効率的な語彙獲得のための英文読解教材の作成

内山　将夫[†]　谷村　緑[†]　井佐原　均[†]

英文読解教材作成のための素材は豊富である．しかし，それらの素材を取捨選択して，1つのコースウェアとしての教材を作成するのは，困難である．本稿では，そのようなコースウェアとしての教材を，学習対象とする語彙とコーパスとから自動的に作成する方法を提案する．その方法によれば，学習対象である語彙をなるべくコンパクトに網羅するような文書集合を選択することができるので，それをコースウェアとすることにより，読解を通した効率的な語彙の獲得ができると考える．実験では，提案手法を，TOEIC学習用語彙と The Daily Yomiuri 新聞記事コーパスについて適用した．そして，作成された読解教材の種々の統計量を，無作為抽出の場合と比べることにより，作成された教材が，コンパクトに語彙を網羅していることが確認された．更に，作成された教材は，実際に，大学の英語の授業で補助教材として利用されており，授業に役立つとの見込みを得ている．

キーワード：　英文読解，語彙獲得，*e-Learning*，コーパス

Organizing English Reading Courseware for Effective Vocabulary Building

Masao Utiyama[†] , Midori Tanimura[†] and Hitoshi Isahara[†]

English reading materials are abundant on the Internet. However, it is still difficult to select proper materials to organize courseware that can be used throughout a one-semester English reading course. We proposed a method for constructing courseware from target vocabulary and corpus. This method was designed to extract a minimal set of articles (from the corpus) that contained the vocabulary. We applied the method to TOEIC (Test of English for International Communication) vocabulary and The Daily Yomiuri newspaper articles. The constructed courseware consisted of articles that had dense occurrence of the target TOEIC vocabulary. The degree of denseness was measured by comparing various statistics of the courseware with those of the randomly sampled articles. It was found that the courseware was efficient in presenting the vocabulary to students through reading. It is also used in English classes in one university as supplementary material and has been shown to be promising.

KeyWords:　*English reading, vocabulary acquisition, e-Learning, corpus*

† 情報通信研究機構, National Institute of Information and Communications Technology

1　はじめに

英文読解教材を作るための素材は豊富である．たとえば，ニュース記事であれば，CNN[1]やTIME[2]やBBC[3]などの記事があるし，また，EFL Reading[4] のように，外国語としての英語教育のための読解教材として，それぞれの学習者の語彙レベルに合せた文章を提供しているものもある．

しかし，これらの素材を組み合せて，1つのコースウェアとしての教材を作成するのは，困難である．ただし，ここでの教材とは，たとえば，大学の半期や通年の授業，もしくは，それに付随する自習教材として使えるような，1つの一貫したものであり，それを学習することにより，英文読解の技能が修得できることを目的とするものを指す．

このような教材の例としては，英文読解に特化したものではないが，たとえば，中学校や高校の英語の教科書がある．これらの教科書では，段階的に，新たな単語や文法事項などを含む文章を提示することにより，その文章を学習することにより，新たな単語や文法事項が段階的に修得できることを目的としている．

このようなコースウェアとしての教材を，自動的に作成するのが我々の目標である．特に，本稿においては，英文を読解するためには，ある程度の語彙を獲得することが必要不可欠であることから，作成された教材を学習することにより，そのような語彙が効率的に獲得可能となる教材を作成することを目的とする．

そのような教材作成の要件あるいは前提として，我々は，以下の2つを想定した．

- 獲得目標としての語彙を任意に設定できること．
- 教材作成に必要な素材としての文章 (文書) の集合がコーパスとして与えられていること．

そして，これらが満されるとき，与えられた語彙をなるべく「効率的」に学習できる教材を作成することを目的とした．ただし，ここでの「効率的」とは，以下のように定義した．

まず，読解を通して語彙を学習するためには，学習対象である語彙が読解教材に出現することが必要である．そのため，語彙全体を学習するためには，その語彙全体が読解教材に出現する必要がある．このとき，それら語彙全体を網羅するような読解教材には，その量に多少がある．つまり，多量の文章を読まなければならないか，あるいは，少量の文章を読まなければならないかがある．そのため,「効率的」の定義として，

定義 1 なるべく少量の文章を読むだけで，与えられた語彙が学習できること

を採用する．ここで，もし，与えられたコーパスから漫然と (無作為に) 文書を抽出して，それを読解教材とした場合には，与えられた語彙全体を網羅するようなものを得るためには，相当

1 http://www.cnn.com/
2 http://www.time.com/time/
3 http://www.bbc.co.uk/
4 http://www.gradedreading.pwp.blueyonder.co.uk/

多くの文書を抽出しないといけないと予想できる．そのため，本稿では，与えられた語彙を多く含むような文書を選択的に抽出する方法を提案する．

また，教材は，一般に複数の文章からなり，それらは，最初から順番に学習していくものである．そのため，場合によっては，教材の最初の方だけを授業でとりあげ，残りの部分を自習課題としたり，あるいは，そもそも，最初の方しか学習対象としないことが考えられる．そのため,「効率的」の定義として，

