言の場（ことのば）

内山 将夫[1]　　三浦 真磁[2]　　谷村 緑[1]　　井佐原 均[1]

1 情報通信研究機構　　2 大阪電通大

1　はじめに

コーパスが外国語学習に役立つことは良く知られたことであるが，それを実際に授業で利用しようとする場合には，克服すべきバリアがある．そのバリアのなかで，我々が英語学習に対してコーパスを利用したときには，次の2つが顕著であった．1つは，コーパス自体のバリアであり，1つはソフトウェアのバリアである．

コーパス自体のバリアとは，授業で利用できるようなコーパス自体を探すことが困難である，ということである．また，ソフトウェアのバリアとは，たとえコーパスがあっても，それを利用するためのソフトウェアがない限りは，利用ができないということである．

これら2つのバリアを克服し，英語学習を対象とした研究成果を，直接英語学習者まで届けるために,「言の場（ことのば）」というウェブサイトを開設した[1]．

本稿では，このサイトに載せている，あるいは，載せる予定のコンテンツについて述べる．なお，言の場自体は，英語学習に限らず，言葉に関連する，研究・学び・実用・表現・教育の場となることを目標としており，そのために，誰もが自由に編集できる Wiki となっている．

2　単言語コーパスの英語学習への利用

コーパス自体のバリアとは，授業で利用できるようなコーパスを探すこと自体が困難である，ということである．我々は，これまでに，語彙学習をサポートするために，学習対象語彙をコンパクトに網羅するような英語記事の集合を用意し，それを授業の補助教材として利用した [3]．この教材の目的は，学習対象語彙を多く含む記事を学習者に提示することにより，文脈を通して，単語を学習することを促すことであった．

そのときの学習語彙としては，中條ら [5][6] により作成された TOEIC[2] 学習用語彙 640 項目[3] を用いた．この語彙は，3 レベルからなり，レベル 1,2,3 が，それぞれ，中学校，高等学校 2 年，高等学校 3 年の英語教科書をマスターした学習者を対象としており，全レベルをマスターすることにより，TOEIC で 700〜800 点程度の得点が期待できるものである．

利用したコーパスは，*The Daily Yomiuri* の 1989 年から 2001 年までの 300 単語以下の約 25000 記事である．そこから学習用語彙をカバーするようなコンパクトな 116 記事を得た．その記事の例を図 1に示す．図 1の記事の全文では，異なり語数では 43 語，延べ語数では 61 語が学習対象の語彙と共通している．それら共通単語については，初出のものを太字，それ以外を斜体で示す．

このように，学習対象語彙が密に存在する記事集合を読むことにより，単語の意味を文脈とともに学習できる．また，異なる記事に何回も使われる単語を重点的に学習できる．

この教材の有効性は，本稿執筆時点では，調査中であるが，アンケートによると，学習意欲が元々高い学生にとっては，この教材は，学びがいがあるようであった．特に,「TOEIC の得点が 150 点以上あがり，700 点を越えた」という感想のように，主観的な評価において，高

[1] http://kotonoba.net/~snj/cgi-bin/wiki/wiki.cgi?page=FrontPage 本稿執筆の時点ではドメイン名は手続き中なので，kotonoba.net が利用できないときには，61.115.230.87 にアクセスのこと．

[2] Test of English for International Communication (http://www.toeic.or.jp/toeic/index.html)

[3] http://www5d.biglobe.ne.jp/~chujo/

Streamlining to cost NTT over 1.4 tril. yen

NTT Corp's restructuring plan, which aims to **transfer** 110,000 workers to subsidiaries, will **cost** the telecom giant a hefty 1.4 trillion yen to 1.5 trillion yen, The Yomiuri Shimbun learned Thursday.

The plan is **expected** to be so **expensive** because of ballooning **retirement** and other **compensation allowances** that will be paid to about 55,000 workers.

NTT will earmark lump-sum **expenses** in its **fiscal** 2001 **account** settlement ending in March to make up for the *costs* of the large-scale streamlining plan scheduled to be **implemented** in spring.

The nation's largest **telecommunications company**, which originally **forecast** after-tax **profits** of 3 billion yen for the **current** *fiscal* year, is **predicting** a loss of hundreds of billions of yen.

Under the restructuring plan, NTT will *transfer* a **total** of 110,000 of its 210,000 workers, mostly from its two **regional** phone **operators**–NTT East Corp. and NTT West Corp.–to other group *companies* to be set up. Among those *transferred*, 55,000 workers aged 51 and above will be **retired** and rehired at **salaries** as much as 30 percent lower than those they are currently **receiving**.

