様式 C-19

科学研究費補助金研究成果報告書

平成２２年　５月２７日現在

研究種目： 基盤研究（C）
研究期間： 2007～2009
課題番号： 19530773
研究課題名（和文）　汎用的データベースの解析処理に基づいた英語構文指導用教材作成システムの開発

研究課題名（英文）　The System Development for the English Sentence Construction Teaching Material based on the Analysis of the General-Purpose Database

研究代表者
岡田　毅（OKADA TAKESHI）
東北大学・大学院国際文化研究科・教授
研究者番号： 30185441

研究成果の概要（和文）：研究代表者と研究分担者の研究室を仮想ネットワークで繋ぎ、分散型処理を実現したリレーショナルデータベースを基幹に据えた英語構文解析用システムを構築し、既存の大規模コーパスおよび研究者、教育者、学習者が独自に選定し構築したコーパスを統一的な形式で蓄積し、動詞の活用形に焦点を当てることから明らかになる日本人学習者に特徴的な英語構文の性格を抽出し、効果的な指導方法と教材の開発に寄与することを目指した。

研究成果の概要（英文）：The labs of the researcher and co-researcher are connected via virtual private network through which a general-purpose database was designed in order to developed a robust system for the statistical analysis of the English sentence constructions written by Japanese writers with reference to the verb conjugation. Not just existing large-scale corpora but small corpora collected and developed by individual researchers, teachers or learners are compiled in the relational database.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	800,000	240,000	1,040,000
2008 年度	1,000,000	300,000	1,300,000
2009 年度	1,400,000	420,000	1,820,000
年度			
年度			
総　計	3,200,000	960,000	4,160,000

研究分野： 社会科学
科研費の分科・細目： 教育学・教科教育学
キーワード：データベース、ネットワーク、分散処理、教材開発

１．研究開始当初の背景

(1) 英語教育用コーパスの相関と整備： 収録数１億単語を超える the British National Corpus(以下 BNC)をはじめとする大規模な英語コーパスを活用した実際の英語の用法に関わる研究は、日本国内でも世界的水準で目覚しい進展を遂げている。母語話者の言語事象を網羅的に提示するこの種のコーパスとともに、日本でも近年は学習者コーパスの整備開発が進んでいる。学習者コーパスは、英語非母語話者である学習者の産出データを体系的に蓄積しそれを分析することによって、効果的で質の高い教材へのフィードバックを企図するもの

である。このような相関関係を示すコーパスの中にあって、学習者への直接的なインプットとなる教材としての教科書の、明確な設計基準に従ったコーパス化が遅れているのが事実であり、本研究で目指すような、個別の研究者、教育者、学習者が収集し構築した英語コーパスの、BNC のような既存で大規模なコーパスとの形式面、属性面での整合性が強く求められるところである。

(2) 母語話者コーパス分析ツールの不備とその克服： 例えば日本人中学生に対して、英語母語話者コーパスと有機的な関係を持った教材とその開発が上述の相関関係からも不可欠であることは明らかであるが、現在世界規模で公開されている BNC や Word Bank 等の大規模コーパスそのものも、またそれらを統括的に処理可能な新しいシステムである PhraseBox などは、必ずしも外国語としての英語教育(English as a foreign language, EFL)を前提として構築されているわけではないし、これらに付属する分析用の各種システムやツールが英語教材作成の立場からの要求に十分に応えられているわけではない。例えば、BNC の日本語版検索プラットフォームである「小学館コーパスネットワーク」上のシステム SAKURA でも、ある階層までの検索は可能であっても、出力された検索結果の一部分を新たな検索キーとして次のステップに進めないという、英語学研究者および英語教育研究者にとっては非常に硬直化した分析ツールしか提供していないといえる。これらをはじめとする様々な分析ツールの不備を一挙に克服するための多面的なアプローチが必要である。そのためにはリレーショナルデータベースマネジメントシステム(RDBMS)の発想と手法とがぜひとも必要となる。