定義 2 教材の最初の方を学習するだけで，必要な語彙の多くが学習できること

も採用する．これは，たとえば，全教材に含まれる語彙の大きさを100%としたときに，たとえば，教材の最初の20%を学習するだけで，50%の語彙を網羅できるようなことを想定している．

このような教材が，もし，作成できれば，それは，特に，English for Special Purposes (ESP) としての英語教育に有効である．なぜなら，ESPにおいては，工学や医学や経済などの，学習者が必要とする特定分野の英語を教育することを目的とするのであるから，学習者に応じた教材を作成する必要があるため，そのような教材は，必然的に，学習者ごとに作成する必要がある．そのため，そのような教材は，学習者や学習対象に応じて柔軟に作成する必要があるのだが，そのような教材は，たとえ，素材としての文章が大量にあったとしても，人手で作成するのは，コストが掛かり，かつ，何らかの作成の指針がないかぎりは，作成も困難である．それに対して，もし，ESPのための語彙とコーパスとを与えるだけで，上述の定義のような効率的な読解教材が自動作成できれば，それは，非常に，教育および学習の役に立つと考える．

以上のような観点から，本稿においては，効率的な語彙獲得のための英文読解教材の作成を目的とする．その具体的な算法(アルゴリズム) と作成された教材の評価については，3節と4節とで述べるが，その前に，そのような読解教材を自動作成するときの一般的な方針として,「最適化としての教材作成」について，2節で述べる．

2　最適化としての教材作成

本節では，コースウェアとしての教材を自動作成するときの一般的な方針として,「最適化としての教材作成」という方針を提案する．この方針自体は，抽象的なものであるので，それを利用して直接教材を作成することはできないが，3節で述べる算法が，この方針の1つの具体化であることから，まず，その方針について述べる．

まず，用語を定義する．複数の文書からなるコーパスをDとする．D中の任意の文書の集合を$A(\subseteq D)$とする．なお，Aは順序付けされたリストであっても良い．Aがリストである場合には，そのリストの先頭の文書から学習を開始し，Aが集合である場合には，学習の順番は任意であるとする．次に，最適化の目的関数を$f(A)$とし，$f(A)$が大きいほど，Aは「良い」教材であるとする．また，Aについての制約を考え，それが満されているときには1，そうでないとき

には0を取る述語として，$\delta(A)$を考える．

以上の準備の下で，最適化としての教材作成においては，次のような$\hat{A}$を求めたい．

$$\hat{A} = \arg \max_{A \subseteq D, \delta(A)=1} f(A) \tag{1}$$

このような$\hat{A}$は，制約が満された「最良」な文書である．

ここで，制約$\delta(A)$としては，たとえば，文書長が300単語以下の文書しか含まないとか，あるいは，一定レベル以下の語彙しか含まないとかが考えられる．また，目的関数$f(A)$としては，たとえば，1節で述べた効率を表現するような関数が考えられる．

次に，$\hat{A}$を探索する算法について述べると，これは一般には，組み合せ最適化の問題である．つまり，コーパス中の全ての文書の組み合せのなかから，$\delta(A)$を満し，かつ，$f(A)$が最大のものを求める必要がある．この問題は，一般的には，効率的な探索方法がない問題であり，3節においても，貪欲(greedy)な算法により，近似解を求めている．

3　効率的な語彙獲得のための読解教材の作成算法

本節では，2節で導入した目的関数$f(A)$と制約$\delta(A)$の具体的な表現を述べ，次に，実際に利用した貪欲算法について述べる．

まず，1節で述べたように，我々は，コーパスDに加えて，語彙Vが与えられていることを前提とし，そこからV全体を網羅するような教材$A(\subseteq D)$を作成したい．そのため，制約としては，以下を考える．

$$\delta(A) = \begin{cases} 1 & (\forall w \in V)(\exists d \in A)(\text{文書}\, d \text{は単語}\, w \text{を含む}) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

この制約は，要するに，語彙V中の全ての単語wについて，それが最低1回は，いずれかの文書dに出現することを表現している．

また，読解教材の文書としては，長い文書は，学習者の負担が大きいことから，コーパスDには，あらかじめ，定められた長さ以下(本稿では延べ語数が300単語以下)の文書しか含まれていないようにする[5]．

次に，$f(A)$については，1節における効率性の定義1から，

$$f(A) = -(A\text{中の文書の文書長の和}) \tag{3}$$

とすることが考えられる．つまり，A中の文書長の和が小さい(短い)ほど「良い」教材であるとする．

5 そのため，本節で述べる算法は，300単語程度の短い文書からなるコーパスにおいて有効であることが(4節で)実証されるが，そのような長さの制限がない場合においても有効であるかは未検証である．しかし，我々が考察する読解教材においては，ある程度短かい(たとえば300単語の)文書しか対象としないので，以下の算法は有効である．

