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

令和　　元 年　　６ 月 １０ 日現在

機関番号：１４４０１
研究種目：基盤研究(C)（一般）
研究期間：2015～2018
課題番号：１５Ｋ０２７１７

研究課題名（和文）分野別特徴語の共起情報を基にした用例表現検索ツールの開発と論文作成支援への応用

研究課題名（英文）Development of an expression search tool based on field-specific collocation information and its application to academic journal article writing

研究代表者

今尾　康裕（Imao, Yasuhiro）

大阪大学・言語文化研究科（言語文化専攻）・准教授

研究者番号：５０６０９３７８

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：本研究では、英語で書かれた工学分野の入門書の電子化および研究誌の論文からテキスト抽出処理をして工学英語コーパスを作成した。そして、分野特徴語を統計的に抽出してその共起語リストを作成したものと、コーパスを文法タグづけして、単語間の文法的関係をもとに共起語を抽出する方法を比較した。統計的に抽出する場合にはデータベース登録前に手間がかかることや、文法情報に基づく特徴語抽出の有用性が高いことがわかったことから、文法的共起情報に基づくデータベースを作成した。

研究成果の学術的意義や社会的意義
大学における英語学習は高学年時においては、すべてにおいて高いレベルを目指すよりも、分野ごとの語彙文法知識や特定のタスクをこなせる能力を身につけることが現実的である。特に専門分野の語彙に関しては、これまで専門用語の単語リスト作成などは行われてきたが、共起情報まで含めたものは多く見られなかった。それらを考慮したデータベースを作成し、学習に必要な共起語を抽出したり、論文作成時に、特定の語の共起語を検索できることは、大学生・大学院生のみならず、研究者に取っても有用なものであろう。

研究成果の概要（英文）：In this study, the researcher digitized introductory engineering textbooks and collected engineering research journal articles and built a corpus of engineering text. From the corpus, the engineering keywords and their collocates were statistically extracted to create a database of field-specific keyword collocations. This was compared with a database created with collocation information based on the dependency-grammar. Since the statistical extraction needs preprocess of the data before storing on the database, and the databased based on the grammatical collocation information was found to be more useful, the final database was created with grammar-relationship collocations.

研究分野：英語教育

キーワード：アカデミックライティング　ツール開発　共起語検索

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

(1) 大学、特に理系での英語教育では、4 年次や大学院、その後の研究者として必要となる、分野の専門性を反映した学術的な英語や内容を扱う重要性が増しているが、1 年次・2 年次の英語教育ではほとんど分野の専門性に重きを置かれていないのが現状であり、一部の理系中心の大学を除いて、高学年次の専門的な英語教育は限られていた。

(2) 日本のみならず海外においても、学術的な英語の語彙リスト（単語・連語）を作成する研究が多く行われていたが、いずれも学術分野をまたがって高頻度に用いられる語彙をまとめたものが多く、実際に論文作成などをする際に必要な分野に特化したリストはほとんど公表されていなかった。また、いずれもリストは作成されているものの、それを学習や論文作成時に活用できる形では提供されておらず、有用性という点で問題があった。

２．研究の目的

(1) 工学分野の英語で書かれた入門的な教科書および学術論文を電子化してコーパスを作成し、特徴語およびそれらの共起語を抽出する。

(2) 工学分野の用語の共起語を検索するためのデータベースを作成し、検索可能なシステムを構築する。

３．研究の方法

(1) 欧米の大学で推薦されているものやそれらの大学の工学系の学部のシラバス、研究代表者の本務校でのシラバス等を参考に、理工学分野の入門的な英語で書かれた教科書を選択し、スキャン後 OCR にかけて電子化処理をしたのちに、本文部分のみを抽出して必要な修正を行い、工学系教科書コーパスを作成した。

(2) 工学系論文誌のランキングや、工学系学術英語を分析した研究などを参考にしてコーパスに含める論文誌を選び、その学術論文を収集して本文部分を抽出し、必要な修正を行った上で、論文コーパスを作成した。

(3) 教科書コーパスおよび論文コーパスから研究代表者が開発している macOS 用コーパス分析アプリケーションである CasualConc に必要な機能を追加した上で、キーワード統計（対数尤度比）を用いて汎用書き言葉コーパスである BNC の書き言葉サブコーパス（BND Written）に対して特徴語を抽出し、CasualConc で TreeTagger を利用して品詞タグづけした上で共起統計（対数尤度比）を用いて抽出した共起語と、Stanford CoreNLP（Manning et al., 2014）を利用して依存文法（dependency-grammar）による文法タグづけを行って抽出した共起語を比較を行なった。

