社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－CE－77 (8)
2004／11／20

BNCを利用した英語教材作成とその提供Webサイトの開発

佐野 洋† 中村隆宏‡

†東京外国語大学 外国語学部 〒183-8534 東京都府中市朝日町 3-11-1
‡小学館コミュニケーション編集局電子辞書編集室 〒101-0051 東京都千代田区神田神保町 2-30 昭和ビル 4F
E-mail: †sano@tufs.ac.jp, ‡takahiro@shogakukan.co.jp

あらまし 筆者等は，中学校，高等学校で利用される英語教科書を調査し，英語教育課程で教授される英語文型を網羅的に調査した。その結果を基に，(株)小学館・マルチメディア局との共同研究により，BNC(British National Corpus：一億語の英文コーパス)から文型を基に英文用例を抽出するための検索式を作成した。検索式を用いて英文用例を抽出した。英文用例はXMLデータベース化し，XMLデータベースからHTMLデータを自動生成して，インターネットを通じて英語教育素材を提供するWebサイトを構築した。本サイトは，小学館コーパスネットワーク(SCN)のサービスとして提供される。

キーワード e-Learning，英語教育，語学教育，教育コンテンツ

Using the BNC to Create and Develop English Educational Materials and a Website

SANO, Hiroshi† NAKAMURA, Takahiro‡

†Faculty of Foreign Studies, Tokyo University of Foreign Studies 3-11-1 Asahi-cho, Fuchu-si, Tokyo, 183-8534 Japan
‡Electronic Dictionary Development Department Shogakukan Inc.
Showa Bldg. 2-30 Kandajinbo-cho, Chiyoda-ku, Tokyo, 101-0051 Japan
E-mail: †sano@tufs.ac.jp, ‡takahiro@shogakukan.co.jp

Abstract We have released a website that allows users to download sentence patterns for English educational purposes. These sentences were extracted from the BNC, and collected by writing search formulas using the LTB. Prior to the creation of the formulas, we surveyed major English textbooks used widely at Japanese educational institutes, for the purpose of covering the most studied English sentence patterns. The data is stored in an XML database and is available to users through the Internet. This project was done in collaboration with Shogakkan and Sano Laboratory at TUFS. The site is provided as a part of SCN services.

Keyword e-Learning, English Education, Language Education, Educational contents

1. まえがき

1.1. 研究目的

本稿の研究目的は，(1) 言語運用データに基づく英語教育用教材の効率的な作成と，(2) 作成した英語教材をネットワークを通じて提供する素材ウェブサイトの開発にある。

実用的な言語運用能力の育成には，4つの技能・能力(「読む」，「書く」，「聞く」，「話す」)を向上させなければならない。我々は，その4つの技能全てに関係する「文法」能力の効果的な育成を支援する教授法と教育教材の開発を目指している。筆者等は，現在，中学校・高等学校で利用される全種類の英語教科書を調査し，英語教育課程で学習されている英語文型を網羅的に調査した。その結果を基に，(株)小学館マルチメディア局との共同研究により，BNC(British National Corpus)から文型検索を通して教育利用のための英文用例を抽出した。抽出した英文用例に文法項目情報等を付加しXMLデータベース化した。XMLデータベースから，インターネットを通じて英語教育素材を提供するウェブサイトを構築した。本サイトは，小学館コーパスネットワーク(SCN)のサービスとして提供される。

1.2. 背景

日本における英語教育の改善・質的向上は，教育課程における問題に留まらず，社会的な要請も強い政策課題でもある。例えば，平成 16 年度に募集された「現代的教育ニーズ取組支援プログラム」(文部科学省)のテーマには、『仕事で英語が使える日本人の育成』が挙がっている。

また，ビジネス分野における業務等の国際展開と経済活動のボーダレス化に伴い，こうした状況下，専門的な分野で活躍する英語力を持った人材が求められている。専門分野の英語力向上を目指した学習は，英語教育分野でも必要性が認知されており，企業内教育では ESP 適合の英語教育教材の需要が高まり，効果的な学習法の確立が望まれている。

それに対し，ゆとり学習のもとに正規に割り当てられる英語の時間数の減少と，それに伴う教科書の教育項目の削減もまた，一方の現実である。この影響が小さい訳がなく，英語教育の改善・質的向上には，教育効果を向上する仕組みが必要である。

筆者等は，減じた学習内容を直裁に遡増するのではなく，投入学習負荷に対して，成果の効率向上を目指している。現実世界で使われる言語形式を教材に用いることが成果の効率向上につながるものと考える。

