ネットワークユーザーによる属性付与機能を実装した英語教材作成支援用データベースの構築

研究代表者　岡　田　　毅　東北大学大学院国際文化研究科教授
共同研究者　坂　本　泰　伸　東北学院大学教養学部准教授

1 序論

本成果報告書は、電気通信具急財団の研究助成を受けて平成 21 年 4 月～平成 22 年 3 月の期間に実施した標題の研究に関するものである。研究代表者の専門分野である英語コーパス研究・英語教育への応用と、共同研究スあの専門分野である汎用データベース研究・情報教育の両分野間の融合を図り、コーパス(大規模な英語データ集)解析システムの構築を、言語学的な視点と情報処理学的な視点から捉えなおしたものである。具体的には、両研究者の研究室間のシームレスなネットワーク接続環境構築からはじまって、ともすれば拡張性や汎用性に乏しくなりがちな文系の研究者や学習者や教材開発者が行うコーパス開発や解析に対して、より大きな拡張性と汎用性を伴った新しいシステムの方向性を模索した。その中で最も重要視されたのは、システムユーザーである英語学習者や教材開発者の要求に柔軟に呼応するために必要なデータベースの設計であり、これを通して、解析対象となる英語データに自らの要求に合致したさまざまな属性情報をユーザー自身が付加することが保証された。これによって、従来にはないコーパス構築・設計・解析の一連の処理に対する新しい視座を求めようとしたものである。

2 英語教材作成支援システムの開発: 技術的側面

2-1 システム開発と VPN 構築

本研究で開発する英語教材開発の支援システム(図 2.1)はコーパスが蓄積されているリレーショナルデータベースマネージメントシステム（RDBMS）を中心として、コーパスのデータを RDBMS 上に構築する部分と、RDBMS 上のデータに対して統計解析を行う 2 種類の部分から構成される。支援システムの解析部分には、自然言語学研究の分野からの要求を満たす統計的解析機能の実装が求められており、その解析結果は外国語学習者へ判りやすく提供することが求められている。この解説キットや応用キット等の補助的なツールの充実化には、一定量以上のデータが RDBMS に蓄積されていることが必要不可欠となる。しかしながら、RDBMS 中のデータ拡充は必要不可欠な部分であるが、やみくもにデータ量を増やすだけでは、我々の望む支援システムの実現は難しい。研究や教育に対して最も効果的な支援を提供するために必要となるデータを、厳選しながら RDBMS の拡充を進めていく必要がある。このように、支援システムの開発では、自然言語研究や教育からのフィードバックを柔軟に取り入れて開発の方針を適宜決定して行く必要がある。

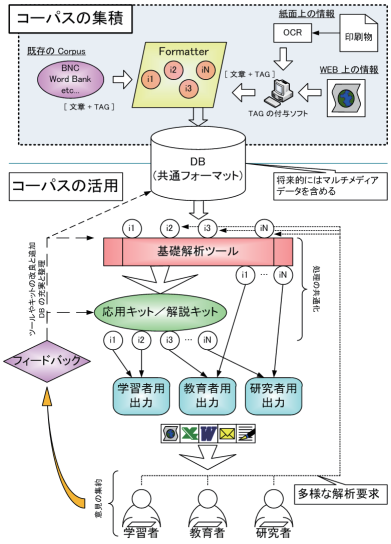

図 2.1

2-2　VPN 構築の意義

本研究組織は、システム開発の責任を担う東北学院大学の研究室と、RDBMS のデータ収集を進める東北大学の研究室の二つの研究室から構成されている。この二つの研究室に所属する開発グループは、開発時から柔軟に意見交換やシステム利用の感想を交換できるような環境が強く望まれている。支援システムの提供する機能は WWW を通じてユーザーへ配信されるが、その際に外部からのセキュリティに対する考慮や、システムを利用するユーザーの管理などを考慮する必要ある。その対処方法には、WEB システムの設定を中心としたユーザー管理や、システムに堅固なユーザー認証機能を実装する事が挙げられるが、前述した通り、統計解析を中心とした機能の開発を進めるフェーズでは、一般に公開できないようなシステムをネットワーク越しに互いに使用しなければならない状況が生じる。一般的に、ネットワーク通信を行う場合には TCP/IP のような通信プロトコルが利用される。通信に利用されるプロトコルは階層構造を持って構成されており、アプリケーションは下位層における複雑なデータ制御処理をあまり意識することなく、ネットワーク通信が可能となる。セキュアーな通信の代表例として HTTPS や SSH の利用が挙げられるが、これらのプロトコルはアプリケーション層で暗号化がおこなわれる。そのため、この安全なプロトコルを利用してアプリケーションを開発する際には、構築しなければならないレイヤーの位置が限定されてしまう。一方、IPSec を用いて Virtual Private Network （VPN）を構築する場合では、インターネットプロトコル（IP）層自身が暗号化されるの、開発者やネットワーク利用者は、その上位層に位置するプロトコルを既存のまま利用しても安全な通信経路を確保することができる。このため、柔軟なシステム開発が可能になる。この様な開発背景から、我々は IPSec を用いてシームレスかつ安全な VPN 環境を両研究室間に実現する。

