されつつあるのが，大規模電子化用例集であるコーパスを活用した方法である．コーパスには，学習者が産出した文章を大量に集積した学習者コーパスという下位カテゴリがある．そのような学習者コーパスに，エラー，不自然な箇所の明示や，校正後の表現等の情報（修訂情報）を付与，分析することで，内省では気付きにくい言語特徴が明らかとなる．

このように，日本人による論文に修訂情報等が付与されたコーパスが構築されれば，コーパス研究の視座から，日本人の学術英語の解明において新たな研究が展開されることとなる．しかし，このようなコーパスは未だ構築されておらず，それは次のような問題にも起因していた．

問題 1．修訂情報を付与すべき，日本人による英語科学論文の収集が難しい

一定の表現上の質が担保された論文は，格式高いジャーナル等から比較的容易に収集できる．その一方で，表現上は改善の余地がある未完成論文ともいえる，修訂情報を加えるような論文は，そもそもそのような段階であることが明示されて公開されることはなく，非常に収集が難しい．さらに，日本人によるもの，と限定すれば，なおさら難しい．

問題 2．学術分野によって論文の規範的な書き方が変わる

学術分野によって論文の形式や書き方，表現まで変わることが知られている．分野バランスを考えずに，特定分野の論文にだけ修訂情報を付与した場合，その分野にかかわる知識しか得られない可能性がある．つまり，構築されたコーパスは，その特定分野以外の他分野には寄与しない，学術分野の汎用性が低いものとなる．

２．研究の目的

目的 1．学術分野に依らない，汎用性が高い修訂情報付き日本人英語科学論文コーパス構築の方法論の提案

修訂情報の付与は，非常に負荷の高い作業である．問題 2 で述べたような学術分野による規範性の違いなどもあるため，できるだけ汎用性が高い，つまり学術分野に依らない論文もしくはその断片に優先的に付与するようなコーパス構築戦略を提案し，そのための要素技術を確立する．

目的 2．修訂情報を付与する可能性がある，日本人によって書かれた英語科学論文の効率的収集

問題１で述べたように，修訂情報が付与されるべき，日本人による英語科学論文を国内の機関リポジトリから収集する．その際，なるべく人手がかからないよう，表現上の質推定などの技術も活用した手立てで実践する．

目的 2．論文の断片的表現に対する「学術分野の共通性」の計量的推定法の開発

さまざまな分野で使われる表現が含まれるような言語断片から優先的に修訂情報を付与するために，論文中の文や一部表現など断片的表現に対する「学術分野の共通性」（適宜，「分野共通性」と略記する）という指標を導入する．たとえば，“ figure shown in Fig. X ”や“ This paper discusses/proposes that... ”など，どの学術分野でもよく使われるものを「分野共通性が高い断片的表現」と考える．適当な表現に対して，このような分野共通性に対する計量的に推定する方法を開発する．

目的 3．分野共通性を優先した修訂情報の付与

目的１で収集した論文に対して，目的 2 で開発した技術で分野共通性の高い箇所を同定し，優先的に修訂情報を付与する．

３．研究の方法

3.1 日本人による英語科学論文の効率的収集

我々の研究組織は，過去の研究課題を通じて，機関リポジトリの処理に関する経験や知見(徳見，2016)と，英語科学論文の表現上の質に関する技術(田中他，2011)を有している．これらを活用しつつ，まずは英文概要を対象とし，研究を進めた．

国内の機関リポジトリに蓄積されている「紀要論文」と，情報系学会のいわゆる SIG といわれる「研究会論文」を収集対象とした．ヒューリスティクスを設定し，収集し，日本人によるかどうか，という判断は，信頼性が求められる作業であるため，適宜，人を介することとした．

3.2 表現（文）に対する学術分野共通性の推定法

「語」の共通性については，コーパス言語学（斎藤他，2005)でも長く研究されており，χ^2 値や対数尤度比などさまざまな定量化が

知られている．実際，近年，論文コーパス等の集積が進み，English for Academic Purposes(EAP)(Hunston&Waters，1987)などの教材開発に，このような知見が活かされつつある．本研究では，まず，人が評価しやすい「文」に対して，このような分野共通性を推定する枠組みを検討した．

3.1.1 英語科学論文の手引きの例文からの着想

英語科学論文の書き方に関する書籍（手引き）が多数出版されている．そのなかで，読者の分野を限定しない手引きの例文をみてみると，次のように著者の専門分野を中心に共通性が低い表現が意外に含まれていることがわかる．

Therefore, it is necessary to develop a low energy-cost process to reduce SiO_2.