図 1: 記事の一部

い評価を与えた学習者もいた．

このことから，学習用語彙をカバーするコンパクトな記事集合は，教材として見込みがあると考えるが，我々は，それを更に確認することを目標とした．そして，そうするために，文献 [3] の手法により作成された教材を，我々以外の不特定多数の人たちにも使ってもらい，それにより評価することを考えた．

しかし，そうするためには，コーパス自体のバリアを克服する必要がある．文献 [3] においては，*The Daily Yomiuri*をコーパスとして利用したが，これを不特定多数の人に利用してもらうことはできない．そのため，(バリア 1) 編集可能でかつ再配布可能なコーパスを用意する必要がある．

また，別の問題として，コンテンツ自体の面白さという問題もある．語彙を学習するだけなら，単語の意味や用法は，10 年程度では変化しないため，10 年前の新聞記事でも，語彙の学習には問題がないはずであるが，実際上，学習者の学習意欲を喚起するには，(バリア 2) 学習者の興味にあったコンテンツを利用する必要がある．

更に，理想的には，学習者の英語習熟度レベルに応じた学習が必要である．そうするためには，学習者の現時点でのレベル，および，到達目標のレベルを測定し，そのギャップを埋めるようなコーパスが必要である．それをするのは，困難であるが，そうするための前提条件としては，(バリア 3) 様々な難易度レベルのコーパスが必要である．このバリアについては，当面，解決するための方法はわかっていないが，その解決には，なるべく規模が大きく，かつ，バリエーションも大きいコーパスを用意することが必要である．

ただし，文献 [3] の方法では，学習対象語彙を多く含むような記事を優先して選択するため，記事に使われる単語の一部は，学習語彙のレベルに自然に適合する．しかし，その他の単語については，学習対象語彙より難しい場合があるが，その点については，辞書の参照が容易にできる環境を用意することにより，対処することが可能と考える．したがって，コーパスの難易度レベルを学習者に適合させることは，結局，学習対象語彙の難易度を学習者に適合させることに帰着する．この点に関しては，文献 [3] で利用した TOEIC 学習用語彙は，日本の大学生をターゲットとし，その語彙を増すことを目的としたものであるため，十分に多くの学習者に適用可能なものであると考えている．しかし，各学習者への個別化も検討したい．

バリア 1 と 2 については，本稿執筆時点では，教材自体は，まだ作成中であるが，Wikipedia の英語版[4]が，再配布可能で，かつ，自由に編集でき，かつ，コンテンツとしても面白いものである，との見込みを得ている．

Wikipedia は，誰でもが自由に編集できる

[4] http://en.wikipedia.org/

事典である．その記事数は，2005年1月7日の時点では，約45万記事である．これらの記事は，GNU Free Documentation License により自由に編集や再配布ができる．また，その記事は，現代の内容を取り扱っているため，新鮮で，記事自体が面白い．そのため，Wikipediaを利用することにより，バリア1を克服すると同時に，バリア2も克服できる可能性がある．

作成される教材は，Wikipediaから，TOEIC学習用語彙を多く含む記事集合を抽出することにより作成される．このような記事集合は，同じ学習用語彙をカバーするようなものであっても，異なるものは何種類もあるので，バラエティに富んだ教材を作ることができる．これにより，学習者の多様性に対処できると考える．

教材として抽出する記事については，275単語以上の長さの記事とし，その中から，冒頭の150単語を抽出し，教材に採用する単位とする．この単位は短かいので，多少難しい単語があっても，辞書の参照が容易であれば，読むことができるだろうし，事典の冒頭であるので，事項全体の要約となっていることも期待でき，更に，興味があれば，Wikipediaの該当項目を直接読むこともできる．

したがって，学習者は，TOEIC学習用語彙のみで十分読むことが可能な記事集合を与えられると同時に，その範囲を越えた，実世界における英語テキストへのアクセスもできる．

ここで,「Wikipedia を教材とするなら，直接 Wikipediaを読めば良い」ということが言えそうであるが，それは，できない．なぜなら，語彙を限らないかぎり，Wikipediaの記事には，中学・高等学校レベルでは学習されない語彙が多いので，やみくもに読んでも，記事の内容が理解できる可能性は低いからである．

そのため，学習対象語彙を多く含むような，学習者のレベルに応じた記事をWikipedia全体から選びだし，教材として提示することが必要である．

この教材は，その他のものと同様，言の場で公開する予定である．

3 日英対応付けコーパスの英語学習への利用

単言語コーパスの英語学習への利用としては，語の用法を検索するものなどが考えられるが，それは，日本においては，まだあまり普及していない．その理由の１つは，前述のように，授業で（安価もしくは無料で）利用できるコーパス自体が少ないことである．しかし，たとえコーパス自体が利用可能であっても，なんらの編集もしない場合には，その難易度が，学習者にとって，高すぎることも，他の理由としてある．たとえば，文献[1]では，99種のテキストサンプルのうちで，1種のみが，中学・高等学校レベルの英語をマスターした段階で内容理解が可能なことを報告している．