2-3　VPN システム構成

教材作成支援システムの開発を行う両研究室には、既に他の研究用のネットワークが構成されている。本研究で設置する VPN は、構成員にとって利用しやすい環境となることが望ましい。そこで、本研究では構成員が VPN 用のネットワークと VPN 以外の既存のネットワークをあまり意識する事なく利用できる環境を提案した。また、東北大学の研究室（岡田研究室）に設置する VPN ルータに関しては、管理者が所属する東北学院大学の研究室（坂本研究室）から物理的に離れた位置に設置される。そこで、東北大学の究室内に設置するルータの管理についても検討を行った。今回、我々の提案したネットワーク構成を図 2.2 に示し、図中におけるルータの型番との対応表 2.1 に掲載する。

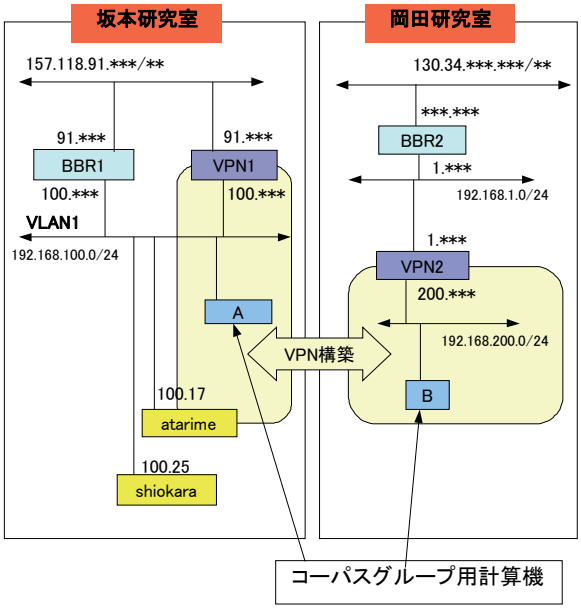

図 2.2

VPN 環境下において、東北学院大学の研究室内で利用する本研究用コンピュータは、ゲートウェイを VPN 用（VPN1）と VPN 以外の接続用（BBR1）にルーティングを分けて使用する。それによって、VPN 以外の接続に対しては 1 Gbps（公称値）の通信速度を確保する事ができ、東北大学の研究室とのネットワーク通信（VPN）

も併用することが可能である。また、VPN１には ACL を設定して、本研究用途以外のコンピュータが VPN を使用することに制限をかける。それによって、東北大学研究室内の計算機に対する不必要な通信や、逆に東北学院大学の研究室のコンピュータに対する不要な通信を防止する。また、東北大学の研究室には、上流にブロードバンドルータが設置され､そのファイアウォール機能を利用してセキュリティ対策もおこなっている。

今回の VPN ネットワークの構築には､YAMAHA 社製の VPN ルータ 2 台と COREGA 社製ブロードバンドルータ 1 台を利用した。東北大学と東北学院大学のローカルエリアネットワーク間の通信は IPSec により暗号化されるが､この暗号化と複合化を行う際にはオーバーヘッド時間が必要とされることが予想される｡そこで､我々の構築した VPN 環境がシステム開発に十分な通信性能を有するかどうかを調査した。調査では、東北学院大学側の研究室に FTP サーバ（計算機 A）を立ち上げ、東北大学側の計算機 B から FTP を用いてファイル転送の速度を IPSec 利用時と非利用時で測定した。FTP のプロトコルバージョンは FTP.0.17-16 である。また、この調査に先駆けて計算機間で IPSec による暗号化が行われているかは、パケット解析で確認済みである。この調査の結果は表 2.2 に記す。

表 2.1

	型版	型版	公称性能値
BBR1	COREGA	CG-BARPRO-X	929.9Mbps
BBR2	BUFFERLO	BHR-4RV	94MBps
VPN1	YAMAHA	RTX1100	200MBps
VPN2	YAMAHA	RT107e	200MBps

表 2.2

	Down Load	Up Load
IPSec	3.0 MB/sec	5.5 MB/sec
非 IPSec	4.2 MB/sec	5.5 MB/sec

実験の結果、我々のシステム基礎開発時に必要と考えられる通信速度（テキストエディタ等がストレスなく利用できる）を充分に確保していることを確認した。また、東北大学と東北学院大学の 2 地点間の IPSec 時のネットワーク速度は、非 IPSec 時の速度と比較して速度差が顕著に見られなかった事から、暗号化を行う際のオーバーヘッド時間がさほどネットワーク通信の性能に大きな影響を与えていないことを確認した。