(3) 汎用的データベース及び分散処理系の必要性: CD-ROMやインターネット上から提供されている既存の母語話者コーパスに加えて、英語教育研究者や英語教員が EFL 教育の現場における教材及び試験問題等として収集し加工したいと考える膨大な量の英語データを、明確な設計思想に基づいた枠組みのデータベースの中に取り込んだ上で、例えば Web ベースで稼動する柔軟な分析処理を実現させない限り、中途半端で実際の教材としては使えない、単なる「膨大な英文データの蓄積」に陥ってしまう。日本の中学校等における実際の EFL のニーズを反映したデータ収集とその処理系を実現させるために、これまでにない広汎な領域をカバーする「教材用素材収集及び解析」・「教材自動作成」のための分散システムが必要となる。

２． 研究の目的

本研究の目的は以下の 3 点に要約することができる。

(1) 研究代表者・研究分担者の大学研究室間を仮想プライベートネットワーク(virtual private network, VPN)で接続し、シームレスなシステム開発の基盤を開発するための通信技術の確保し、多様な構造と付加情報を伴った多くの既存コーパスおよび個々のユーザーが求める形式や属性を柔軟に反映することのできる RDBMS 開発に繋げること。これは、これまでに指摘した、汎用性を保証した柔軟なデータベースシステムの設計と開発の必要性が大きな背景となっている。

(2) 既存の大規模コーパスの分析を通しての、開発に必要なデータベース設計指針の確立。特に BNC に典拠を求め、その属性付与や文書管理の手法を体系的に精査し、そこから本研究にとって求められる汎用的な性格を検討する。特殊な研究目的にも、英語の代表性(representativeness)を追求するというような幅の広い研究にも供することのできる属性体系をコーパスに一律に与える、という発想を覆し、利用目的や分析目的に即した属性、とりわけ英語の品詞(part-of-speech, POS)標識(tag)をユーザー側で自由に定義し、データに付与するという概念を採用した。この際に、ユーザーの要求レベルの多様性が予想され、それに対応するための RDBMS 内でのデータテーブルの緻密な設計を目指した。

(3) 日本人 EFL 学習者に対する指導において、システム的にその教育用素材を抽出することの困難な、構文にかかわる特徴的な情報を、特に動詞の活用形の統計的分析に着目することによって可能とすることを目指した。構文解析(parsing)や意味標識(semantic tag)付与の研究は、近年、英国 Lancaster 大学、米国 Pennsylvania 大学等において盛んであるが、本研究では敢えて品詞標識(POS tag)のみの付与と連鎖解析に基づいて、英語の構文特徴を抽出することを目指した。このために、新しい標識付与プログラム(TAGASS)の開発を企図した。これに前後する形で、ユーザーが収集した幅広い使用域(register)に属するデータ英文を統一的な形式で蓄積するためのトークン化(tokenizing)の一環としての行整形プログラム(Line Formatter, LF)の開発と運

用を目指した。

３．研究の方法

(1) 汎用的コーパス・データ英文処理システムの構築: 研究組織の各ネットワークをVPN(Virtual Private Network)を利用することによって仮想的に結合する。これによって、東北大学で推進する(i)汎用的なデータベースに要求される属性や分析形態の研究、(ii)Web のみならず、印刷物も含めた多様な目標英文の収集と統一フォーマットによる蓄積作業と、東北学院大学で推進する(i)データベース用計算機の整備、(ii)データベース体裁統一のためのフォーマッタプログラム及び各種の処理系プログラムの開発という 2 つの連携作業が極めて効率的に行われることとなり、不特定多数のユーザーに対する柔軟で強力な教材作成支援サービスを提供することを可能ならしめる。

(2) 質的向上に寄与する教材収集・教材作成システムの開発と Web 上での公開: 教材作成者のニーズを幅広く反映したインターフェイスを装備したシステムを開発する目的を達成するためには、多様な利用目的をもったエンドユーザーの層や、利用形態についての考察が必要となる。これに基づいて、日本の EFL 教育を支援するためのシステムに、独自に要求される要素を明らかにすることになる。また、ICAME 等の海外で開催される国際学会において研究成果を積極的に公表し、フィードバックを得ることに努める。例えば学習者の習熟度や勉学意欲等を最も正確に把握しているのは、教材や試験作成者である教員自らであり、その意見を集約した上での支援システムを Web 上から開発することによってより大きな貢献が期待できる。