しかし，この定義による$f(A)$では，貪欲算法により制約を満しつつ近似解を求めるのは困難である[6]ので，実際には，後述するように，これまでに網羅されていないようなV中の単語を，なるべく多く含むような文書を順次選択するという貪欲算法により，近似解として$\hat{A}$を得た．この算法によれば，なるべく少ない文書数で与えられた語彙を網羅することができるので，結果として，$f(A)$が大きい文書集合を近似解として得ることができる．

実際の算法を，図1に，擬似コードにより示す．図1において，入力である，語彙Vは，学習対象である単語の集合であり，コーパスDは文書の集合である．また，後述する関数$g(d|V_x)$は，語彙V_xに対する文書dのスコアであり，$W(d)$は，文書dに含まれる単語の集合である．また，出力である文書リスト$\hat{A}$は，作成された教材である．$\hat{A}$に含まれる文書の順序は，それらが抽出された順番になっているので，$\hat{A}$を利用する際には，その先頭から利用する．それにより効率性の定義2を満すことができる．なお，図1の処理が停止する前提として，Vは，コーパス中の文書の単語集合の和集合に含まれなければならない．

入力：語彙V, コーパスD, 関数$g(d|V_x)$, $W(d)$
出力：文書リスト$\hat{A}$
前提：$V \subseteq \cup_{d \in D} W(d)$
処理：
$\hat{A} \leftarrow \phi$
$V_{\mathrm{todo}} \leftarrow V$
$V_{\mathrm{done}} \leftarrow \phi$
$\alpha \leftarrow 1$
while $V_{\mathrm{todo}} \neq \phi$
　$\hat{d} = \arg\max_{d \in D, W(d) \cap V_{\mathrm{todo}} \neq \phi} \left\{ \alpha g(d|V_{\mathrm{todo}}) + (1-\alpha) g(d|V_{\mathrm{done}}) \right\}$
　$\hat{A} \leftarrow \hat{A} \cup \{\hat{d}\}$
　$V_{\mathrm{done}} \leftarrow V_{\mathrm{done}} \cup (V_{\mathrm{todo}} \cap W(\hat{d}))$
　$V_{\mathrm{todo}} \leftarrow V_{\mathrm{todo}} - V_{\mathrm{done}}$
　$\alpha \leftarrow \frac{|V_{\mathrm{done}}|}{1+|V_{\mathrm{done}}|}$
　$D \leftarrow D - \{\hat{d}\}$
end

図 1　効率的な語彙獲得のための読解教材の作成算法

図1の処理においては，V_{todo}には，まだ$\hat{A}$中の文書に網羅されて(含まれて)いないような

6 この問題の厳密解を求める方法としては，たとえば，整数計画法による集合被覆問題に対する解法がある (ウィリアムス 1995)．本稿の実験で，その解法を用いなかった理由は以下のようである．すなわち，まず，我々は，実際に大学の授業で使うことを意図して教材を作成したため，その授業での運用に間に合うように教材を作成する必要があった．そして，その必要のために，まず，実装が容易である貪欲算法により教材を作成した．そして，その結果として作成された教材をみたところ，その教材は，十分使用に耐えるようであった．そのため，新たな方法を試すより，その教材を授業で使うとともに，その品質を評価することを優先した．

単語を格納し，V_{done}には，既に網羅されている単語を格納する．そして，これら2つを利用して，4式に示す各文書のスコアをもとめて，そのスコアが最大の文書$\hat{d}$を，$\hat{A}$に追加する．その後，$\hat{d}$に含まれるV_{todo}中の単語については，既に網羅されたとしてV_{done}に移して，次の文書を求める，ということを与えられた語彙中の全ての単語が網羅されるまで繰り返す．

図1の算法が，効率性の定義1 , 2に沿うような$\hat{A}$を出力するためには，文書dについてのスコア

$$G(d|\alpha, V_{\mathrm{todo}}, V_{\mathrm{done}}) = \alpha g(d|V_{\mathrm{todo}}) + (1-\alpha) g(d|V_{\mathrm{done}}) \tag{4}$$

が，これまでに網羅されていないV中の単語V_{todo}を，なるべく多く含むようなときに，大きい値とならなければならない．そうするために，まず，$g(d|V_x)$を以下のように定義する．

$$g(d|V_x) = \frac{k_1+1}{k_1((1-b)+b\frac{|W(d)|}{E(|W(\cdot)|)})+1}|W(d) \cap V_x| \tag{5}$$

ここで，k_1とbとは，経験的に定める定数であり，本稿の実験においては，$k_1 = 1.5, b = 0.75$である．なお，5式の$g(d|V_x)$は，情報検索で使われているBM25 (Robertson and Walker 2000)という尺度を簡略化したものである．$g(d|V_x)$がBM25に基づいた理由は，BM25は，情報検索において，高精度に質問と関連した文書を検索できることが実証されているので，それに基づいた$g(d|V_x)$を利用することにより，V_xと良く関連した文書が得られると考えたためである．また，簡略化の概略は以下の通りである．すなわち，BM25においては，各単語の重みとして，IDF (Inverse Document Frequency)や，単語の文書d中での出現頻度等が考慮されているのであるが，5式では，それら重みを全て1としている．なお，重みを全て1にしているということは，語彙V中の単語を，全て同等に扱っているということである．しかし，もし，特に優先したい単語がある場合には，その単語の重みを大きくするということも考えられるが，本稿の実験では，それは試みてはいない．