3　コーパス解析システムの全体像

3-1　3 階層モデル(three-tier model)とコーパス解析システム

多階層構造(multi-tier architecture)の一種である 3 階層モデル(three-tier model)はソフトウェア工学の分野では広く普及している概念モデルである(Petrou, *et al.* (1999))。各階層間には高度にデザインされたインターフェイルが実装されている｡システムのユーザー側に最も近い階層はしばしばフロントエンド(プレゼンンテーション)階層と呼ばれ､中間に位置する階層はミッド(ロジックあるいはアプリケーション)階層と呼ばれ、ユーザー側からみれば最も遠い位置にある階層はバックエンド(データ)階層と呼ばれる。このモデルの最大の利点は、3 つの階層のうちのいずれでも、ユーザーの要求や技術の進歩に伴って、他の階層とはまったく別個に、独立した形で修正されたり大幅な更新を受けたりすることが可能なことである。図 3.1 で示すように、我々はこの 3 階層モデルの考え方を独自のコーパス解析システムに応用しようとするものである。フロントエンドにおいて Java や Web ベースで稼働する GUI(graphical user interface)は整備され、主にインターネット通信を介してユーザーやコーパスの研究者は GUI を通してシステムとやり取りをする形式になっている。この図からも明らかなように、ミッド階層は複数のモジュールから構成されており、RDBMS(relational database management system)の中で互いに相関関係を持ちながら機能するように設計されている｡そしてこのアーキテクチャがデータベースへのデータの入力形式や入力に関わるロジックと共に、結果的には、コーパス解析に必要とされるシームレスな解析ロジックをユーザーに提供することになるのである。バックエンド階層は RDBMS から成り、これを通して SQL によってコーパスデータが蓄積され更新され検索されることになる。データはコーパス解析プログラム群とは別に格納されているために、多様な利用要求を持つと想定されるシステムユーザーたちの間でいかなるデータであってもそれらを共有することが可能

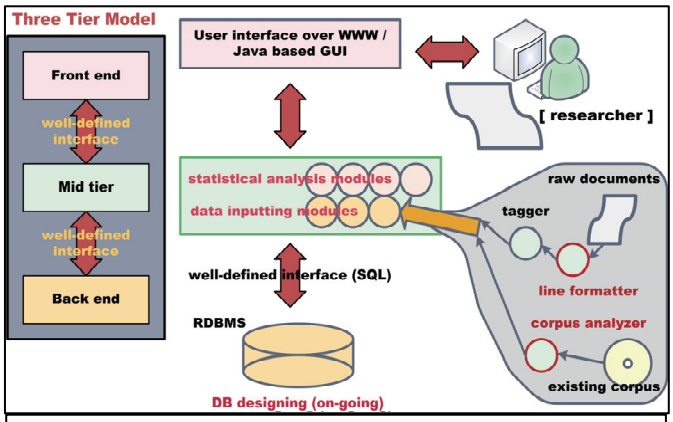

図 3.1　3 階層モデルとコーパス解析システム

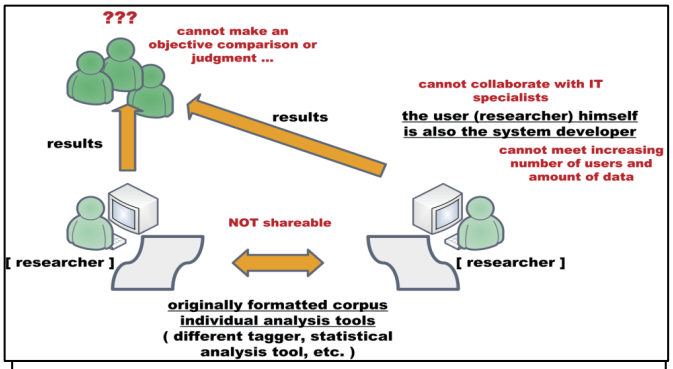

図 3.2　コーパス研究における低い拡張性

となる。

世界的に認知された研究機関の率先によって、コーパス研究に関わる大規模なコラボレーションが盛んになってきているとは言え、多くのコーパス研究者たち、とりわけ EFL(English as a foreign language)/ESL(English as a second language)の研究者たちの間では、独自の研究目的のために独自の小規模なコーパスを開発することが依然として主流である。本研究ではこのような個別のコーパスを独自に開発することの多い研究者たちにシステムユーザーとして焦点を当てる。図 3.2 に示されるように、個別の研究者たちには、それぞれが独自で他の研究者たちとの間での互換性の乏しい個別コーパスを構築するという傾向がある。