(3) 英語の動詞活用形とそれが用いられる構文の相関関係を網羅的に把握した研究はこれまでに存在せず、本研究でのように、さらにその研究結果を日本における EFL 教育の構文指導に応用することには斬新で大きな意義がある。動詞の頻度のみに着目するのではなく、個々の動詞の活用形(BNC では 6 種類)中の最多のものを学習の早い段階で提示する必要がある。しかし、単独のリストなどで活用形を示すことは、特に中学校レベルの EFL においては無意味であり、それらは『学習指導要領』で示されるような、実際の言語使用の現場に則した英語表現の中で「学習段階に応じた」理解の度合いに準じた構文という環境内で提示されるべきである。これらの点に配慮しながら、日本人研究者の産出する英文における構文的な特徴を、動詞活用に注目することによって明らかにする。

４．研究成果

(1) 研究の全体像: 信頼性の高い RDBMS を中心に据えた新しいコーパス解析システムを開発することが主眼となり、そのシステムによって、英語の構文指導用の効率的な教材が作成できるような斬新な知見を得ることも射程に含まれる。

(2) 既存コーパスと独自コーパスの整合: BNC や Word Bank のような既存の大規模コーパスの構造を精緻に分析し、それらを汎用性のある形式に整えてデータベースに蓄積し、これと並行して、個々のユーザーが収集した独自の英文データに対して品詞標識のような属性を付与し、既存コーパスと同一の形式でデータベースへ蓄積する。

　独自のデータをコーパス化する際には、「1 行・1 文」形式にレコードを統一するためのプログラムと品詞標識付与プログラムが重要な役割を果たすが、とりわけ品詞様式付与に関しては、新しい概念に基づいた手法を構築した。これはリレーショナルデータベースにおけるテーブル間の関連性という特徴を最大限に活用したものであり、これによって、既存のコーパスに付与された数多い標識のセットに準拠できることはもちろん、ユーザーが自らの分析目標や学習到達度に基づいて独自の標識を自由にデザインすることが保障される。また、個々のコーパスを別個に扱い、アドホックな解析を施すのではなく、本研究でのシステムにあっては、コーパスは RDMBS 中でドキュメントとして管理されることから、ユーザーが任意のコーパス(群)を指定し、それをコーパスセットという形で統計処理の対象として指定できるという画期的な特徴がある。

(3) 研究の推移と成果: 研究初年度に、YAMAHA 社製 VPN ルータ RTX1100 を用いて、東北大学と東北学院大学の 2 研究室の間に IPSec を利用した仮想ネットワーク（VPN）を構築した。これによって、物理的に離れた二つの研究組織を、単一のセキュアーなネットワークとして融合する事に成功した。VPN 環境下では、すべてのネットワークパケットが暗号化されるので、通信するデータの隠匿性が確保されるが、その暗号化に必要となるオーバーヘッド時間は、ネットワーク環境を利用した共同開発に際しては重要な要素となる。そこで、実際に構築したネットワーク環境下で、この VPN を利用した通信性能の測定を行った。この測定では、両研究室間のネット

ワーク通信性能が、ほぼ当初の予測通り（数 M byte/sec）の速度に達することを確認し、VPN 環境下において共同開発を進める事に支障が生じないことを明らかにした。

