【３】

氏名	後藤（ご とう）　一（かず）　章（あき）
博士の専攻分野の名称	博　士（言語文化学）
学位記番号	第　２２３９０　号
学位授与年月日	平成20年６月25日
学位授与の要件	学位規則第４条第１項該当 言語文化研究科言語文化学専攻
学位論文名	**電子コーパス解析に基づく英語コロケーション研究－コーパス言語学がもたらす効果的なコロケーション学習の可能性－**
論文審査委員	（主査） 教　授　細谷　行輝 （副査） 教　授　村岡　貴子　　准教授　田畑　智司

論文内容の要旨

本研究の目的

社会の国際化・情報化が進む中，英語は様々な領域において事実上の共通言語として用いられており，自らの意思や思想を英語で的確に表現するための発信能力の育成が強く求められている。こうした背景の下，単語と単語を適切に結合するためのコロケーション知識の習得が改めて重要視されている。1語の単語で伝達される情報量は少なく，その単語の状態や活動などを別の単語と結合することで詳細化し，それらを統合していくことが文の作成プロセスに他ならないためである。しかしながら，どのようなコロケーションを優先的に学習すべきか，という基本的かつ重要な点については未だ明らかにされていない。さらに，コロケーションの効果的・効率的な学習を支援するための学習リソースについても，ほとんど開発されていないのが実情である。この主たる原因として，コロケーションの使用実態を正確に解明できる分析手法が，コーパス言語学研究において確立されていない点が指摘される。そこで本稿では，効果的なコロケーション学習リソースの開発を視野に入れ，自然言語処理分野における言語解析技術を導入した高精度なコロケーション分析手法の確立を試みる。

各章の総括

第1章では，まず本稿の主要テーマである「コロケーション」の定義を行った。本研究では，学習者にとっての直感的な捉えやすさや，コーパス処理との融合性などを考慮し，統語的に関係付けられる単語と単語の結合をコロケーションであると定義した（「主語+述語」や「形容詞+名詞」など）。また，frequency-based approach と呼ばれる立場に基づき，コーパスにおける各コロケーションの生起頻度の多寡を，それらの重要性を測る指標として捉えることとした。続いて，英語学習におけるコロケーション学習の重要性について言及すると共に，それを効果的に行うための学習リソースが不足している現状を指摘した。これを改善するために，コロケーションの実態を正確に解明できるコロケーション分析手法の必要性を主張した。

第2章では，コロケーションの実態を解明するために，コーパスからコロケーションを機械的に抽出するための手法について議論した。最初に，コーパス言語学分野において代表的な分析手法である「KWIC Concordance の観察」「ノードスパン分析」「N-gram分析」を概観し，これらがいずれも単語の生起位置という表層的な情報に依存しているため，正確なコロケーション抽出のためには不十分であることを指摘した。高精度なコロケーション抽出を実現するためには，単語の生起位置ではなく，単語間の統語関係という深層的な情報に基づくことが不可欠であると考え，統語構造の解明が可能な

統語解析器を活用したコロケーション抽出手法を提案した。ただし，統語解析器を用いたコロケーション抽出が効果的に行われるためには，統語解析自体が高精度である必要がある。よって，最適な統語解析器を選定するため，Machinese Syntax と Link Grammar Parser という代表的な2種類の統語解析器の精度を比較した。その結果，Machinese Syntax の方がより高精度で解析が行われることが明らかとなり，本研究では Machinese Syntax を用いてコロケーション分析を行うこととした。

第3章では，具体的なコロケーション学習リソースとして，コロケーションデータベース，及びコロケーションリストの開発を行った。特に，学術分野に携わる大学生，及び大学院生にとって有用となるコロケーション学習リソースの構築を目標とした。コロケーションデータベースは，British National Corpus の一部である「BNCアカデミックコーパス」から4種類のコロケーション，すなわち「主語+述語」コロケーション，「述語+目的語」コロケーション，「形容詞+名詞」コロケーション，「名詞+名詞」コロケーションを，統語解析器を用いた手法によって網羅的に抽出したものとなる。本コロケーションデータベース開発の本来の目的は，コロケーションリスト作成のための基礎資料として用いることであったが，データベース内の項目を検索するプログラムと連携させることで，それ自体もコロケーション辞書として有効に活用される可能性が示された。一方，コロケーションリストは，学習者にとって特に重要度が高いと考えられるコロケーションのみを選定したものである。リストに記載されたコロケーションは，コロケーションデータベース内の項目から，「生起頻度 (frequency)」と「使用分布 (range)」を基準として選出されたものとなる。その結果，「主語+述語」，「述語+目的語」，「形容詞+名詞」，「名詞+名詞」の各統語構造別に，学術分野に携わる学習者にとって最低限必要となるコロケーションのリストが作成された。

第4章では，単語によって生起しやすい統語構造が異なる，という「語の文法」の考え方に基づき，200語の高頻度名詞を対象に，各名詞が典型的に生起する統語構造の解明を試みた。そのために，工学系英語論文で構成された「工学系コーパス」を統語解析し，各名詞の生起頻度を「主語としての名詞」や「前置修飾語としての名詞」など統語機能別に計測した。こうした各名詞の生起傾向を多変量解析によって統計学的に分析することで，200語の名詞が3種類のグループに大別化されることが明らかとなった。「グループ1」には「前置修飾語」として生起しやすい名詞群が類型化され，「グループ2」には「形容詞句内」または「副詞句内」に生起しやすい名詞群が類型化された。さらに，「グループ3」には，「主語」「(受動態)主語」「目的語」として生起しやすい名詞群が類型化されることとなった。また，各名詞が典型的に生起する統語機能を明らかにすることで，それらの重要なコロケーションが推定されることが示された。すなわち，「前置修飾語」として生起しやすい「グループ1」に属する名詞であれば後続する名詞とのコロケーション，「形容詞句内」または「副詞句内」に生起しやすい「グループ2」に属する名詞であれば前置詞とのコロケーション，また，「主語」「受動態主語」「目的語」として生起しやすい「グループ3」に属する名詞であれば述語動詞とのコロケーションに注目することで，各名詞の特徴的なコロケーションが抽出されることが示された。