この例文は，“ therefore ”の語法に関する解説箇所で掲載され，下線部のような専門的な語が含まれる．実際には訳が付いており，それも一助となるだろうが，この分野に無関係な読み手の多くは，下線部のような箇所は適当に読み飛ばし，要所を理解することになる．いくつかの例文を観測していると，次のようにまとめられる．

(a) 文の構造が単純である．
(b) 専門的な語や表現が使われていても，それを受ける語は比較的基本的か，共通性が高いものである．

(a)は，解説すべき項目に焦点化されることが重要な例文という性格からも容易に推測されることである．(b)は，本研究の基本アイディアに関わる性質である．上述の例文であれば，“ energy-cost ”を受けるのは“ process ”で，“ SiO_2 ”を受けるのは“ reduce ”と，いずれも比較的，基本的で共通性が高い語である．人の文理解，適切な読み飛ばしという点からみても，難解な語が密集するのではなく，一方は難解でも他方は基本的であることは非常に重要な特徴と考えられる．

また，実際に論文で使用された文では，共通性が高い語だけで構成されることは稀で，専門的な語や表現が含まれることも勘案した文の共通性の推定法が求められる．そこで本研究では，文内の係り受け構造を考慮した推定法を提案した．

3.1.2 文の学術分野共通性の推定法

語wが語w'と係り受け関係を結んでいることを<w,w'>と表し，文sに含まれる係り受け関係の集合をR(s)と表す．R(s)に対して，なんらかの条件を課した結果，制限された係り受け関係の集合$R'(s) \subseteq R(s)$を考える．このとき，s の共通性 DC(s)を次のように与える．

$$DC(s) \equiv 1/|R'(s)| \Sigma_{<w,w'> \in R'(s)} F(w,w'). \quad (1)$$

ここで，F(w,w')は係り受け関係<w,w'>の共通性で，別途与えられる語の共通性F(w)に基づき，次のように計算する．

$$F(w,w') = \max \{ F(w), F(w') \}. \quad (2)$$

ここでF(w)は，分野$C_1, C_2, \ldots, C_M$から語wが発生するようなモデルを考えた際の，$P(\cdot|w)$に対するエントロピーH(w)とした．H(w)は，

$$H(w) = -\Sigma_{i=1}^{M} P(C_i|w) \log P(C_i|w) \quad (3)$$

と表され，$P(C_i|w)$は分野別コーパスから最尤推定する．

3.3 分野共通性を優先した修訂情報の付与

当初は，論文中の任意の断片で，分野共通性の高いものから修訂情報を付与することを考えていた．しかし，実際の修訂作業では，前後の文脈，特に当該の箇所にいたるまでの論文の内容や書き方を勘案しなければならない．そこで，ところどころ抜き出し，それだけで修訂を施すのは不適切であると判断し，論文の冒頭から中途までの一定のまとまりを対象とすることとした．

また，実際の修訂作業では，分野共通性が高い箇所以外だけでなく，通常通り，そのまとまり全てに対して修訂作業を施した．効率・手間を求めるならば，分野共通性が高い箇所のみを対象とすれば良いが，本研究では分野共通性推定自体がまだ実験的であることも勘案し，前文の通りとした．つまり，分析の際に，分野共通性が高い領域の修訂情報のみに注目し，処理することになる．

