様式 C-19

科学研究費補助金研究成果報告書

平成　22　年　3　月　16　日現在

研究種目：若手研究（B）
研究期間：2008 ～ 2009
課題番号：20720145
研究課題名（和文）　自然言語処理技術を利用した英語コロケーションリストの作成

研究課題名（英文）　Development of an English Collocation List Using NLP Techniques

研究代表者
後藤　一章（GOTO KAZUAKI）
摂南大学・外国語学部・講師
研究者番号：90397662

研究成果の概要（和文）：本研究では，コーパス言語学において伝統的な「数語以内の位置で生起する単語をコロケーションと見なす」という素朴な手法の問題点を指摘し，統語解析技術を利用したより高精度なコロケーション抽出手法を提案した。これにより，キーワードと直接的な統語関係を有する単語のみをコロケーションとして抽出することが可能となった。当該手法を利用することで，大規模な英語コーパスからそこに生起するコロケーションが網羅的に抽出された。すべてのコロケーションから特に高頻度で使用される項目を選定し，効率的な英語コロケーション学習のためのコロケーションリストを構築した。

研究成果の概要（英文）：In the field of corpus linguistics, a conventional method of collocation extraction, often called "the window-span model," has been widely acknowledged This method, however, fails to capture syntactic relationships of collocations extracted from computer corpora. To cope with this problem, this research has proposed to employ a parsing technology that can reveal syntactic structures of sentences automatically. The proposed technique enables more accurate and comprehensive extraction of collocations. From British National Corpus, I have extracted important collocations based on their frequencies and sorted them out to create a "collocation list." The list can be expected to be used for effective collocation learning.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,100,000	330,000	1,430,000
2009 年度	600,000	180,000	780,000
年度			
年度			
年度			
総　計	1,700,000	510,000	2,210,000

研究分野：人文学
科研費の分科・細目：言語学・外国語教育
キーワード：コロケーション，コーパス

１．研究開始当初の背景

英語学習者が自然な英文を作成するためには，語と語を適切に共起させるための知識，すなわちコロケーション知識の習得が不可

欠だと言われている。実際的な英語運用能力の育成が期待される昨今の英語教育において，コロケーション学習の重要性は一層高まっており，効果的なコロケーション学習のためのリソース開発は喫緊の課題と言える。

語彙学習リソース開発の分野では，コーパスという大規模言語データに基づいた「語彙リスト」の開発が注目されている。従来の人手による主観的な選定方法では，どうしても恣意的な側面が残される。これに対し，各単語の生起頻度を計測し，それらを重要度の指標として使用するというコーパスベースドな手法であれば，単語の学習優先度を客観的に推定することが可能となる。こうして，効率的な単語学習を促進する「語彙リスト」が広く開発されることとなった。

一方コロケーションに関しては，コーパスに基づいて開発された「コロケーションリスト」はほとんど見あたらない。

その原因として，コーパスからのコロケーションの抽出手法に関する問題点が挙げられる。コーパスに基づいてコロケーションの重要度を推定するためには，まずコロケーションをコーパスから正確に抽出する必要がある。コーパス言語学の分野の伝統的なコロケーション抽出手法は，キーワードの前後数語以内の単語を網羅的に抽出し，その共起頻度が高い単語を共起語として見なすというものである。この方法には，単語と共起語間の統語的な関係性を問わないという問題点がある。例えば，(1) の文で実際に統語関係があるコロケーションは gold + medal や won + medal などあるが，当該手法では medal + in や medal + figure なども抽出してしまう。こうした組み合わせは意味的な連想関係は有しているかもしれないが，統語的な関係性は認められないため，学習者に提示する形として適切とは言い難い。

(1) He won the gold medal in men's figure skating.

また，当該手法では win + victory に関して，直接的な統語関係が存在しない(2) と，直接的な統語関係が存在する (3) を区別することができないため，各コロケーションの正確な生起頻度が計測されないことになる。

(2) She experienced a sweet victory when she won the gold medal.

(3) He was guaranteed to win a complete victory.