さて，5式において，$|W(d) \cap V_x|$は，文書d中の単語とV_xとの共通単語数であるので，$g(d|V_x)$は，共通単語数が大きいときに大きい値となる．更に，$|W(d)|$は，文書dにおける異なり単語の数であり，$E(|W(\cdot)|)$は，そのコーパス全体における平均値である．そのため，$g(d|V_x)$は，$\frac{|W(d)|}{E(|W(\cdot)|)}$の影響により，異なり単語の数が少ない(文書長が短かい)文書において大きい値となる．すなわち，$g(d|V_x)$は，V_xと共通単語数が多く，かつ，文書長が短かい文書において大きい値となる．

したがって，もし，3式の$f(A)$だけを貪欲に近似しようとするなら，$g(d|V_{\mathrm{todo}})$のみを，$G(d|\alpha, V_{\mathrm{todo}}, V_{\mathrm{done}})$の代わりに，図1の算法において利用することも考えられる．しかし，我々は，まだ網羅されていない単語V_{todo}との共通単語数に加えて，既に網羅されている単語V_{done}との共通単語数も大きいような文書のスコアを大きくすることを考えたので，$g(d|V_{\mathrm{todo}})$と$g(d|V_{\mathrm{done}})$の組み合せである$G(d|\alpha, V_{\mathrm{todo}}, V_{\mathrm{done}})$をスコアとして利用した．

次に，重みαの計算法について述べる．αの計算においては，V_{todo}との共通単語を，V_{done}

との共通単語よりも，ずっと優先させたいので，V_{todo} での 1 単語の共通が，V_{done} 全体での共通と同程度になるように α を設定することにした．そのためには，計算を簡単にするために，5 式において，文書長の影響を除いて，$|W(d) \cap V_x|$ のみを考えることにすると，

$$G(d|\alpha, V_{\mathrm{todo}}, V_{\mathrm{done}}) = \alpha|W(d) \cap V_{\mathrm{todo}}| + (1-\alpha)|W(d) \cap V_{\mathrm{done}}|$$

となる．ここで，V_{todo} での 1 単語の共通が，V_{done} での全単語での共通と同程度なことから

$$\alpha \times 1 = (1-\alpha) \times |V_{\mathrm{done}}|$$

とした．そして，これを満すものとして，

$$\alpha = \frac{|V_{\mathrm{done}}|}{1 + |V_{\mathrm{done}}|}$$

を採用した．

以上より，図1の算法が，これまでに網羅されていない単語を多く含む文書を，優先して教材に採用することが言える．

4　実験

本節では，まず，学習対象語彙とコーパスについて述べる．次に，それらから作成された教材 (以下では作成教材と呼ぶ) の性質を調べる．

4.1　語彙

学習対象とする語彙としては，(中條 2003; 中條, 牛田, 山崎, マイケル・ジナング, 内堀, 西垣 2004) により作成された TOEIC[7] 学習用語彙「レベル 1(補習向け)」「レベル 2(初級向け)」「レベル 3(中級向け)」を用いた[8]．この語彙は全部で 640 項目からなり，それぞれのレベルの項目数は，レベル 1 が 200, レベル 2 が 200, レベル 3 が 240 である．これらの選定の詳細は (中條 2003) に記述されているが，そのリストの性質は，端的には，以下の通りである (中條 内山 2004).

> これらは語彙選定を専門とする英語教育者が，頻度・分布度を基準に語彙リストを作成した後，学習の容易性・学習の必要性等の主観的判断を加え，中学校教科書出現語を除去した後，さらに，高校・大学英語教科書語彙等の外部資料との比較を繰り返して学習段階別に選択・配列したものである．

これらの語彙は 3 レベルに分かれているが，本実験においては，特に，それらを区別することなく，全語彙 640 項目を学習対象の語彙とした．なお，これら項目において，複数単語を含むものについては，その全ての単語を学習対象とし，また，後述のコーパスに出現しない単語

7 Test of English for International Communication (`http://www.toeic.or.jp/toeic/index.html`)
8 これらの語彙は http://www5d.biglobe.ne.jp/~chujo/ で公開されている．

については，それを語彙から除去した．その結果，全部で 642 単語を学習対象の語彙とした[9]．

なお，本節の実験において TOEIC 用の学習語彙を用いた理由は，本稿における作成教材を授業で利用する大学においては，TOEIC を念頭において授業が構成されているので，その補助教材として利用することを考えたためである．また，TOEIC を対象にした語彙として，特に (中條 2003) の語彙を用いた理由は，それが教育・研究資料として公開されていることに加えて，実際に，学習効果が高く，信頼性の高い語彙であることが報告されているからである．