４．研究成果

4.1 日本人による英語科学論文の効率的収集

機関リポジトリとして国立情報学研究所 CiNii を採用した．CiNii は多くの学術刊行物がデータベース化されており，API も提供されている．

紀要論文については，学術誌名が著者の所属機関および所属部局名と「紀要」を含む論文のメタ情報を得た．また，研究会論文については，情報処理学会・電子情報通信学会・人工知能学会の研究会名をあらかじめ各学会の Web ページより得，それをクエリ化し，論文のメタ情報を得た．

これらの論文のメタ情報の著者名・所属を人目で確認し，日本人名ではないと思われる著者，日本人名とみられるものでも所属先が日本国内以外のものについては収集対象から外した．さらに，英文概要が，短すぎるもの，ピリオド等できちんと終わらず，中途で

切れていると考えられたもの，田中他(2011)らの表現上の質推定で G クラス（表現上の質が十分である）と判定されたものをのぞいた．
その結果，30,592 編の日本人が書いたと考えられる論文の英文概要を得た．

4.2 文に対する学術分野共通性の推定法
分野別の論文コーパスとして，京大論文コーパスを採用した．13 分野の一流英語ジャーナルを電子化したもので，延べ約 1,607 万トークンが含まれている．
これらの英文を TreeTagger(Shmid, n.d.)で形態素解析し，原形で(1)-(3)式を算出した．R'(s)は<w,w'>∈R(s)で<w,w'>がともに，軽動詞を除く内容語のものに限った．
同分野で文が含む語という観点でダイス係数が 0.3 より大きな 2 文を組としてランダムに抽出し，そのうち 47 組に対して EAP 英語教育担当者が各組で「いずれかの文が共通性が高いか」ということを独立に付した．その結果，41 組(87.2%)について DC の関係が反映されていることを確認した．

4.3 分野共通性を優先した修訂情報の付与
4.1 で得た英文概要のうち，情報系学会の研究会の論文を対象とし，修訂情報を施した．分野は情報学のなかでも自然言語処理を中心に修訂作業を専門家に依頼した．英文概要は 4.2 の DC が高い 200 編を対象とした．修訂対象は，1,064 文で延べ語数は 19,599 語であった．
修訂は，10 編単位で 20 部に区切り，各部をそれぞれ異なる修訂作業の専門家に依頼した．修訂作業では，一から抜本的に書き直した方が最終的な表現上の質は高くなる場合もあっても，元の英文を活かせる場合にはそれを優先してもらった．また，部毎に修訂者を分けたのは，分析時の修訂者を考慮したためである．
これらのデータは，別課題の構文的分析を施し，修訂が語の交換なのか，構造的な修訂なのか，といった類型化も行った．
また，公開にむけて，語の情報をマスクし，元の英文を辿れないようなかたちに変換したものも準備した．

参考文献

Hutchinson, T. and Waters, A.: English for the Specific Purposes, Cambridge University Press (1987).
斉藤俊雄，赤野一郎，中村純作 編: 英語コーパス言語学 改訂新版，研究社 (2005).
Shmid, H. (n.d.) TreeTagger, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html .
徳見道夫 (2016) 機関リポジトリを活用した大学別発信型語彙リストのオーダメイド作成法，科学研究費事業データベース，https://kaken.nii.ac.jp/ja/grant/KAKENHI-PROJECT-24520625/ .
田中省作，柴田雅博，冨浦洋一 ()Web を源とした質的情報付き英語科学論文コーパスの構築法，英語コーパス研究，第 18 号，pp.61-71 (2011).