(4) データベースのテーブル設計: セキュアーなネットワーク環境を利用して研究グループで打ち合わせを進め、コーパス解析システムの中核を占める RDBMS (PostgreSQL)のテーブル設計を進めた。この設計では、英語コーパス研究の一つの標準と見なされている British National Corpus(BNC)を参考として、その BNC に取り込まれている全ての英文書とその単語に付与されている品詞を RDBMS に取り込む事を第一の目標として設定した。BNC には、約 4,000 余りの英文書が蓄積されており、それぞれの文書の容量が 1 メガバイト以下のサイズであるものが 98%とその大半を占めている。また、この文書に含まれる単語には約 70 種類の品詞タグが付与されている。コーパス解析システムで取り扱うデータセットには、非常に特殊な場合を除いて、BNC の場合と同じように、蓄積する文書個々の容量は小さいが蓄積する文書の数は多く、その結果として総量が増加する傾向にあると推測される。このような特徴を持つデータセットに対して、利用者が解析対象となる文書を独自に管理する方法及び、利用者が使用する品詞タグセットを自由に管理選択できる方法を、テーブル設計から可能にする事を明らかにした。

(5) アプリケーション開発: コーパス解析システムのコンテンツとなる英文データの収集方法には、前述の BNC の様な既存コーパスのデータを利用する方法と、WEB 等の英文書（プレインテキスト）を蓄積する二種類の方法に大別することが可能である。前者の場合、既存コーパスのフォーマットを我々の設計した RDBMS のテーブル形式のフォーマットに変換するアプリケーションが必要となる。特に BNC のデータを使用する際には、このフォーマット変換のアプリケーション（BNC アナライザ）を開発し、フォーマット変換を行って実際の解析システムの RDBMS に英文書の蓄積を行った。我々の RDBMS のテーブル設計では、利用者からの幅広い検索に柔軟に対応する事を目指しており、文書の構造を、単語、文、文書といったように階層化して表現をしている。BNC のデータ蓄積に際しては、RDBMS に蓄積されているデータの容量の増加に伴い、文書の蓄積に要する時間の大幅な増加が発生した。今後、この問題を解決するためには、文書構造の簡略化や RDBMS 自身のシステムチューニングも必要であることが本研究によって示唆された。

一方、プレインテキストを我々のコーパス解析システムで利用するには、文書中の個々の英単語に対して品詞を付与した後に RDBMS に蓄積する事が必要不可欠である。そこで、英文書を一行一文の形式に成形する Line Formatter(LF)と、LF によって成形された文書中の単語に対して品詞を自動的に付与するアプリケーション Tag Assigner(TAGASS)のプロトタイプ版を、Java 言語を用いて開発した。LF や TAGASS における文字列処理や品詞タグの自動付与に関するアルゴリズムは、その最適解が未だ確立されていない。この点を配慮して、LF や TAGASS の処理ルーチンは、フィルターと呼ばれる概念で定義されており、Java のダイナミッククラスローディングの機能を利用して、アプリケーションの実行時に、処理方法の内容や順序を変更可能に設計されている。これにより、プレインテキストを我々のコーパス解析システムに対して蓄積するだけではなく、文字列処理や品詞タグの自動付与に関する最適解自身の探求にも、これらのアプリケーションを利用する事が可能である。

このように、RDBMS のテーブル設計から始まり、コーパス解析システムにデータを蓄積することが可能なプロトタイプ版のアプリケーション開発を終了した。このプロトタイプ版の解析システムを利用して、一般外部利用者に対するサービス公開に際して、システムが実装すべき最低限の機能に関する考察を進めた。特に我々が注目したのは、データベースに蓄積されている文書の利用を中心とする単純な（初心者的な）利用者と、データベースに対して逐次データを蓄積して自分の研究に利用する研究者レベルの利用者の二種類の利用者である。単純な利用者に対しては、WEB ブラウザの様な一般化されたインターフェイスを利用して、極力機種依存性を排して利用できることが望ましい。また、コーパス解析システムの出力する一次の解析結果は主に統計解析の結果となるが、これを柔軟に解釈して二次的な情報を作りだすようなユーザインターフェイスも必要になる。研究者レベルの利用者に対しては、柔軟なパラメータを設定しながら英文書を蓄積することが必要になるため、WEB ブラウザの利用よりも、独立した Java のサーバクラインとシステムが必要になると考えられる。また、英文書を蓄積した後に、登録した品詞情報を再編集するような枠組みを提供することで、利用者の要求に