この種の低い互換性や統一性の欠如という性格が、独自に開発された個別のコーパス間で最も明白となるのは POS(品詞(Part-of-Speech))標識の付与のレベルである。Atwell (2008)が指摘するように、多くの異なった POS 標識のセットが存在しているからである。英国 Lancaster 大学で開発された CLAWS(Constituent Likelihood Automated Word-tagging System)に属するような POS 標識ファミリーであってさえも、POS 標識が示す品詞の「繊細さ・詳細さ」(delicacy)については多様性が存在する。このような表示形式における詳細さの多様性加えて、解析用のツール、とりわけ統計解析用のプログラム等は、目標とする分析自身に大きく依存するという性格を帯びている。言うならば、解析ツール間にも豊かな互換性は保証されてはいないのである。このことは、それぞれの研究者が同時にシステム開発者であることを意味している。従って、多様化し増加するコーパスシステム使用に伴う需要や要求やニーズといったものに個々のシステムが対応できないということを示している。この種のシステム開発のスタイルは本質的に共同開発者として多くのソフトウェアエンジニアの参画を困難にしているからである。小規模な個別コーパス解析システムが内包する低い拡張性と互換性の結果、分析結果を公表された側の他の研究者や論文の読者たちはその分析の信頼性や妥当性に対して公平な判断ができなくなることが多い。それらの公表された結果は、個別に構築された小規模コーパスに対する個別の手法に基づいた分析の産物だからである。コーパス解析システムに 3 階層モデルの概念を適応することによって、このような「個人開発」の小規模コーパスの解析結果に見られる低い拡張性や低い共同開発性や低い互換性というような短所を克服することをここでは目的としている。

3-2　開発の現状

ここでは、おもにバックエンド階層に相当するデータベースとデータ入力モジュールを含むミッド階層の開発を概観する。データベースのテーブルについて詳述する前に、ミッド階層が担う、データ入力と統計的解析という 2 つの主要な役割について言及する。データ入力は POS を含めた付加情報標識の付与によって実現され、それは Schmid (2008)で述べられているトークン化作業(tokenisation)と BNC (British National Corpus)のような既存大型コーパスの解析プロセスから始められる。

実際のコーパス分析で用いられる統計処理モジュールの開発はデータベース自体が十分な量の情報を含んでいるか否かに大きく影響を受けるために、我々は BNC のような既存コーパスの構造的マークアップを最初に解釈し、続いてコーパスのテキスト本体とそれに付加された属性情報を別個に抽出するプログラムを開発した。このプログラムはコーパスアナライザ(corpus analyser)として図 3.1 で示されている。このプログラムと並行して、トークン化プログラムであるラインフォーマッタ(line formatter)も開発した。このライン

フォーマッタに関しては後に詳述する。このプログラムとコーパスアナライザが共働して、次の開発段階で実現される POS 標識付与プログラム(POS tagger)が要求する豊富な量の入力データを生成することになる。つまり、本システムではデータベースが既存コーパスから豊富な量のデータを入手するのに加えて、ラインフォーマッティングプロセスを経た十分な量のデータを併せて入手することが可能となっているのである。プレーンな初期データに対して、ユーザーは独自のコーパス解析ニーズに合致した POS 標識や追加的な属性標識を付与することができる。例えば、ユーザーは CLAWS5 や CLAWS7 といった定評のある標識セット(tag set)を選択することもできるし、POS 標識変換器(POS tag convertor)と呼ぶプログラムを介して、既に付与されている POS 標識を他の標識セットに変換して付与し直すことも可能となる。これに加えて、POS 標識デザイナー(POS tag designer)と称するプログラムを用いることによってユーザー自らが望む POS 標識セットを選定し設計することができるようになる。このようにして、入力となったデータが既に POS 標識情報などを伴っている場合、その情報は再解釈され、必要に応じて異なった標識セットと置き換えるというプロセスが実現されるのである。結果的に、柔軟な POS 標識や属性を付与することが可能なデータが統一された形式を伴ってデータベースの中に増加し続けることが保証される。

多くのシステムユーザー、特に EFL/ESL 研究者や学習者たちは、特定の言語使用域やジャンルに特化して、個々のニーズを満たすために比較的に小規模なコーパスを構築する傾向がある。独自のコーパスを構築するために、そのようなユーザーたちは、印刷物や Web 文書や PDF 文書というような多岐にわたる「生の」データ資源からテキストを収集しなければならず、収集の後も後続の蓄積や分析プロセスに供するために、それらのデータを適切な形式にフォーマットし直さなければならない。例えば、紙に印字された英文のデータなどは通常 OCR(光学的キャラクタ認識)プロセスを介さねばならないし、Web 文書などには HTML や XML などのマークアップ言語によって付加的な属性が付与されているために、それらを慎重に扱い、不必要な情報は削除されなければならない。この両方のプロセスにおいて、Mikheev (2000), Mikheev(2002), Ratnaparkhi (1996), Reiley (1989)などが指摘するように、トークン化は常に付きまとう深刻な問題である。トークン化はソフトウェアのみでは完璧に実現することができないからである。