このように，従来のコロケーション抽出手法は言語学習リソース開発の点からは不十分であると考えられ，これがコーパス研究分野においてコロケーションリストの作成が推進されてこなかった原因だと推測される。

そこで本研究課題では，高精度なコロケーション分析手法を確立し，効果的なコロケーションリストを開発することを目的とする。

２．研究の目的

(1) コロケーション分析手法の確立

文の統語情報に基づき，コーパスからコロケーションを正確かつ機械的に抽出する手法を確立する。そのために，自然言語処理技術，特に統語解析技術について調査を行い，コロケーション分析への応用の可能性を探る。

(2) コロケーションリストの作成

コーパスからコロケーションを抽出し，各項目の生起頻度を計測する。それらの頻度を手がかりに各項目の重要度を推定し，学習用コロケーションリストを構築する。

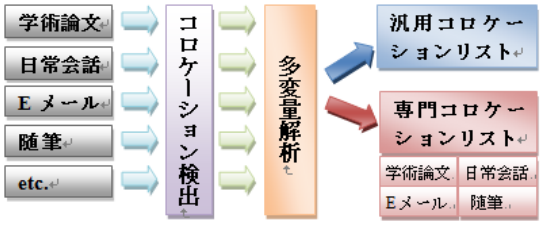
３．研究の方法

本研究課題において重要な役割を担うのが文の統語構造の検出技法，すなわち統語解析である。統語解析を行うツールは統語解析器と呼ばれ，現在複数の解析器が開発，公開されている。これらの中から，本研究のコロケーション抽出処理に最適な統語解析器を調査した。特に，コーパス言語学研究への利用という観点から，解析精度だけでなく，その操作性も考慮して選定を試みた。

統語解析器によってコーパスに統語解析を行い，それに基づいてコロケーションを機械的に抽出した。特に，本研究ではコーパスデータとして約 1 億語の British National Corpus（BNC）を使用した。抽出された各コロケーションの生起頻度を計測し，各項目の重要度を推定した。

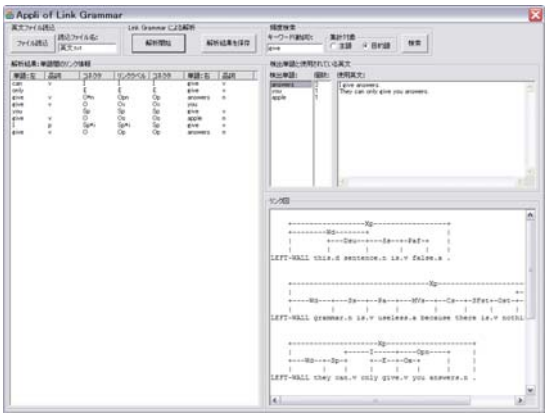
また，各コロケーションの重要度を言語使用域の観点からも推定した。言語使用域とは，適切な使用が可能な範囲のようなものであり，様々な状況で幅広く使用されるコロケーションもあれば，特定の状況でのみ使用されるコロケーションも存在する。母語話者であれば言語使用域を柔軟かつ適切に使い分けられるが，学習者にとってこれを自由に行うことは難しい。そこで，初中級の学習者には，多様な状況で幅広く使用される汎用的なコロケーションを学習することが効率的であると考えられる。

BNC は様々なジャンルのテキストで構成されているため，各コロケーションの総頻度に加え，各ジャンル別の生起頻度も合わせて計測し，言語使用域の調査を行った。こうして，複数のジャンルに共通して生起する汎用的なコロケーションと，特定の分野に特化して生起する専門なコロケーションを明らかにした（下図参照）。



４．研究成果

まず，コロケーション抽出に最適な統語解析手法を調査するため，Apple Pie Parser, Link Grammar Parser, Machinese Syntax, Enju という 4 種類の統語解析器によるコロケーションの抽出実験を行った。その結果，解析精度には顕著な差異は見られなかったが，出力形式や操作性に少なからず差異が見受けられた。本研究課題の目的に照らし合わせると，単語間の直接な統語関係を捉えることが重要となる。また，操作性が簡便であることも重要となる。以上の 2 つの観点から，Machinse Syntax が最適であると判断した。ただし，Machinse Syntax は年度ごとのライセンス契約が必要であり，継続的な使用を行うことは難しい。そこで，Machinse Syntax をメインの統語解析器として使用しながらも，今後を見据え，Link Grammar Parser の改良も合わせて取り組んだ。Link Grammar Parser はオープンソースの統語解析器であるため，自由にプログラムに組み込むことが可能である。そこで，情報科学分野の研究者の協力を得て，Link Grammar Parser を利用し，以下のような GUI 型コロケーション分析プログラムを開発した。本プログラムによって，簡易なコロケーション抽出手法が実現され，コーパス言語学分野におけるコロケーション研究のさらなる発展が期待される。