1.3. アプローチ

対話行為をある種の問題解決過程と考えると，発話者能力の育成は，発話状況における発話タスクの遂行に十分な「どのような」種類の知識(言語知識)を「どのように」使うべきか(言語運用知識)という計画遂行の能力の取得に帰することができる。従来の教材は、「どのような」種類の知識の学習と習得に力点が置かれ，したがって言語知識の理解と習得のための訓練が中心の教材内容であった。教育項目は言語学的な抽象化が過ぎていたのである。

発話状況毎に，発話戦略(解決方略)の中で使われる表現は，現実の状況で利用された(言語知識を含む)言語形式が対応する。言語運用知識は，これらの言語形式を訓練することで習得できるだろう。すなわち教材内容が言語知識だけを含むのではなく，同時に言語運用知識も含むようにすればよい。現状の学習項目規模に合わせ，学習負荷を効率化するほうが，学習者への負担は軽い。

上記の教材は，特に ESP 教育教育には適している。これらの教材の特徴は，専門分野の語彙と文型が意識されていること，リーディングとライティング能力の育成・向上に焦点あること，学習者の言語運用目的と到達目標に応じた内容を持つこと，効率的な学習が可能であることなどである。

具体的に，我々は，教科書に記載の用例だけでは不足する思われる演習用例文を，教科書の学習項目を特徴パターンとして，言語運用データから特徴パターンを含む用例を自動抽出する仕組みを研究している。

学習負荷の効率化に加えて，この取り組みの背景には，(1) 教材作成が労働集約的な作業であって，短時間に多様で且つ大量の教材が作成できないこと，(2) 作成された教材内容の品質は作成者個人の能力に大きく依存してしまうこと，(3) 作成者の多くは英語母語話者ではないので，英文表現に多様性が乏しいこと，など演習用例文の品質が定性的に保証できない問題がある。こうした問題の解決も視野に入れている。

1.4. Web サイトの位置づけ

筆者等はこれまで，多様な学習要求に適合する語学教育方法論(N-Cube)を研究してきた[1,2]。N-Cube は，ESP(English for Specific Purposes)適合の教材作成の効率化とその教授法確立を目指した語学教育支援枠組みである。この枠組みは，(1) 教育メソッドとして，認知力と母語の運用能力を活用することを特徴にした Cognition-Based Rule-Driven/Data-Driven (RD/DD) Interactive Learning 方法論の研究と，(2) 学習支援として，言語運用データと自然言語処理技術を使った効率的な教育教材作成を目指すものである。

図 1 には研究目的に対応する全体的な作業工程を示している。本稿がカバーする事項は，図中の(3), (4), (5), (6), (7)である。(4)の LTB(Language Tool Box)は(株)小学館が開発したコーパス利用のためのワークベンチである。

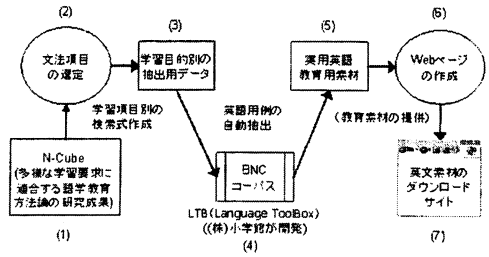

図 1. 全体構成図

我々は，教科書の学習項目を特徴パターンとして(3)，LTB を利用し，BNC から特徴パターンを含む用例を抽出した(4),(5)。抽出した英文用例に文法項目情報等を付加し XML データベース化した上で，インターネットを通じて英語教育素材を提供するウェブサイトを構築した(7)。本サイトは，小学館コーパスネットワーク(SCN)のサービスとして提供される。

以下，2 章では，小学館コーパスネットワークと LTB について説明する。3 章では，教科書の学習項目に対応する検索式の作成と，LTB を使った BNC からの用例抽出について説明する。4 章は，教育素材提供のための Web サイトの構築について，作業効率の向上を含めて説明する。

2. コーパス抽出システム

2.1. 小学館コーパスネットワーク

小学館コーパスネットワーク(SCN)は，コーパスを広く言語研究や語学教育に役立てることを目的に，(株)小学館が 2003 年に設立した商用サイトである。現在，コーパス検索がサービスの中心である。設立の背景には，コーパスの活用経験の蓄積や利用技術の進展を通じた討究によって，ひいては，辞書・辞典や語学教材の品質向上に貢献したいとの思いがある。