(4) 共起語検索のためのデータベースを SQLite を利用して作成した。

４．研究成果

(1) ウェブ上の大学の工学系の教科書採用ランキングや研究代表者本務校やその他の日本の大学のシラバスを確認して、大学の授業で採用されているものを含めて、基礎科学および基礎工学の入門的な英語で書かれた教科書を 20 冊ほど選び、そのうち分野の重複などを考慮して 10 冊ほどの電子化を進め、最終的に 8 冊の本の電子化が完了し、1 冊の電子化が未完となった。科学・工学系の教科書では変数名や化学式、数式などが多用されており、それらは OCR での認識が正確ではなかったり、複数行に渡っていたりするなど、文字認識と確認に時間がかかった上、作業に当たった大学院生の専門は英語でのコーパス分析であったため、予定していたよりも時間が必要であった。

ここで作成した教科書コーパスに含まれるテキストは、基礎科学の教科書には、生化学、物理学、有機化学（未完）を含み、工学の教科書には、環境化学、流体力学、建築学、化工熱力学、固体物理学、生産工学を含む。コーパスの総語数は約 150 万語となった（表 1）。

表 1 教科書コーパス

	冊数	総語数	異なり語数
基礎科学	3（未完 1）	548,359	19,385
工学	6	1,029,848	27,094
合計	9（未完 1）	1,578,207	35,851*

*異なり語数は重複するものがあるため合計行の値は 2 つのサブコーパスの合計とは異なる

(2) 工学系論文コーパスは、当初の予定では 3–5 誌ほどを選び、それぞれ複数年にわたって収集しテキスト化処理を行った上でコーパスにする予定であったが、教科書の電子化処理や、そ

れと同じく抽出した論文テキストの数式や変数名などの修正処理に予想以上の時間が必要となったことから、Hyland and Jiang (2016) などでも工学系論文誌として分析対象とされた歴史のある論文誌である Automatica の 2017 年度一年分の論文のテキスト化が完了した。論文コーパスに含まれる論文数は 442 で総語数は 200 万語弱となった。当初の予定と異なり、1 誌のみとなったことから、論文コーパスの内容には分野としての偏りがあることは否定できない。

表 2 論文コーパス

	論文数	総語数	異なり語数
Automatica (2017)	442	1,951,188	25,189

(3) 教科書コーパス、論文コーパスそれぞれの単語リストから、BCN Written を参照コーパスとして対数尤度比を用いた特徴語抽出を行なった。表 3 は、対数尤度比上位の単語から、機能語（ストップワーズ）や figure, equation など、分野の特徴語というよりも学術文書全般で多く見られる単語を除いた上位 5 単語のリストである。教科書・論文コーパスのどちらのリストも、名詞が多くなっている。

表 3 教科書コーパス・論文コーパスのキーワード上位 10 単語

教科書コーパス		論文コーパス	
特徴語	頻度	特徴語	頻度
energy	7,006	theorem	5,021
surface	4,198	system	11,908
temperature	3,160	matrix	5,406
velocity	2,026	control	8,613
particle	1,814	function	5,793

次に、CasualConc 上で TreeTagger (Schmid,1995) を利用して品詞タグづけをしたコーパスで、CasualConc の Collocation ツール（共起語抽出ツール）を使って対数尤度比および頻度で抽出した特徴語の共起語と、Stanford CoreNLP で文法タグづけをしたコーパスから文法関係で頻度順に抽出した共起語を比較した。この処理を行うために、CasualConc には、TreeTager をアプリケーション上でインストールし、バッチタグ付け処理をする機能を追加した。

CasualConc の Collocation ツールでは、中心語の左右の指定範囲での共起語を検索でき、TreeTagger で単語に品詞タグづけしたコーパスとレマに置換してタグづけしたコーパスを用意し、左右 5 単語までの共起語の頻度を集計したのち、Word Count ツール（単語リスト作成ツール）で作成した単語頻度の情報を使って共起統計値（対数尤度比）を計算した。ここで特定の品詞タグのついた共起語を抽出するために、CasualConc には共起語の文字列を指定してフィルター処理をする機能を追加した。