統語解析器の選定に続いて，Machinese Syntax を用い，British National Corpus に統語解析を行った。解析結果からコロケーションを網羅的に抽出し，その中から特に，名詞と動詞によって構成されるコロケーションの生起頻度を計測した。また，総頻度とは別に，British National Corpus における各ジャンルごとの生起頻度を測定した。

名詞と動詞で構成されるコロケーションの中で，特に「述語＋目的語」の統語関係で共起するコロケーションに着目した。このコロケーションは約 17.5 万種類の項目が抽出された。ただし，各項目のジャンル別の頻度を観察すると，多数のジャンルに共通して生起するコロケーションの数は非常に少なく，約 17 万種類のコロケーションは，共通して生起するジャンルの数がわずか 3 以下であった。また，このうちの約 15 万項目は，わずか 1 ジャンルのみにしか生起していなかった。一方，全ジャンルに共通して生起するコロケーションは 254 項目であった。複数のジャンルに共通して生起するコロケーションは文脈を問わず自然に使用できる可能性の高い汎用的な項目として捉えられ，学習者が優先的に習得すべき項目であると考えられる。こうした汎用的コロケーションの例を以下に示す。

answer + question
ask + question
do + job
draw + attention
find + way
form + part
give + chance
give + opportunity
give + way
have + access
have + advantage
have + difficulty
have + effect

一方，専門的なコロケーションとしては以下のような項目が挙げられる。特徴として，汎用的コロケーションと比べると，名詞が具体的な事物であることが多い。これは，各専門領域における用語の使用方法を把握する上で有用となる。

add + garlic
add + tax
develop + software
integrate + system
increase + profit
preheat + oven
run + application

こうして，生起頻度や使用域の情報に準拠した，2 種類のコロケーションリストが構築された。両者の性質はそれぞれ異なり，汎用的コロケーションは可能な限り多数の項目

を習得することが望ましく，専門的コロケーションについては，各学習者が携わる分野に特化して習得することが効率的と言える。

本研究の意義は，英語コーパス研究分野において従来見過ごされてきた統語解析手法に焦点を当て，新たなコロケーション分析手法を提案したことにある。統語解析によるコロケーション分析がコロケーション学習リソースの開発に効果的であることを示したことは，当該分野に重要な貢献を果たしたと考えられる。また，本コロケーションリストは，数十万単位のデータ量を備えているため，学習リソースとしての役割だけでなく，他のコロケーション研究においても有用な基礎データになることが期待される。

なお，現状は希望者に対してのみデータの配布を行っているが，より広く配布するため，Web サーバから自由にダウンロードできる環境を構築する予定である。

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 4 件）
①後藤一章、歴代米国大統領就任演説における統語構造の変異、英語コーパス研究、査読有、17 号、2010 年、pp161-175.
②後藤一章、BNC の言語使用域間におけるコロケーションの共通性と多様性、統計数理研究所共同研究リポート、査読無、231 号、2009 年、pp.63-76.
〔学会発表〕（計 7 件）
①後藤一章、多変量アプローチに基づく BNC における名詞と使用域の分析、言語研究と統計、2010 年 3 月 27 日、大妻女子大学.
②森真幸、後藤一章、竹蓋順子、Link Grammar Parser を活用した英文構造出力ソフトウェアの開発、外国語教育メディア学会、2009 年 8 月 5 日、流通科学大学.
③後藤一章、構文研究と統計、日本英語学会、2008 年 11 月 16 日、筑波大学.

６．研究組織
(1)研究代表者
後藤 一章（GOTO KAZUAKI）
摂南大学・外国語学部・講師
研究者番号：20720145