サイト設立当初，コーパスとして BNC を，2004 年 4 月には WordBanksOnline をリリースした。今後 SCN では，学習者コーパス(JEFL)，科学技術コーパス(PERC)，American National Corpus(ANC)のリリースを予定している。なお，提供が予定されているいずれのコーパスでも，利用方法が変わらないように，タグ付けには同一ソフトウェアを使い，検索インタフェースも共通化した。

小学館では，SCN の公開に先立ち，90 年代の後半から，コーパスに基づく辞典編集の手法を調査，研究してきた。小学館内部での電子編集のインフラを構築する目的で，コーパス検索のシステム開発も行ってきた。システム開発の中心に小学館 LTB(Language Tool Box)が位置する。SCN は，小学館社内における研究開発の成果を，利用者に使いやすい形態で公開したものである。なお，次節で紹介する LTB は原則的に非公開であるが，研究成果の共同利用を前提とすることで利用が可能である。

2.2. LTB の概要

小学館 LTB は，コーパス検索エンジンをコマンドインタプリターでラッピングしたサーバー/クライアント・システムである。検索コマンド(fcql)は，コーパス検索言語(CQL：Corpus Query Language)から構成される。検索コマンドだけでなく，コーパス検索で有用な種々のコマンドが提供されている[5,6]。

- kwic：検索結果を KWIC 形式に整形するためのコマンド
- cluster,clusterC：検索結果からクエリーの特定の文法カテゴリーと長さを指定することで出現語彙を集計するコマンド
- colloc,collig：検索結果からクエリーの任意の文法カテゴリーと相互位置を指定して，共起の統計情報を計算するコマンド
- cql_and,cql_or，cql_diff,cql_symdiff：複数の検索結果間の集合演算を行うコマンド
- random：検索結果からランダムに用例をサンプリングするコマンド

検索結果は，"最詳細 KWIC 形式"と呼ばれる書式付データとして出力される。最詳細 KWIC 形式は，クエリーの項の位置番号，ノード語の特定情報，テキストの形態・構文情報(文と語の属性群)を全て備えたデータ形式である。このデータ形式から，例えば，クエリーの特定の文法カテゴリーに出現する語の集計を後処理で行うことができる。なお，一行目には，クエリー情報がヘッダーとして追加されている。

社外利用の場合の小学館 LTB は，機能別にブラウザー上を複数のフレームで切ったグラフィックインタフェースが提供されている。ダイアログボックスの利用によって，コマンドに付帯するオプションやスイッチ指定の煩雑さが軽減する。コマンドは CGI 経由でサーバーに送られ，処理結果は，表示エリア(標準出力)に表示される。図 2は，LTB のスクリーンショットである。

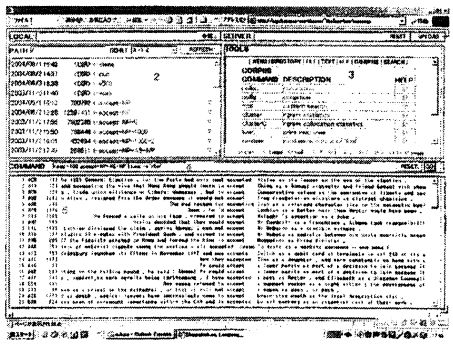

図 2. LTB スクリーンショット

(1) クライアントからサーバーへのファイルのアップロード
(2) サーバー上のファイル管理。ディレクトリーの表示とディレクトリー間の移動もできる
(3) コマンドに付帯するオプションやスイッチ指定を行うダイアログボックスの呼び出し。コマンドは機能毎に分類整理されている
(4) コマンドインタプリター用のインタフェース(コマンドライン)
(5) 結果の表示エリア(標準出力，標準エラー出力)

2.3. CQL とその機能

BNC の付加情報は，以下の 2 つに区分される。

- ヘッダー部の分類情報，文の属性
- 語の属性

ヘッダー部の分類情報は，サブコーパス情報を使った検索のために利用する。文の属性は，話し言葉か書き言葉なのかの特定とその属性を示す。語の属性は，タガー(CLAWS)によってタギングされたもので，品詞コードとレマ(辞書の見出し語の形態)[1]が付与されている。