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）
〔雑誌論文〕（計 7 件）
① 天野翼，宮崎佳典，田中省作，長谷川由美 (2018) 簡略化を用いた例示型英文書作成支援ツールの開発と検証，統計数理研究所共同研究リポート，397，13-24.
② 田中省作，宮崎佳典，徳見道夫 (2018) 句構造を考慮した学術分野共通性の高い頻出表現の抽出，統計数理研究所共同研究リポート，397，1-12.
③ Hasegawa, Y., Miyazaki, Y., Amano, T. (2017) A Web Application to Support Technical Writing for Non-native (EFL) English Speakers, The 31st Annual Hawaii Association of Language Teachers (HALT) Conference, 2017, 6.
④ 田中省作，宮崎佳典，徳見道夫 (2017) 構文変化検出のための校正英文対データベースの設計と試作，統計数理研究所共同研究リポート，382，29-40.
⑤ 天野翼，宮崎佳典，田中省作，長谷川由美 (2017) コーパスを用いた技術英文書作成援用ツールの開発とその評価(その 2)，統計数理研究所共同研究リポート，382，41-52.
⑥ 渡部孝幸，田中省作，宮崎佳典 (2016) 論文標題:英文汎化における語の品詞化と構文木の非冗長化，統計数理研究所共同研究リポート，356，17-25.
⑦ 宮崎佳典，戸沢信晴，田中省作 (2016) コーパスを用いた技術英文書作成援用ツールの開発とその評価，統計数理研究所共同研究リポート，356，1-16.

〔学会発表〕（計 12 件）
① 天野翼，宮崎佳典，田中省作，長谷川由美 (2018) 構造的簡略化を用いた例示型英文書作成支援 Web アプリケーションの開発と評価，情報処理学会第 80 回全国大会 (2018)
② 田中省作，徳見道夫，宮崎佳典，金丸敏幸，田地野彰 (2017) 構文構造を活用した学術論文における頻出コリゲーションの抽出，英語コーパス学会第 43 回大会.
③ 天野翼，宮崎佳典，田中省作，長谷川由美 (2018) 構造的簡略化を用いた例示型英文書作成支援 Web アプリケーションの開発と評価，2017 年度 JSiSE 学生研究発表会.
④ 徳見道夫，田中省作，田辺利文，宮崎

佳典（2017）係り受け構造に基づいた実例文の学術分野共通性の推定，平成29 年度電気・情報関係学会九州支部連合大会．

⑤ 宮崎佳典（2017）コーパスを用いた技術英文書作成援用ツールを用いた実験とその評価，言語研究と統計 2017．

⑥ 田中省作（2017）構文構造の変化情報（構文変化)が付与された校正英文対データベースの試作，言語研究と統計 2017．

⑦ 田中省作，宮崎佳典，坂本泰伸，日野友貴，岡田毅（2016）ハイライティングを活用した英語リーディング授業向けCMCシステムiBELLEsの開発，教育システム情報学会第 41 回全国大会．

⑧ 宮崎佳典（2016）リーダビリティ式自動生成による英文リーディング用 e ラーニングソフト開発，東北大学研究科共催合同シンポジウム「ICT を利用した英語教育支援ツールの開発とその活用方法」．[招待講演]

⑨ 宮崎佳典，田中省作（2016）コーパスを用いた技術英文書作成援用ツールの開発とその評価，言語研究と統計 2016．

⑩ 天野翼，渡部孝幸，田中省作，宮崎佳典（2016）共起関係ならびに構文情報を考慮した英文汎化と英作文支援，2015 年度 JSiSE 学生研究発表会（東海地区）．

⑪ 田中省作（2016）学術情報マイニング，第 4 回九州大学異分野融合テキストマイニング研究会シンポジウム．[招待講演]

⑫ 戸沢信晴，宮崎佳典，長谷川由美，田中省作（2015）コーパスを用いた技術英文書作成援用ツールの開発とその評価，日本e-Learning学会2015年度学術講演会．

〔図書〕（計１件）

① 石川有香，石川慎一郎，清水裕子，田畑智司，長加奈子，前田忠彦（編著）田中省作，宮崎佳典他 19 名（著）(2016) 言語研究と量的アプローチ，307 (145-155，229-240)，金星堂．

〔その他〕

６．研究組織

(1) 研究代表者

徳見　道夫（TOKUMI，Michio）
九州大学・大学院言語文化研究院・名誉教授
研究者番号：90099755

(2) 研究分担者

田中　省作（TANAKA，Shosaku）
立命館大学・文学部・教授
研究者番号：00325549

宮崎　佳典（MIYAZAKI，Yoshinori）
静岡大学・情報学部・教授
研究者番号：00308701