我々のラインフォーマッタは Java 言語によるアプリケーションとして開発されている。生の入力データが持つ高度な多様性に対処するために、実際のトークン化プロセスは一連のフィルタ群に分割されている。それぞれのフィルタは Java のクラスとして実現されているために、「ダイナミック・クラス・ローディング」(Dynamic Class Loading)と呼ばれる Java の機能によって、ユーザーは必要なタイプのフィルタ群を選定し、それらの適応順序までをも指定することが可能となる。このような柔軟なフィルタリングは形態素解析に立脚し、オブジェクト指向型言語の特性を十分に反映したものといえる。別の言い方をすれば、ラインフォーマッタはフィルタ・クラスの連続体として実現されているのである。それぞれのフィルタは個別に機能するので、処理対象となるインプットデータの性格によってシステムユーザーは、自分の分析の目的に沿ってフィルタを選定し、その適応順序をキューの中で指定することができるのである。

トークン化を可能にするこのような柔軟な形態素解析に加えて、このモジュールには新しい拡張性を実現したスペルチェック機能を実装することができる。これはデータインプットの対象となる英文が EFL/ESL 学習者のような非英語母語話者の産出によることを射程に入れているためである。フィルタの一種として Mitton and Okada (2007)で論じられているような、非英語母語話者向けに改良されたスペルチェッカがユーザーによって選択され指定されることを可能にしている。

3-3 RDBMS 中のテーブルの設計概説

支援システムが利用するコーパスを RDBMS 内に構築するには、さまざまな文書から目的とする単語や構文といったような、利用者が求める情報を柔軟に抽出可能となるようにテーブル設計を進める必要がある。例えば、研究者がコーパス利用する際には、母集団となる文書の集合の中から、必要とする単語やその連鎖を含む文を抽出してその出現回数を計数する操作が必須となる。対象となる文書の中には、段落や文、さらにその内部構成要素として単語と云ったものが存在する。これらの階層構造をもつ情報を、検索に際しても十分な時間性能を持つように、平面的な表形式のテーブルを利用して RDBMS の内部に構築することが必要となってくる。

文書を管理するテーブルの設計を行うにあたり、一つのテーブルを利用して全ての文書を蓄積する手法と、1 つの文書毎に対応するテーブルを準備して管理する 2 つの手法を提案した。解析等の文書に対する統計処

理ルーチンの開発の面から考えると、文書毎にテーブルを準備した方が 1 つのテーブルで全文書を管理するよりも開発効率が高く、文書を削除したり、一部の内容を書き換えたりする際も容易であると考えられる。そこで、この 2 つのデータベースのテーブル構成に関してベンチマークテストを行った。

ベンチマークテストを行った際に使用したテーブルの構成について概観すると、データベースには、英単語を蓄積した source テーブルと文書を管理するテーブル、文書の内容を表すテーブルが存在する。文書の内容を表すテーブルは 2 種類存在し、データベース中に存在する個々の文書からなる granular テーブル群と、データベース中の全ての文書の中身を蓄積する gathered テーブルである。これらのテーブルは、英単語に関連付けられた一意の key 値の連鎖を持ち、これによって英文書を表現している。データベースには BNC に含まれる文書数と概ね同数の 4000 文書を蓄積する。

このベンチマークは、表 3.1 で記される環境下で実行した。ベンチマークでは、4000 の文書に対して、英単語を保存している source テーブルから無作為に単語を選び出し、その単語に関連付けられた key を含む文書を抽出する処理を行い、この処理に要する時間を求めた。本研究で利用した RDBMS である PostgreSQL には、データの検索を高速化するために各テーブルに含まれるレコードのインデックスを作成する機能がある。ベンチマークでは、この機能のオン・オフに関しても調査を行った。この調査結果を表 3.2 に記す。データベースの検索を高速化するために、データベースに含まれるテーブルのインデックスを作成してベンチマークを行った場合も、インデックスを作成しない場合でも、Gathered テーブルを利用した方が高速化される事が明らかになった。

OS	Linux (Debian5.0.3 lenny)
CPU	Intel(R) Xeon(R) CPU E5450 @ 3.00GHz
CPU数	2
core数 / CPU	4
メモリ	24 Gbyte
DB	PostgreSQL 8.3
プログラミング言語	JDK 1.6

表 3.1

OS	Linux (Debian5.0.3 lenny)
CPU	Intel(R) Xeon(R) CPU E5450 @ 3.00GHz
CPU数	2
core数 / CPU	4
メモリ	24 Gbyte
DB	PostgreSQL 8.3
プログラミング言語	JDK 1.6