小学館では，上記の 3 つの付加情報のうち，ヘッダー部の分類情報と，語の属性を任意に組合せたクエリーを書くことができるコーパス問合せ言語(CQL)を設計した。

語連鎖に対する検索要求は，語の並び(シンタックス)と属性の並び(パラダイム)の表に対する任意のパターン照合と考えればよい。例えば「"give up + Gerund"」の用例を，give の語変化パラダイムを無視して調べたい場合には「レマ="give"，レマ(活用がないので表層でもよい)="up"，品詞コード="VVG"(VVG は BNC における -ing 形を指す品詞コード)」で示されるパターンを記述できればよい。

連鎖する語数が正確に分からない場合，例えば，"give up + 名詞句"の用例を検索する場合には，名詞句の構造を近似的に「修飾語の有無＋名詞」で表現できればよい。“修飾語の有無”は，いわゆるワイルドカードで指定することになる。

このように，CQL は二次元配列された語と属性に対するパターンマッチを行う簡易言語として仕様化することができる。

2.4. DB の実装方法

小学館では，SGML の全文検索エンジンとしてすでに実績があった OpenText(ver.5)を視野に入れ検索機能とパフォーマンスの検討を行ってきた[7]。

ヘッダー部を検索キーにして，本文中のキーワードを重ねて条件にする場合，例えば「話し言葉で，"give up"の用例を探す」場合は，OpenText の PAT コマンドの region 指定(タグ名を指定してサブインデックスファイルを作成しておくと，そのタグ内だけを検索の範囲に特定することができるオプション)を使って直接，クエリーを記述できる。

ヘッダーのような単純なレコード形式のデータへのアクセスは問題がない。しかし，品詞，レマ，ワイルドカードを用いて，ある程度の長さの文のパターンを指定する場合，そのままでは PAT コマンドを上手く利用できない。すなわち，PAT では，ワイルドカードに相当する近傍検索は，その長さをバイト単位でしか指定できないためで，文パターンの任意の場所に多数のワールドカードが指定されるような複雑なクエリーでは，偽の照合も大量に成功してしまう。

そこで，ワイルドカードの解釈を，照合数を指定する仕様とした。同時に，本文部分のデータは，トークンを 4 バイトの固定長 ID に置換し，インデックスファイルを作成した。コーパスは更新されないから，置換表の最適化も行うことができる。

固定長 ID 化によって，データサイズも圧縮された。冗長な書式指定部を全て削除できること，英単語の平均語長は 4 バイト以上であることなどから，BNC コーパスのファイルサイズ(約 1.3GB)を 30%程度までに縮小した。メモリーコストの安い PC-UNIX 上では，インデックスファイルやソースの全てをメモリー上に展開できる。複雑なパターンサーチもファイルアクセスなしで実行でき，レスポンスの大幅改善が可能になった。

3. 例文自動抽出の方法

3.1. 文法項目の選定

教科書に関する市場調査(売り上げ高)を基準に選択した 31 種の英語の教科書(中学校英語教科書：6 種，高校英語教科書：英語 I ：8 種，英語 II ：8 種，ライティング 9 種)，及び日本の英語教育で広く使われている参考書 4 種を対象に文法項目を調査した。次の 2 点の方針，(1) 主要教科書に共通して現れる項目であること，(2) そうでない場合，(1)の条件を緩和し，各教科書を幅広くカバーすると同時に，言語運用規則として生産性の高い項目であることに基づいて 144 の文法項目を整理した。この 144 は，"I am+名詞"といった単純な文法項目から，"SVC"，"強調構文"，そして"仮定法未来"などの難しい文法項目まで含む。

144 の文法項目は肯定文を基本とする。そして各文法項目に対し，否定文や疑問文など 14 の下位項目を設けた。144×14(2016)項目から，英語構文として存在しない 634 のパターンを除き，総計で 1,382 文型を整理した。

3.2. 検索式の作成と用例抽出

1,395 の文型に対応する CQL 式を作成した。例えば「how を使った感嘆文」であれば次のような検索式になる。

```
^{W="how"} {P="AJ0|AV0"} [0,10] {L="!"}$
```

この式は，『"How"という単語で文が始まり，形容詞

[1] 一般に配布されている CDROM 版 BNC のデータにはレマの情報は付加されていない。CLAWS を使って付加した。

もしくは副詞が続き，0個以上10個以下の単語を間に挟んで"！"で終了する文を検索』することを意味する。上記に示すような検索式を使い，1382文型にすべてについて，英文用例を抽出した。