第5章では，意味的に関係するコロケーション (これを「連関コロケーション」と呼ぶ) を学習者に提示することで，より効果的なコロケーション学習が可能になるとの考えから，コロケーション間における「連関関係」の機械的な推定手法を提案した。共起動詞の共通性が高い2語は類似していると仮定するDistributional Hypothesisに基づき，200語の高頻度名詞における連関関係を，共起動詞の一致度から推定した。具体的には，各名詞の共起動詞を「主語+述語」などの統語構造別に抽出し，その種類と頻度分布を変数として，各名詞間の「相関係数」を算出した。相関係数の高い名詞間を対象に，実際に連関関係が認められるかをシソーラスなどを用いて検証したところ，48組の名詞間に連関関係が存在することが示された。こうした連関関係にある名詞群が同一の動詞と共起している場合，名詞間における連関関係がそのまま拡張していると捉え，これらを連関コロケーションであると判断した。こうした連関コロケーションを第3章で開発したコロケーションデータベースに取り入れることで，コロケーションの自律的な学習の促進や，独力で思い至らなかったコロケーションの発見支援などの効果が見込まれ，有用なコロケーション学習リソースとして用いられる可能性が示された。

本研究の意義

本研究の目的は，学習効果の高いコロケーション学習リソースを開発するために，コロケーションの使用実態を多角的に分析する手法を確立し，コロケーションを使用する上で認識しておくべき「頻度」「使用域」「統語構造」「他コロケーションとの関係性」などのプロファイル情報を解明することであった。そのためには，英語教育分野，コーパス言語学分野，及び自然言語処理分野における各知見の有機的な統合が必要となり，こうした学際的研究の実現化を試みた点に本研究の意義が存在する。また，本分析手法は，主観的な判断に依存する部分を可能な限り縮小し，コンピュータ処理による作業の自動化を追及することで，客観性及び汎用性の高いものとなっている。従って，本稿では学術分野に携わる英語学習者，特に工学系分野の大学生や大学院生を対象とした学習リソースの開発を行ってきたが，こうした個別分野に限定されず，これをモデルとし，様々な分野にも応用が可能となっている。例えば，BNC における70分野ごとのコロケーションリストの開発や，分野別コロケーション間の「連関関係」の解明なども低コストで実現されると考えられ，コロケーション学習リソースの不足が課題となっている現状において，本手法がもたらす波及効果は極めて大きいと考える。

本研究がコロケーション研究のさらなる興隆と，コロケーション学習リソースの充実化の一助となれば幸甚である。

論文審査の結果の要旨

本論文では，「生起頻度」「使用域」「統語構造別使用傾向」「コロケーション間の意味的関係性」など，これまで実態が十分解明されていなかった英語コロケーションに関する様々な側面に焦点を当て，主に実証的な観点から，電子コーパス解析に基づくコロケーションの分析を行っている。

本論文は、5章からなるが、とくに、第3章、第4章、第5章について、略述すると、第2章では，統語解析技術を活用したコロケーション分析手法を提示した。統語解析を用いることで，従来の手法では困難であった「主語+述語」や「前置詞句」などの特定の統語構造に限定したコロケーション抽出が可能となり，さらに，コーパス内に生起するコロケーションの網羅的な抽出も高精度で行われることが示された。第3章では，コーパス内に生起する全コロケーションを抽出し，各項目の重要度を，その生起頻度と出現分布という指標によって推定し、これにより，無数に存在するコロケーションから特に重要な項目が選定され，効率的かつ効果的な学習が見込まれるコロケーションリストが作成された。第4章では，コーパス内に生起する高頻度名詞を対象に，各単語における統語構造別の生起頻度を計測し、その結果，主語・目的語として使用されやすい名詞もあれば，前置詞句内で使用されやすい名詞も存在するなど，同じ「名詞」というカテゴリーに属していても，その生起傾向は決して一様ではないことが明らかとなった。すなわち，典型的なコロケーションを解明するには，各単語の典型的な統語構造に即した分析手続きが重要であることが示された。第5章では，意味的に関係する単語同士はその共起語が類似するという仮説を応用し，相互に関連性のあるコロケーションをコーパスから機械的に抽出する手法について検討している。意味的に関係するコロケーションを学習者に提示することで，それらの項目の相乗的な学習効果が期待され，さらにそうした情報が活用されたコロケーション辞書などは，一種のコロケーションのシソーラスとして利用できる可能性が示されている。

コーパス言語学分野における既存のコロケーション分析手法を無批判に適用するのではなく，より的確かつ高精度な分析手法を追究し，プログラミング処理や統語解析技術を活用した新たな方法論を確立した点に本研究の意義が認められる。さらに，単に方法論の提示に留まるのではなく，具体的な成果として，重要項目を選定したコロケーションのリストやコロケーションのシソーラスなど，英語学習に有用と思われる学習リソースを実際に作成した点も評価に値すると言える。

以上により、本論文は博士号学位（言語文化学）論文として十分価値あるものであると認める。