表 3.2

このテーブルに、BNC に含まれるデータを蓄積した。BNC に含まれる各文書の容量は平均で数キロバイトの大きさである。Gathered テーブルを利用した場合、DBMS に含まれる文書の数に比例して文書の蓄積に要する時間が長くなる傾向が見られ、最長で 20 分程度要することが明らかになった。一方、Glandular テーブル型を利用した場合は、この傾向は顕著に見られなかった。今後は、データ蓄積に関する調査を進め、データ蓄積と検索の間のトレードオフの取捨選択を進める必要がある。

3.4 単語リストテーブルと品詞標識リストテーブルの設計

このベンチマークを踏まえて、我々は次のようなテーブルの構成を提案した。ここでは便宜上、例文として *This system is useful for a wide range of researchers.* という英文を利用する。次の Table 1 は単語

リストのテーブルで語彙項目が連続番号を付与された形でランダムに列挙されている。Table 2 は品詞標識のリストでPOS key という名称でCLAWS5 タグセット中のそれぞれの標識が連続番号を伴って格納されている。ここで重要なことは、ユーザーの要求に沿ってこの品詞標識のリストは修正したり変更したりすることができるという点である。POS tag デザイナーと呼ぶプログラムのおかげで、品詞標識間の親子関係(包含関係)や相関関係を自由に定義し設計することが可能となる。

Word Key	word
1	the
2	of
3	and
4	a
5	in
6	to
7	it
8	is
9	to
10	was
11	I
12	for
13	that
---	---
23	not
24	this
25	's
---	---
196	about
197	system
198	local
---	---
486	century
487	range
488	European
---	---
914	numbers
915	wide
916	appropriate
・・・	・・・
1038	prices
1039	useful
1040	conference
---	---
3864	judges
3865	researchers
3866	equivalent
---	---

Table 1

POS Key	POS tag
1	AJ0
2	AJC
3	AJS
4	AT0
5	AV0
6	AVP
7	AVQ
---	---
12	DPS
13	DT0
14	DTQ
15	EX0
16	ITJ
17	NN0
18	NN1
19	NN2
20	NP0
---	---
27	POS
28	PRF
29	PRP
30	PUL
---	---
40	VBN
41	VBZ
42	VDB
43	VDD
44	VDG
45	VDI
46	VDN
47	VDZ
48	VHB
49	VHD
50	VHG
51	VHI
52	VHN
53	VHZ
---	---

Table 2

3.5 センテンステーブルの設計

例文中の各単語には単語キーと品詞キーと共に連続番号が付与されている。Table 3 が示すように、これによってコーパスデータ中の文が数値化して表現される。

Table 4 のセンテンス管理テーブルは行同士の関係を表現しており、具体的には連続番号 1 の文は単語番号 1 を与えられた単語から始まり、単語番号 11 の単語で終了する。そして次の文は 12 番の単語から開始され 21 番で終了するというように、である。単語の連続番号 1 は単語キーとして 24 を付与されており、これが具体的な単語 *this* にリンクしている。そしてこの単語は品詞キーとして 13 の(DT0)を有しており、2 番目の単語は単語キーとして 197 の *system* を持ち、品詞キーとして 18 番の NN1 にリンクしているという具合に、それぞれのテーブル間で有機的な関連付けがなされている。

Word sq No	Word Key	POS Key	---	---
1	24	13		
2	197	18		
3	8	41		
4	1039	1		
5	12	29		
6	4	4		
7	915	1		
8	487	18		
9	2	28		
10	3865	19		
11	.	.		
12	---	---		
13	---	---		
14	---	---		
15	---	---		
16	---	---		
17	---	---		

Table 3

Sentence sq No	Word seq No of Initial wrd	Final wrd
1	1	11
2	12	21
3	22	29
4	30	39
5	40	51
6	52	62
7	63	70
8	71	97
9	98	107
10	---	---
11	---	---

Table 4

3.6 文書管理テーブルと文書属性テーブルの設計

上述のように、単語と文の属性間の関連を表現するテーブルを作成することによって RDBMS はそれらを容易に処理することができるようになる。そして、システムはまた、それぞれの文書に与えられた属性を次のようなテーブルで表現できれば、効率的にそれらを扱うことが可能となる。

Doc No.	Sent seq No of Initial sent	Final sent
1	1	154
2	155	290
3	291	489
4	490	723
5	724	975
6	976	1201
7	1202	1322
8	1323	1766
9	1767	2003
10	2004	2183
---	---	---