4.2　コーパス

教材作成のためのコーパスとしては,「読売新聞記事データ」における「The Daily Yomiuri」の 1989 年から 2001 年までの約 11 万記事を元データとした．これらの記事から，読解のときの負担が軽くなるように，300 単語[10]以下のものを選ぶと約 4 万 1 千記事である．

ここで，The Daily Yomiuri の記事の特徴として，それと内容が対応するような読売新聞記事が，記事によっては存在するというものがある．そのため，そのような特徴を生かした教材を作成する可能性も考慮して，上述の 300 単語以下の記事から，更に，読売新聞において対応記事が存在する可能性が高いもののみを選ぶと約 2 万 5 千記事である．ただし，対応記事については， (内山 井佐原 2003) において対応付けされた記事対のうちで，対応付けの信頼性が高いと報告されている，対応付けスコア 0.111106681 以上の記事対を利用した[11]．

この約 2 万 5 千記事に対して，以下の処理を適用した結果のコーパスに対して，3節の算法を適用した．その処理は，

- 各記事を 1 文単位に整形したあとで，
- 各文の各単語に品詞をタグ付けし，基本形を得る

というものである[12]．なお，ソフトウェア上の制限として，基本形は，名詞・動詞・形容詞・副詞のいずれかについてしか求めることができなかった．

次に，図 1の算法を実際に利用するという観点から，図 1における変数と上述の語彙とコーパスを関連づける．まず，図 1の V は，上述の 642 単語からなる TOEIC 学習用語彙である．次に，コーパス D は，上述の約 2 万 5 千記事である．あとは，D 中の d に対する $W(d)$ を定義すれば，図 1の算法を実際に利用することができる．その定義は，本実験においては，記事 d 中の全

9 元のリストとの違いは以下の通りである．まず，削除した項目は「interoffice」と「complimentary」の 2 項目である．また，複数単語を含むものは「advertisement (ad, ads)」「incorporate (INC)」「PIN (personal identification number)」の 3 項目である．このうち，1 番目については,「advertisement」「ad」「ads」を学習対象とし，2 番目については,「incorporate」「Inc」を学習対象とした (「INC」はコーパス中に出現しないが「Inc」は出現したため)．3 番目については,「PIN」「personal」「identification」「number」が含まれるが，このうち「personal」「identification」は既に別の項目としてあるため,「PIN」と「number」のみを学習対象とした．(「number」については既習とみなして学習対象から除去することも考えたが，元の語彙をなるべく変更しないという方針から，これも学習対象に加えた．)

10 この単語数は，The Daily Yomiuri に記事情報の 1 つとしてつけられている単語数をそのまま利用した．

11 対応付けデータは `http://www2.nict.go.jp/jt/a132/members/mutiyama/jea/index.html` で公開されている．

12 この処理においては，ソフトウェアとして，`mxnl`, `tokenizer.rb`, `mxetag`, `elemma` を利用した．これらは `http://www2.nict.go.jp/jt/a132/members/mutiyama/software.html` で公開されている．

ての単語，および，それに加えて，各単語の可能な基本形の全てからなる集合である．たとえば，ある単語が，動詞「saw」の場合には，その基本形には,「see」と「saw」の2通りの可能性があるので，これら単語「saw」と基本形「see」「saw」の集合として，{see, saw}が$W(d)$に寄与する．$W(d)$は，そのような各単語に関する単語集合の和集合である．このような和集合を$W(d)$に利用する理由は，基本形の同定には誤りが含まれるからである．すなわち，もし，各単語の基本形の同定が，曖昧性なく正確にできるならば，$W(d)$として，各単語の基本形の和集合のみを利用することも考えられるが，実際には，基本形の同定は，誤りを含むため，そのような誤りが含まれている場合であっても，V中の単語が，なるべくもれなくコーパス中の単語に一致するように，基本形に加えて，元の単語も$W(d)$に含めることにした．

以下では，これらのコーパスと学習対象語彙に，3節の算法を適用した結果として得られる作成教材の性質について述べる．

4.3　作成された教材の例

本節では，作成教材における記事の例として，その最初の記事を図2に示す．図2の記事の本文では，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．

図2の記事は，3節の算法を，上述の学習対象語彙とコーパスに対して適用した結果として得られる最初の記事であるから，その算法における，最良の記事である．したがって，2番目以降に教材として採用される記事については，共通単語数は減少していく．しかし，減少していったとしても，ただ無目的に記事を集めるのに比べれば，学習対象語彙中の単語を多く含む記事を抽出できる．そのことについては，次節でより詳しく調べる．