本報告書においては、比較したすべての特徴語の結果を示すことはできないため、教科書コーパスでの特徴語として対数尤度比の値の最も高かった energy の共起語をのみを表 4 に示す。単語とレマのそれぞれの対数尤度比における共起統計と頻度の上位 5 単語までを掲載するが、be 動詞と have は、動詞と助動詞としての区別が品詞タグのみではできないため除外してある。文法関係による共起語は、範囲ではなく文法的な関係での抽出ができるため、energy を主語とした動詞と目的語とした動詞を分けて頻度上位のレマを示している。こちらは、動詞としての have と助動詞としての have が区別できるが、範囲指定の共起語との比較を行うため除いてある。文法関係による共起語リストの「主語」とは、energy が主語として用いられている動詞、「目的語」は目的語として energy が使われていた動詞のリストである。また、受動態と能動態での区別もできるため、受動態のものは be をつけて集計した。

表 4 教科書コーパスでの energy（名詞）の動詞の共起語

位置による共起語 (L5–R5)				文法関係による共起語	
単語		レマ		レマ	
対数尤度比	頻度	対数尤度比	頻度	主語	目的語
stored	stored	store	use	be transfer	use
transferred	transferred	transfer	store	be conserve	provide
transformed	associated	transform	transfer	be use	transfer
associated	required	associate	require	be transform	require
required	used	require	transform	depend	absorb

位置による共起語では、対数尤度比においては単語、レマのいずれも同じ単語が抽出されたが、頻度による抽出についてはほぼ同じ単語が異なる順位で抽出された。一方、文法関係による共起語のリストでは、energy が主語で用いられる場合は、受動態での使用が多いが、能動態の

depend も含まれた。目的語としてのリストは、位置による頻度集計とは異なる単語が抽出されている。形容詞の共起語は、名詞の場合は直前に位置することが多いことから、位置による抽出でも文法関係による抽出でも同じ単語が抽出された。他の特徴語でもこれと同様の傾向が見られた。

この結果から、用例検索のデータベースとしては、単純な位置での共起語よりも文法関係を考慮した共起語の方が有用度が明らかに高いことや、品詞タグと位置での抽出では抽出後に人の手で処理をしたものを記録する必要があるのに対して、文法関係タグをつけたものは、タグの精度の問題はあるにせよ、検索結果そのままでも情報として扱えることがわかったため、当初の計画とは変更して、あらかじめ特徴語の共起語を抽出してデータベースに記録するのではなく、Stanford CoreNLP による依存文法の文法タグをデータベースに記録し、コーパスに含まれるすべての単語の共起語を検索可能なシステムを構築することにした。

(4) Stanford CoreNLP では、依存文法による文法タグだけでなく品詞タグやレマ処理なども行えるが、コマンド入力が必要な CUI のアプリケーションであり、テキストファイルを読み込んでテキストファイルや XML 形式のファイルとして書き出すため、コンピュータになれたユーザ以外には使い勝手の面で問題がある。当初の計画では、特徴語や共起語を抽出したものをデータベースに保存して検索する予定であったため、市販の汎用データベースアプリケーションである FileMaker を使用する予定であったが、将来のデータ追加を考慮すると人の手を入れなければデータの追加が行えないことは現実的ではなく、また、単純に保存したものを検索して表示する以上の自由度を確保するのが難しいという点なども考慮して、SQL でデータベースを構築する方針に変更した。

データベースの構築は、単語、レマ、品詞タグのいずれでも検索ができ、さらには検索語の文法的な関係を検索できるように、単語、レマ、品詞タグ、文法タグをデータとして保持し、リレーショナルデータベースとして扱える SQL の強みを生かして複数の文法項目を組み合わせて柔軟な検索が可能となるように、スクリプト言語である Ruby で SQLite データベースを作成して検索のテストを行いつつテーブルの仕様を決定した。

実際のデータベースの作成は、今回作成したコーパスのテキストファイルからデータベースを作成して恒久的に使用するより、テキストデータを追加したり、別の分野のコーパスで同様の検索システムが作成したりできるように、データベースを容易に作成するための macOS 用 GUI アプリケーションを開発した。このアプリケーションでは、コーパスのテキストファイルを読み込み、Stanford CoreNLP でタグづけして XML ファイルとして書き出し、さらに、その XML ファイルを SQLite データベースに落とし込むという一連の作業を GUI で行えるようにしたものである。