Table 5

Doc No	Title	Author	Publisher	Date
1	The Price of Glory	Horne, Alistair	Penguin	1993
2	R&D Management	Bergen, S.A.	Blackwell	1990
3	Winning Karate Competition	Mitchell, David	A&C Black	1991
4	Media and Voters	Miller, William	Clarendon	1991
5	Britain's Defence Dilemma	Jackson, William	B T Batsford	1990
6	Kites	Moulton, Ron	Argus Books	1992
7	Aspects of Language Teach	Widdowson, H.G.	Oxford Univ Pre	1990
8	A Short History of Sussex	Lowerson, John	William Dawson	1980
9	Against Liberation	Leahy, Michael	Routledge	1991
10	Rival States	Henley, J.S.	Cambridge Univ	1992
---	---	---	---	---

Table 6

Table 6 で表にされて表現された文書からユーザーは、自らの分析の目標としての文書をコーパスセットとして指定し活用することができる。Table 4 の中の、Sentence sq No という名称のコラムにそれらの情報はリンクされており、これは 1 番目の文書中の 1 番目の文は 11 個の単語から構成され、それぞれの単語自身や品詞標識も併せて検知することができるために、Table 5 によって表現された文書に関わるあらゆる情報をシステムは利用することができるようになる。

3.7 語彙的関連テーブル

ユーザー自身が必要性を認める語彙間の関連を反映させるテーブルを作成するのを支援するプログラムによって、幅広い語彙項目間の相関を関連付けることができる。ユーザーが仮に EFL や ESL 学習者である場合を想定すると、コーパス解析システムを用いて例えば同意語や反意語、動詞の活用形等について学べる機会を提供されているということは、とりわけ初級学習者にとっては益するところが多く、またこの機能を最大限に利用した有効な教材作成が効率的に行われることを保証することにもなる。大切なことは、語彙項目間の関連性についての緻密さの度合いをユーザー自身が定義することができるという点であり、言語研究者の要求するような詳細で綿密な語彙項目相関関係などを必要としない外国語学習者の、直観的で簡便な相関情報をテーブルの中で表現されている階層関係を操作することによって自在に抽出することができる。下の

Table 7,8,9,10 はお互いにリンクし合っている語彙的関連テーブルの例である。Table 7 は先述のセンテンステーブルの詳細であるが、それぞれの単語間の関連を示す 2 つの追加的なフィールドを伴っている。第 2、第 3 コラムの中で( )に入った項目は、説明上の便宜であり、実際は全て数値で表現されているものである。さらに、Table 9 と 10 のそれぞれのフィールのラベルは固定された POS 標識の名称ではなく、ユーザーによって定義することが可能なものである。こうして、システムはテーブルの中から特定単語の語彙的な関連情報全てを取りだすことができるのである。

Word sq No	Word key	POS key	Rec No(A)	Rec No(B)
1	(the)	(AT0)	3	---
2	(data)	(NN0)	1	2
3	(which)	(DTQ)	4	2
4	(I)	(PNP)	5	3
5	(love)	(VVB)	2	3
---	---	---	---	---
68	(love)	(NN1)	1	2
---	---	---	---	---
158	(get)	(VVB)	2	1
---	---	---	---	---

Table 7

Lexical Tbl No	Category	No of items
1	Noun	2
2	Verb	7
3	Article	2
4	Rel Pro	5
5	ProNoun	8
6	---	---

Table 8

Lexical sq No	NN1	NN2
1	love	loves
2	datum	data
3	---	---

Table 9

Lexical sq No	VVB	V-pres	VVZ	V-part	VVD	V-pass	V-perf
1	get	get	gets	getting	got	got	got
2	go	go	goes	going	went	-	gone
3	love	love	loves	loving	loved	loved	loved
4	---	---	---	---	---	---	---

Table 10

これらのテーブル間の関連性は、例えば名詞としての *love* は 2 つのトークンを持ち、POS 標識はそれぞれ NN1 と NN2 であり、動詞としての *love* は、それぞれに対応する POS 標識を持つ合計 7 つの活用形で出現するということを表している。最も重要なことは、それぞれのテーブルがユーザー自身によって設計することが可能であるという点である。これはユーザーが単語同士の関連性をコントロールすることができ、POS 標識の詳細さの度合いをも規定できるということを示している。

4 結論

IT 技術をはじめとする関連する研究分野の目覚しい発展に伴って、従来にもまして大規模なコーパスが構築されるようになり、量的な側面はもとより質的にも十数年前には創造もつなかかったような情報を伴った

言語データが身近な存在となった。言語コーパスが重要視しなければならないのは、その代表性(representativeness)と斬新性(up-to-datedness)である。無限ともいえる言語使用の実態を有限の量のコーパスに反映させようとする際に、言語のどの部分をどれだけの分量、切り出してきたかが大きな問題となる。また、日々変化を続け、膨張を続ける自然言語のいつの時代をどのタイミングで切り取っているのかという問題もこれと平行して存在する。これらの問題を内包するコーパスを言語教育、とりわけ外国語教育の分野に応用しようとする考えは数十年前から存在し、現在では「コーパスで学ぶ○△語」などという教材や放送番組なども豊富にアクセス可能な存在となっている。