そのような難易度の問題を避ける方法として，日英対訳コーパスを利用する方法がある[7]．文献[7]では，情報通信研究機構により公開されている「日英新聞記事対応付けデータ」[4]を利用して，語彙の学習をしている.「日英新聞記事対応付けデータ」は，約12年間分の「読売新聞」と *The Daily Yomiuri* とについて，翻訳関係にあるような文の対応を付けたものである．語彙を学習するためには，大規模なデータでないと，学習対象語彙自体が出現しないため，規模の大きな対訳データが必須である．

学習方法の概略は以下の通りである．まず，注目する日本語あるいは英語の単語について，それを当該言語のコーパスで検索する．次に，その検索単語と顕著に共起するような相手言語の単語を目印にして，日英の単語の用法を調べるというものである．この方法には，以下の利点がある．

- 学習者自らが，検索の過程において，語の多義性に気づく.（１つの日本語が複数の英語に対応する等）
- 辞書とは別の言語資源としてのコーパスの有効性に気づく.

このように日英対応付けコーパスを利用する方法は有望であるが，以下のようなソフト

ウェア上のバリアがある．

まず，一般に，コーパスとソフトウェアとは別個に開発されるため，当該ソフトウェアが当該コーパスを上手く扱えるとは限らない．また，計算機の初心者の場合には，ソフトウェアを使うまでに時間がかかる．更に，ソフトウェアは高価な場合が多いうえに，授業だけでのライセンスの場合には，自習することができない．また，日本語を上手く扱えないソフトウェアも多い．

これらソフトウェア上のバリアを除くためには，サーバサイドで検索プログラムを実行し，そこに，Webブラウザからアクセスすることが，現時点では，良い解決策である．したがって，言の場では，そのように運用している．

2005年2月現在の検索ソフトウェアは，SUFARY[5]をベースに作った文字列検索システムである．インターネットに接続できるところであれば，ユニバーサルにアクセス可能であり，利便性は高い．今後は，これを使いやすいものとすることを予定している．具体的には，前述のような使い方，すなわち，日本語と英語の共起関係を同定することや品詞による検索などは実装する予定である．

また，現在の検索対象は，「日英新聞記事対応付けデータ」に，「日英対訳文対応付けデータ」（再配布可能な小説などのデータ）[2]を加えたものであり，全部で，約25万件のデータがある．予定としては，来年度には，2002年以降の新聞記事データも対応付けデータとして追加し，30万件程度にしたい．更に，もし，検索された回数が数万件を超えるようになった場合には，検索ログも公開する予定である．検索ログの有用性は今のところ不明であるが，それを調べることにより，皆が調べたい英語や日本語表現が分かるのではないかと期待している．

[5] http://nais.to/~yto/tools/sufary/

4 おわりに

研究成果をエンドユーザーに直接届けることを目標にして，言の場（ことのば）を開設した．このサイトには，現在のところ，限られた言語資源しかないが，今後最低5年間ほどは運用し，言葉に関する各種の活動の場となることを目標としている．

謝辞

本稿に対して有益なコメントを下さった日本大学中條清美氏に感謝する．

参考文献

[1] Kiyomi Chujo, Masao Utiyama, Sachiko Sone, and Chikako Nishigaki. Increasing the effectiveness of parallel corpora through text analysis. In *The Sixth International Conference on Teaching and Language Copora*, 2004.

[2] Masao Utiyama. Japanese-English bilingual corpora and their applications. Asialex 2003, 2003.

[3] Masao Utiyama, Midori Tanimura, and Hitoshi Isahara. Constructing English reading courseware. In *PACLIC-18*, pp. 173–179, 2004.

[4] 内山将夫, 井佐原均. 日英新聞の記事および文を対応付けるための高信頼性尺度. 自然言語処理, Vol. 10, No. 4, pp. 201–220, 2003.

[5] 中條清美. 英語初級者向け「TOEIC語彙1,2」の選定と効果. 日本大学生産工学部研究報告 Vol.36, pp.27–42, 2003.

[6] 中條清美, 牛田貴啓, 山崎淳史, マイケル・ジナング, 内堀朝子, 西垣知佳子. ビジュアルベーシックによるTOEIC用語彙力養成ソフトウェアの試作III. 日本大学生産工学部研究報告, Vol. 37, pp. 29–43, 2004.

[7] 中條清美, 西垣知佳子, 内山将夫, 原田康也, 山崎淳史. 日英パラレルコーパスを利用した英語語彙指導の試み. 投稿中, 2005.