3.3. 用例評価と検索式の精緻化

教育教材の品質を確保するために検索式の精度向上を行った。

LTBで利用するBNCは，形態素解析結果までの情報を参照することができる。CQLを利用して，特定できる言語的な情報は，一つの語について，表層形，基底形(辞書形)，そして品詞分類の3つである。例えば，品詞分類では，自動詞と他動詞の区別がない(BNCの仕様)。そのため，例えば，SVO(O=that～)『目的語にthat節を取る他動詞構文』文型は，Vに対する他動詞品詞の指定ができないにも関わらず，Oの表層形の指定が可能なために，CQL式で検索すると，93%の精度で用例を抽出することができる。それに対して，SVOC構文『目的語に名詞をとる構文』文型は，Vの品詞指定ができず，Oが名詞であることを頼りにCQL式で検索するために，50%の精度でしか用例を抽出できない。

構文の抽出は，本来は構文解析結果を用いて行うべきであるが，1億語という大規模コーパスが利用できる利点があること，形態素解析の精度に比べ，構文解析結果の精度は低く，抽出精度が向上しても解析精度が高くなければ結果としての抽出精度は上がらないことなどから我々は，解析精度の高い形態素解析結果を基にして，(1) 検索式を作成し，用例を抽出する，(2) 抽出用例を評価し，検索誤りが生じている語連鎖を特定する，(3) 評価結果をもとに，誤り部分の語連鎖だけを検索する減算式を作成する，(4) 減算式を含めて再抽出を行って誤用例を正用例から除いた用例を得た。抽出用例の精度向上を示す概念図を図2に示す。赤い部分が，減算後の用例抽出結果になる。

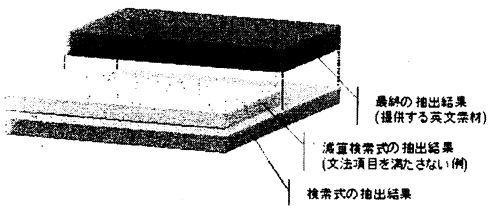

図 3. 用例評価と精度向上の概念図

4. XMLによる教材データの管理

4.1. 検索式による抽出およびHTML化

複数の作業者で，同時平行的に且つ，共同で検索式を開発する。作業者間で情報共有する仕組みが必要となるため，データ管理にXMLを使用した。

図3に用例抽出の作業工程の模式図を示す。図に示すように複数の作業者が検索式を分担して記述し，検索を行う用例を抽出する。個々の作業者が作成した検索式を使い，直接LTBを使って用例検索するのではなく，各作業者はXMLファイル上で検索式を編集したり更新したりする。XMLファイルからスクリプトを用いて，LTBで用いる具体的な検索式を生成し，バッチ的に実行することで用例を抽出した。データ管理にXMLを使うことで全員が，常に最新の情報を知ることができ，知識の散逸の防止・品質の統一を図ることができる。

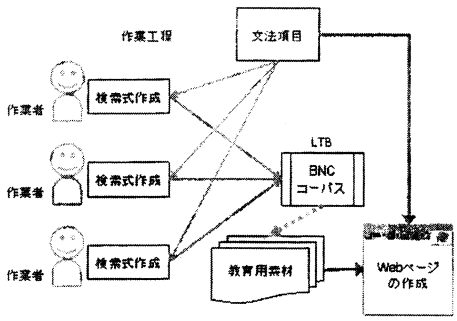

図 4. 作業の工程の模式図

同時に，検索式だけでなく，文法項目の解説などもタグを利用して管理が可能である。また，XSLTテンプレートに適用することで，自動的にHTMLを生成でき，視覚的な確認も容易になるなど利点がある。データ修正時にはXMLデータを書き換えれば，HTMLも連動して書き換わる(図4を参照)。

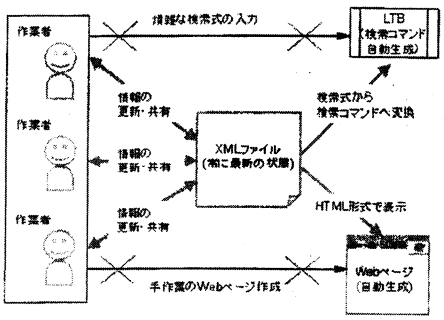

図 5. 教材データ管理

4.2. 英語教材提供のための Web デザイン

英語教育素材を提供する Web サイトの利用者として，筆者等は現在，中学生，高校生や大学生に英語を教授する立場にある人を想定している。文型に対応する用例データは，恐らく利用者の要求として，補助資料やテストなどを作成する際の英文参照や，作為例文のためのサンプル利用が考えられる。