そのような現状を踏まえながら、現在のコーパス研究とその言語教育への応用研究が依然として抱えているいくつかの問題を浮き彫りにし、それに対する解決策を、情報科学の知見に裏付けられたコーパス研究の立場から明らかにしようとしたのが本研究であった。Web を中心としたリソースが極めて身近なものとなった以上、現在では個々の研究者や教材作成者や外国語学習者が、自らの手で相当量にのぼるデータを入手し、それをコーパス化して研究や学習の素材とするという環境が日常化してきている。しかし、この場合、個人個人が構築し、分析しようとするコーパスには、量の問題よりも、汎用性や拡張性の欠如というような、情報処理の観点からの欠陥がつきまとっている。また、独自に開発したコーパスに対する独自の分析の結果に基づいた成果を公表した場合には、それを受容した側での客観的な判断や批判や後続すべき研究の発展の可能性は十分に保証されていないことになる

インターネットを中心とした電子通信が世界規模で発展している今日にあって、このように個別・固有のデータやコーパスや分析が無数に存在していることは一種の皮肉な現象といえるかもしれない。本研究では、本論で述べたような 2 点に焦点を当てて、これまでに存在していなかった新しいコーパス構築システムをリレーショナルデータベースの概念を元に提案した。第 1 点は、データベースのテーブル間で実現される柔軟な関連性の確保であり、2 点目はユーザー自身によるコーパスセットの収集と構築、属性付与と属性そのものの定義を許す柔軟性である。これらによって、個人が収集し構築したコーパスは、他の研究者やユーザーと共有することが容易になり、統一された解析手法に基づいて、汎用性のある分析が可能となり、得られた結果も多くのユーザーによって共有することが可能となるのである。

【参考文献】

Atwell, E. (2008). “Development of tag sets for part-of-speech tagging”. In A. Lüdeling and Kytö, M. (ed.) *Corpus Linguistics: An International Handbook (Volume 1)*. Berlin: Walter de Gruyter, 501-527.

Granger, S. (1998). “The computer learner corpus: a versatile new source of data for SLA research”. In S. Granger (ed.) *Learner English on Computer*. London: Longman, 3-18.

Mitton, R. and T. Okada (2007). “The adaptation of an English spellchecker for Japanese writers”. In *Proceedings of the Symposium on Second Language Writing: Second Language Writing in the Pacific Rim,* 6, 15.

Mikheev, A. (2000). “Tagging sentence boundaries”, In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics,* 264-271.

Mikheev, A. (2002). “Periods, capitalized words, etc”. *Computational Linguistics,* 28, 289-318.

Okada, T. (2003). “Spelling errors made by Japanese EFL writers: a corpus-based approach”. paper presented at Colchester Second Language Acquisition Workshop 2003, University of Essex, Colchester, UK.

Okada, T. (2005). “Spelling errors made by Japanese EFL writers: with reference to errors occurring at the word-initial and the word-final position”. In V. Cook and Bassetti, B. (ed.) *Second Language Writing Systems*. Clevedon: Multilingual Matters, 164-183.

Okada, T. (2009). “An adaptation of English spellchecker in new corpus analysis system based on the three-tier model”, *Journal of the Graduate School of International Cultural Studies*, 17, 93-109.

Petrou, C., S. Hadjiefthymiades and D. Matrakos (1999). “An XML-based, 3-tier scheme for integrating heterogeneous information sources to the WWW”, In *Proceedings of Tenth International Workshop on Database and Expert Systems Applications: 1999*, 706-710.

Ratnaparkhi, A. (1996). “A maximum entropy model for part-of-speech tagging”, In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 133-142.

Reiley, M. D. (1989). “Some applications of tree-based modelling to speech and language indexing”, In *Proceedings of the DARPA Speech and Natural Language Workshop,* 339-352.

Ramirez, A. O. (2000) “Three-tier architecture”, *Linux Journal (July 1st, 2000)*
Available at: http://www.linuxjournal.com/article/3508 (accessed: 2 Sept 2009)
Schmid, H. (2008). “Tokenizing and part-of-speech tagging”, In A. Lüdeling and Kytö, M. (ed.) *Corpus Linguistics: An International Handbook (Volume 1)*. Berlin: Walter de Gruyter, 527-551.

〈発 表 資 料〉

題 名	掲載誌・学会名等	発表年月
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 第 5 号	平成 22 年 3 月
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 第 17 号	平成 22 年 3 月
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	平成 22 年 2 月
A New DBMS and Flexible POS Tagging for EFL Learners and Researchers	Corpus Linguistics 2009 大会	平成 22 年 7 月
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	平成 22 年 7 月
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	平成 22 年 5 月