4.4　作成された教材の性質

本節では，作成教材が，1節で述べた効率性の定義1 , 2を満すことを示す．そのときに，効率性の比較の対象として，無作為に記事を抽出したときと比べての有効性を示す．ここで，無作為抽出と比較する理由は，もし，作成教材が，無作為抽出の記事集合と比べて，それほど効率的でないのであれば，わざわざ3節の算法を利用する必要はないため，比較のベースラインとして利用する．

まず，作成教材における基本的な統計量を調べる．まず，その教材を構成する記事数は116である．次に，各記事の基礎統計量を表1に示す．

表1において,「記事長」とは，各記事における延べ語数である[13]．次に,「Vとの共通延べ語

13 延べ語数を計数するときには，空白で区切られた，アルファベットを含む文字列のみを単語として計数し，その他の，たとえば，ピリオド等の記号や数字のみからなる文字列については，単語としては計数していない．このことは，以下で単語の数に言及するその他の場合においても同様である．こうした理由は，ピリオド等の記号や数字のみからなる文字列については，特に学習を必要とせずともその意味が分かる(あるいは既習と考えられる)ため，それを計数しない方が，学習が必要な単語のみを対象とした場合における，作成教材の統計的性質を調べるのに妥当であると考えたためで

296 words – 2001/11/09

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**.

The move comes amid sluggish **demand** and deteriorating **earnings** by the two *regional* phone *operators*, which are currently cutting **rates due** to intensifying competition.

NTT's **labor union**, which reached a broad **agreement** on the *company*'s restructuring plan at its August **convention**, **approved** the **management**'s plan in its extraordinary central **committee meeting** on Thursday.

The *union* also *approved* the *management*'s **offer regarding** the **amount** of *compensation* to be paid to *transferred* **employees**.

The restructuring plan will *cost* NTT about 1.1 trillion yen in *retirement allowances* and about 300 billion yen in *allowances* to compensate for average cuts of 55 percent in workers' lifetime **wages**. NTT will likely cover the *expenses* with **bank loans** or by selling land and other **corporate** assets.

図 2　作成教材における記事の例

表 1　各記事の基礎統計量 (記事数=116)

	平均	標準偏差	最小値	最大値
記事長	180.2	65.2	44	296
Vとの共通延べ語数	25.3	14.6	1	65
Vとの共通異なり語数	17.4	8.8	1	43
Vとの共通新出異なり語数	5.5	7.2	1	43

数」とは，上述したTOEIC学習用語彙Vについて，各記事において，Vに含まれる単語の延べ語数のことである．ただし，ある単語がVに含まれるとは，その単語自体，あるいは，その単語の可能な基本形のいずれかがVに含まれることと定義する[14]．また,「Vとの共通異なり語数」とは，計数の方法を延べ語数の場合と同様にした場合における，異なり語数のことである．最後に,「Vとの共通新出異なり語数」とは図1におけるV_{todo}と各記事との共通異なり語数の

ある．

14 たとえば，前述の「saw」については,「see」と「saw」の2単語が該当する．しかし，実際のところ，V に含まれる語彙で，このように，記事中の1つの単語がV中の2つの単語に相当する例は,「ads」の基本形が「ad」であり，その結果「ads」と「ad」の双方が該当する例しかなかったので，このようにしても問題はなかった．なお，この場合でも,「ads」は延べ語としては1単語として計数される．

ことである．これは，要するに，教材を最初の方から読んでいって，当該記事に到達したときに，その記事において，これまでにV中で網羅されていないような単語が，異なり語数で何単語網羅されているかである．

次に，これらの統計量，および，その他の統計量が，無作為抽出の場合と比較して，どの程度の量であるのかを調べることにより，作成教材が，定義1の意味で効率良くVを網羅していることを示す．

その比較にあたっては，まず，無作為抽出の場合の記事集合(以下では「無作為記事セット」と呼ぶ)を得ないとならない．そして，このときに，無作為記事セットにおける延べ語数は，作成教材と同じでなければならない．その理由は，延べ語数が変わると，それにともない，種々の統計量も変動するので，適切な比較のためには，延べ語数を揃える必要があるからである (影浦 2000)．そのため，ある1つの無作為記事セットをつくるときには，コーパス全体から，作成教材における延べ語数20900語と同じだけの延べ語数となるまで記事を無作為に抽出することにした．このとき，最後の記事の途中で，20900語を越えた場合には，その記事については，20900語で打ち切った．このようにして，1000個の無作為記事セットを作成し，それらについて，種々の統計量を計算し，作成教材と比較した．その一覧を表2に示す．

表 2　統計量の比較

	傾向	教材	平均	SD	上側	下側	超	以下	以上	未満
記事数	多い	116	111.1	3.4	0.076	0.924	62	938	99	901
平均記事長	短い	180.2	188.3	5.9	0.918	0.082	901	99	938	62
平均共通延べ語数	多い	25.3	19.3	1.1	0.0	1.0	0	1000	0	1000
平均共通異なり語数	多い	17.4	12.8	0.6	0.0	1.0	0	1000	0	1000
平均共通新出異なり語数	多い	5.5	3.6	0.1	0.0	1.0	0	1000	0	1000
網羅率	高い	1.0	0.616	0.016	0.0	1.0	0	1000	0	1000