対して柔軟に対応できることが明らかとなり、このインターフェイスに関する実装設計は今後の課題となっている。

(6) システムを用いた分析実例: 以下は、2009 年 7 月にロンドン大学で開催された第 3 回国際現代英語言語学会(International Conference of Linguistics of Contemporary English (学会発表③))での研究発表を一例とし、本研究で培われ開発されたコーパス収集・構築・整備・分析という一連のシステムによる英語構文における非英語母語話者の使用傾向分析を紹介する。

この研究では、主に伝達動詞に焦点を当て、それらの動詞が英語の学術論文で用いられる際の、英語母語話者である研究者と、非母語話者である日本人研究者が採用する構文上の特徴を対照的に明らかにした。コーパス研究の分野にあって近年、大きな注目を浴びつつある「特定領域・目的における英語」(English for Specific Purposes)の、さらに具体的な下位分野としての「学術目的の英語」(English for Academic Purposes)にその分析対象を限定することは、本研究において開発された RDBMS を中心とし、コーパスを柔軟に管理することによって、分析の対象を「コーパスセット」という単位で抽出・管理可能なシステムによるものである。

英語の構文をコーパス分析の立場から把握する際には、従来では構文解析(parsing)に基づいたアプローチが主流であるが、ここでは動詞の活用形(BNC 等で採用されている Lancaster 大学の研究チーム UVREL 開発の品詞標識セットでは 6 種類の活用形が認定されている)に着目することによって所期の対照的特徴抽出という目的を達成しようとするものである。

動詞の活用形を示す 6 種類の品詞標識の連鎖に分析を集中し、そこから日本人研究者に特徴的な英語構文の使用特徴を数量的に算出し、さらにはこのような傾向が生じる原因を、日本国内で広く販売されている各種の英文作成マニュアル本を吟味することによって質的に解き明かそうとした。結果的に、*it is suggested that...* というような構文が、ある意味では定型表現のようにマニュアル本で取り上げられ、文法的には正しくとも同じ構文が圧倒的に多用されるという傾向が浮き彫りとなった。この発表に対しては国際会議参加者の中でとりわけ非英語母語話者の研究者からの反応が大きく、どうしても定型表現に頼りがちになるのは非母語話者に共通する特徴である可能性と、その定型表現の種類の多様性に関しての、書き手の母語圏内で流布している指導書や教科書の与える影響に関しての将来の研究につながる議論と示唆を得ることができた。

本研究で目指した目的の一つに、多様なシステム利用者のニーズに柔軟にこたえるシステムの開発があった。これは、分析のレベルや統計処理の種類というような、いわばシステムの出口側に近い部分でのみ重要な要素ではなく、システムの入り口側ともいうべき、基幹データの取り込みと整形、そしてデータ分に対する品詞標識等の属性付与という段階にも当てはまる開発目標である。今後は、本研究で得られた知見をもとに幅広い英文データを対象としたコーパス研究成果を公表し、システムユーザーのニーズに応えるための柔軟性と汎用性をさらに追及していく必要がある。

(7) 総括: VPN を介した共同研究体制を構築すること自体が、本研究のような汎用的で柔軟な性格を持つ新しいコーパス解析システムの開発という目的に沿ったものと言える。コーパス研究には、コンピュータ可読式の言語データの機械処理という性格を当てはめれば 40 数年の歴史がある。その歴史を言語研究全般のそれと対比して浅いと見るか相応のスパンと見做すのかについては意見は分かれるが、総じて近年は、コンピュータやネットワーク通信の技術的な進展のペースに、コーパス研究そのものが対応できていないような様相を呈している。即ち、安易に入試可能な電子データを無作為にそして無計画に集積し、発達した統計処理ソフトウェアにその分析を委ねることによって、何らかの数量的な結果が容易く得られるようになってきてはいるものの、基幹となるコーパスそのものと、それが付随して有するべき属性に関する議論や知見が甚だしく欠落したままで、コーパスという言葉や、その分析に基づいた、言語教育分野がその主流となるが、応用研究が膨大に公表されている。