実際の教科書では，文法項目名が目次として強調されていることが多い。利用者がキーワードとして探しやすいのは，文法項目名であると考え，文法項目名(展開項目を含む)からナビゲートし，文型説明が確認でき，そして英文用例がダウンロードできるようなデザインを行った(図 6 を参照)。

文法項目毎に，その項目の番号と項目名を表記する。同じ項目名でも教科書によって多少内容が異なることもあるので，項目名の定義を明確化するために，各項目について各種教科書で使われている用語を用いた文法説明データを付加した。その上で，その文法項目に対応した抽出用例文をダウンロードできるようにする。

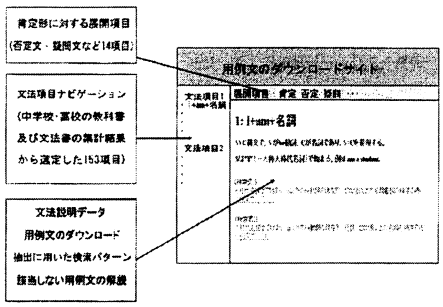

図 6. Web デザイン

4.3. 構築した Web サイト

図 7 に構築した Web サイトを示す。

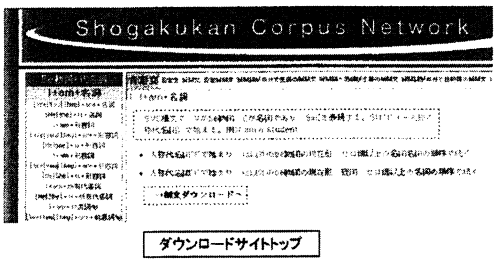

図 7. Web サイトの様子

画面右は，14 の下位項目をタブによって選択する部分と，提供する素材についての情報を説明する部分である。文型情報などに加えて，提供する用例文を，利用者が実際に教材として利用する際の参考データとして，用例の抽出精度評価の結果を示すようにした。評価内容は，抽出 100 文中で文法項目に該当しない例文についての説明，誤用例についての定性的な分析結果の記述からなる。参考として，文を抽出するために用いた検索パターンを示した。

5. 今後の課題

ダウンロードサイトの公開を行い，利用者評価を実施する。必要文法項目とその数，1 文法項目あたりの必要用例数，文長や語彙レベルの制御等の評価が必要だろう。利用者の要求はサイト改善に反映される。

謝辞

本研究は以下の助成を受けた。

(1) 平成 14-16 年度文部科学省科学研究費 (基盤研究(B)(2))「全電子化検定済み教科書データの解析と大規模日本語コーパスの構築」 (研究代表者：佐野洋)

(2) 平成 15 年度 (株) 小学館・マルチメディア局委託研究

文　献

[1] 佐野洋：「ESP 適合の教材コンテンツを実現する語学教育支援システム」,『最新外国語 CALL の研究と実践』, コンピュータ利用教育協議会(CIEC)・外国語教育研究部会(34～44,10 頁),2003 年 3 月.

[2] 佐野洋，猪野真理枝，宇野陽一郎：「多様性適合の学習環境を実現する語学教育支援システム」，情報処理学会，情報学シンポジウム講演論文集(55～62,8 頁),2002 年 1 月.

[3] 新井 雅之，渡辺 亜美，佐野 洋：「言語運用に基づく英語教育教材とその提供 Web サイト開発」，教育システム情報学会第 29 回全国大会講演論文集，pp.257-258，教育システム情報学会,2004 年 8 月.

[4] 岩倉隆幸，新井雅之，佐野洋：「言語運用データを使った英語教育教材の作成」，FIT2004 第 3 回情報科学技術フォーラム,2004 年 9 月.

[5] Nakamura, T. and Tono, Y. (2003) Lexical profiling using the Shogakukan Language Toolbox. In Murata, Yamada & Tono (eds.) ASIALEX 2003 Proceedings. Dictionaries and Language Learning: How can Dictionaries Help Human & Machine Learning?, pp. 170-176.

[6] Nakamura, T., Tateno, J. and Tono, Y. (2004) Introducing the Shogakukan Corpus Query System and the Shogakukan Language Toolbox. Williams, G. and Vessier, S. (eds) EURALEX 2004 Proceedings . The Eleventh EURALEX International Congress, July 6-10, 2004, Lorient, France, pp. 147-152.

[7] 中村隆宏 相澤弘 渡辺亮嗣 (2004)「自然言語文の検索方法および検索装置」特許出願 特願 2004-047377