表2において，各行は，各種統計量を示す．各列については,「傾向」は，作成教材の当該統計量が，無作為記事セットと比べてどういう傾向にあるかであり,「教材」の数値は，作成教材における当該統計量の値である．次に,「平均」と「SD」は，当該統計量の無作為記事セット全体における平均と標準偏差である．また,「上側」と「下側」は，当該統計量が，無作為記事セットの平均と標準偏差による正規分布をすると仮定したときに， 作成教材の数値を超える値となる確率(上側確率)と，それ以下の値となる確率(下側確率)である．なお，上側確率が「0.0」となっている欄における確率は，10^{-7}未満である．最後に,「超」「以下」「以上」「未満」にある数値は，それぞれ，作成教材の数値「を超える」「以下」「以上」「未満」の数値であるような無作為記事セットの数である．そのため,「超」と「以下」の和，および,「以上」と「未満」の和は1000である．

表2においては，前述の通り，各記事セットの延べ語数は20900語と一定である．そのため，記事数が多いということは，平均記事長が短いということと同じことである．これら記事数と平均記事長について，上側確率と下側確率は，それぞれ，0.076 と 0.082 であるので，これらの

傾向は，統計的には有意というほどではないが，記事数は多い，あるいは，平均記事長は短いといえる．このことは，図1の算法の説明で述べたように，5式が短い記事を優先することを示している．

次に，表2の「平均共通延べ語数」「平均共通異なり語数」「平均共通新出異なり語数」の行には，それぞれ，表1の対応する統計量の各記事あたりの平均値についての数値がある．これらについては，上側確率等を見ると分かるように，無作為記事セットが，作成教材よりも大きい数値をとることは，ほぼないと言える．すなわち，作成教材の数値は，無作為記事セットの場合と比べて，統計的に極めて有意に大きい．これより，作成教材は，学習対象語彙を多く含む記事集合を選択していると言える．

最後に,「網羅率」とは，学習対象語彙Vのなかで，実際に教材に出現した単語の，V全体に占める割合である．これは，作成教材については，図1の算法の性質から，かならず1になる．しかし，無作為記事セットについては，そうなるとは限らないし，実際に，網羅率の平均は0.616である．すなわち，無作為記事セットにおいては，Vの約40%が出現しない．また，上側確率も 0.0 である．そのため，作成教材は，無作為記事セットと比べて，定義 1 の意味で効率良くVを網羅していると言える．

以上においては，無作為記事セットとの比較により，効率性の定義 1 の観点から，作成教材の性質を述べた．以下においては，作成教材のみに注目して，効率性の定義 2 の観点，つまり，「教材の最初の方を学習するだけで，必要な語彙の多くが学習できること」という観点から，その性質を調べる．

まず，図3に，図1の算法により教材に採用された記事の順位 (article ranking) が下がるにつれて，学習対象語彙Vとの共通異なり語数 (num. of types) が減少する様子を示す．図3の横軸は，記事の採用された順位 (順番) であり，縦軸は，異なり語数である．そして，下部の実線 (num. of new types) は，Vとの共通新出異なり語数であり，十字の点 (num. of common types) は，Vとの共通異なり語数である．これより分かるように，共通単語数は，順位が下がると徐々に減少する．このことは，特に，共通新出異なり単語について，顕著である．このことは，4式のスコアが，共通新出異なり単語の少ない記事において小さいことを裏付けている．しかし，図3 より分かるように，新出単語が少ない記事であっても，その他のVの語彙は出現するので，新出単語と既出単語を関連づけることはできると言える．

次に，図4に，全語彙Vについて，網羅された異なり語数の増加の様子を実線で示す．図4の横軸が記事の順位 (article ranking) や延べ語数 (num. of tokens) で，縦軸が，V で網羅された異なり語数である．これより，最初の方における，異なり語数の増加が大きいことがわかる．図4では，延べ語数が100増えるごとに，Vでの異なり語数を数えたときに，初めてVの50%および90%を超えるところに横線を引いてある．それらは，それぞれ，異なり語数が326語と578語のところである．そして，このときの延べ語数は，3700語と13900語であり，記事の順位は，