その意味において、本研究では、①コーパスという情報単位をどのように汎用的な形式で蓄積し幅広い研究や分析に供するべきか、②コーパス研究の応用分野の中でも、言語教育という極めて間口の広い領域に視座を据えて、システムのユーザーである研究者、教育者、教材作製者、学習者という多様なニーズを持った利用者層に対して、どのような工夫と仕組みをシステムに実装させなければならないのか、という問いに答えようとしたことになる。

単体のデータの集積としてコーパスを捉えてしまっては、それに対する統計的分析等も必然的に単体的でアドホックなも

のとなる。コーパスを統合的に管理する上位の属性を反映させるような情報を、RDBMS 中のテーブルに実現することによって、分析手法のみならず個々のコーパス構築と整備に関わる手法が汎用的な利用に供されることになり、個別のユーザーであっても、研究者や教育者や教材作製者や学習者それぞれが有機的な連携の一環を成すことになり、言語教育のみならず、コーパスを用いた自然言語処理と分析の発展に大きく寄与することになる。本研究はそのような遠大なビジョンに向けての意欲的な挑戦と位置付けることができよう。

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 Takeshi Okada、Yasunobu Sakamoto、A New RDBMS and Flexible POS Tagging for EFL Learners and Researchers: Designing a Corpus Analysis System Based on the Three-tier Model、東北大学高等教育開発推進センター紀要、査読有、第 5 号、2010、1-19
② 岡田毅、3 階層モデル準拠の新しいコーパス解析システムにおける英語スペルチェッカーの改良、東北大学国際文化研究科論集、査読有、第 17 号、2010、91-107
③ 岡田毅、コーパス研究と中学校英語教科書: 新しい動詞分類基準に向けて、東北大学国際文化研究科論集、査読有、第 15 号、2007、99-114
④ 岡田毅、eラーニングにおけるinteractionについて: SCPDから学べること、e-Learning教育研究、査読有、第 2 巻、2007、1-12

〔学会発表〕（計 7 件）

① 坂本泰伸、コーパスに基づく外国語研究と学習・教育支援システムの開発: 英単語に対する自動品詞タグ付与アプリケーション TAGAssignerの開発、平成 21 年度情報処理学会第 5 回東北支部大会、2010 年 2 月 12 日、仙台市
② Takeshi Okada、Yasunobu Sakamoto、A New DBMS and Flexible POS Tagging for EFL Learners and Researchers、Corpus Linguistics 2009 大会、2009 年 7 月 23 日、連合王国リバプール
③ Takeshi Okada、A Corpus-based Study of Sentence Patterns Peculiar to L2 Writing、ICLCE (International Conference on the Linguistics of Contemporary English)第 3 回世界大会、2009 年 7 月 15 日、連合王国ロンドン
④ Takeshi Okada、Yasunobu Sakamoto、The ESP Corpus POS Tagging Based on Users' Preference: Annotation and Statistical Analysis、ICAME (International Computer Archive of Modern and Medieval English)第 30 回世界大会、2009 年 5 月 29 日、連合王国ランカスター大学
⑤ 坂本泰伸、コーパスに基づく外国語研究と教育支援システムの開発: 再定義可能な品詞タグ付与に関する考察、平成 20 年度情報処理学会第 5 回東北支部大会、2009 年 2 月 13 日、仙台市
⑥ 坂本泰伸、コーパスに基づく外国語研究と教育支援システムの開発: 自動品詞タグ付与ソフトTAGASSの開発、平成 20 年度情報処理学会第 5 回東北支部大会、2009 年 2 月 13 日、仙台市
⑦ 坂本泰伸、コーパスに基づく外国語教育と教育支援システムの開発、平成 19 年度情報処理学会第 5 回東北支部大会、2008 年 2 月 14 日、仙台市

６．研究組織
(1)研究代表者
岡田 毅（OKADA TAKESHI）
東北大学・大学院国際文化研究科・教授
研究者番号： 30185441

(2)研究分担者
坂本 泰伸（SAKAMOTO YASUNOBU）
東北学院大学・教養学部・准教授
研究者番号： 60350328