最終的なシステムは、データベース検証のために Ruby を用いたため、そこでのスクリプトが転用可能であること、ウェブブラウザ上で SQL のデータベース検索ができ、しかも構築がしやすいことなどから Ruby On Rails で構築することにした。そのための検索 GUI のテストを行うため、CasualConc に今回作成した SQLite データベースを検索する機能を追加している（図 1）。

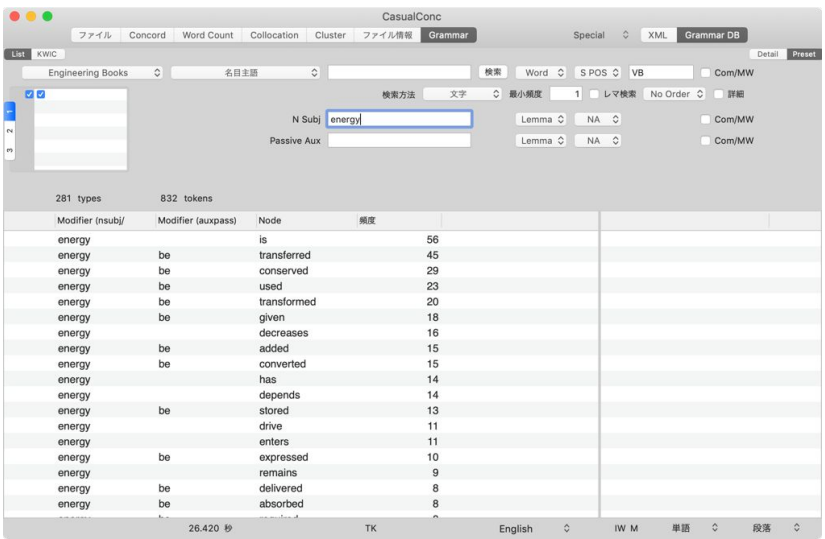

図 1 CasualConc での検索 GUI テスト

CasualConc は、内部のテキスト処理に Ruby を使っているため、Ruby で書いたテストのスクリプトが使えたことや、最終的なシステムである Ruby On Rails にも転用可能であるため、この

ような形のテストを行っている。CasualConc 上では、作成した SQLite の共起語データベースを登録し、複数の登録してあるデータベースから使用するデータベースを選び、文法関係を選択して検索語を入力し検索するような形式になっている。この機能を追加したバージョンは今後リリース予定である。

CasualConc に追加した機能では、将来研究用に使用できるように細かな設定をした上での検索が可能にしてあるが、一般の共起語検索のシステムとしての使用目的では操作が煩雑であるとのフィードバックを受けたため、最低限の設定で容易に検索できるインターフェイスが必要であることがわかった。

本研究の研究期間終了時には、日本語で検索をするためのデータベースの修正までは完了させたものの、ウェブインターフェイスは未完成である。現在 CasualConc に追加した機能のインターフェイスををもとにさらに使いやすいインターフェイスを作り、今後、さらにコーパスデータの拡充をするとともに、検索システムの完成および公開を目指したい。

＜引用文献＞

Hyland, K., & Jiang, F. K. (2016). “We must conclude that…”: A diachronic study of academic engagement. *Journal of English for Academic Purposes*, *24*, 29-42. DOI: 10.1016/j.jeap.2016.09.003

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings from Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*: System Demonstrations.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

５．主な発表論文等

〔雑誌論文〕（計 2 件）

① Mizumoto, A., Hamatani, S., & Imao, Y. (2017). Applying the Bundle-Move Connection Approach to the Development of an Online Writing Support Tool for Research Articles. *Language Learning*, *67*(4), pp. 885-921. DOI: 10.1111/lang.12250

② 水本篤・浜谷佐和子・今尾康裕 (2016).「ムーブと語連鎖を融合させたアプローチによる応用言語学論文の分析–英語学術論文執筆支援ツール開発に向けて」,『英語コーパス研究』, 23, pp. 21-32.

〔学会発表〕（計 3 件）

① Imao, Y. (2018). Going beyond simple word-list creation using CasualConc. *The 14th Conference of the American Association for Corpus Linguistics (AACL)*, Georgia State University, Atlanta, Georgia.

② 今尾康裕 (2016). Exploring corpus data with CasualConc. *ICAME 37*, The Chinese University of Hong Kong, Hong Kong.

③ 今尾康裕 (2015).「コーパス分析ツールの選択肢の一つとしての CasualConc」, 英語コーパス学会第 41 回大会、愛知大学名古屋キャンパス.

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。