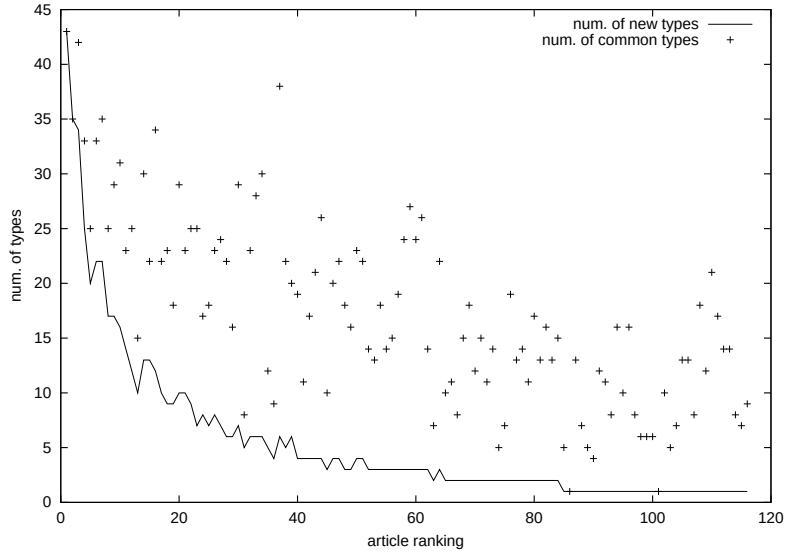

図 3　記事の順位と重複異なり単語数

17位と 68 位である．そしてこれらは，延べ語においては，18%および 67%に相当する．つまり，作成教材の最初の 18%や 67%で，全学習対象語彙の 50%や 90%を網羅している．

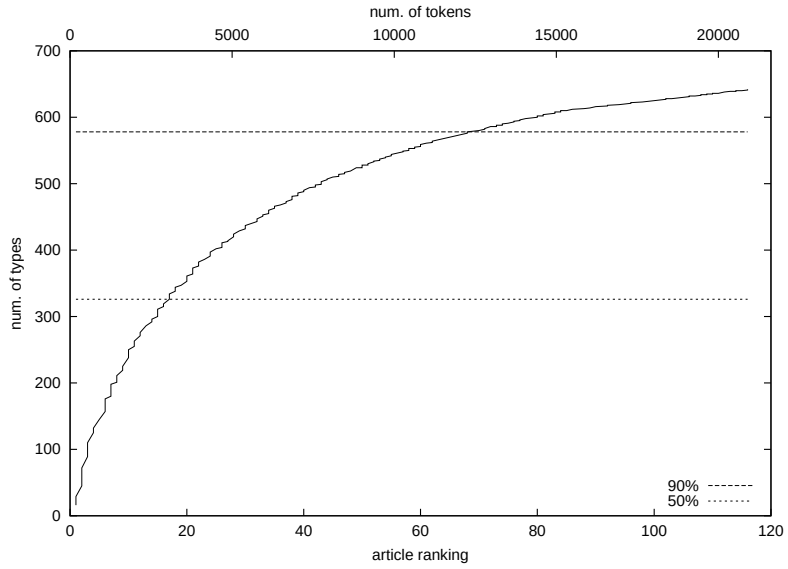

図 4　記事の順位や延べ語数と異なり語数(全体)

以上より，定義 2 の意味での効率性があることが言える．しかし，このことは，逆に言えば，教材の後ろの方では，学習対象語彙の異なり単語の提示の効率が悪いことを意味する．実

際，図3の最後の方では，新出異なり単語は，各記事あたり1単語程度になる．このことから，もし，この教材の目的を，語彙の獲得のみであると考えた場合には，教材の最後の方は，それほど有効なものではない．しかし，もし，語彙の獲得については，何らかの単語学習を主とすることとし，この作成教材は，学習された単語の定着を目的とした補助教材として利用するのであれば，このように最後の方で，新出異なり単語が少なくなったものであっても，十分に有用であると考える．

5　議論

5.1　作成教材の利用例

作成教材は，2004年5月から，関西地区のある大学において，実際の授業で，補助教材として利用されている[15]．授業においては，自習課題の記事における，TOEIC語彙についての確認テストをしている．

そのときの経験から以下のことが言える．まず，学生に与える教材の動機付けは非常に高い．学習者自らが「確認テストを毎回授業でして欲しい」と申し出たほどで，中級程度の学生は常に高得点を収めている．次に，単語の学習方法として，単語のリストではなく，文脈のなかで単語を学習することにより，学習者は文脈により単語の意味が変化することを実感している．たとえば，ある学習者は，”Takashi Kitaoka, president of Mitsubishi Electric Corp., said...”という文に遭遇したときに,「初めてpresidentに大統領以外の意味があるのに気がついた」と述べ，自身の小さな発見に喜びを示した．また， 読解自体に対する抵抗感も減少するという効果もあるようで,「教科書が簡単に思えるようになった」というコメントも得ている．

5.2　作成教材の問題点

作成教材の問題点としては，以下の2点がある．まず，図1の算法においては，単語のみしか考えていないため，熟語が上手く扱えない．たとえば,「ad hoc」という熟語において，この「ad」と,「advertisement」の略語としての「ad」とがマッチするという問題がある．これに対処するには，テキスト中の熟語を認識する必要がある．次に，教材作成に利用されたTOEIC語彙とThe Daily Yomiuriとにおける，語義のズレの問題がある．たとえば,「agency」という単語は，TOEICでは,「代理店」という意味となることが多いが，The Daily Yomiuriでは,「政府機関」という意味での「agency」も多い．これに対処するには，与えられた語彙に良くマッチしたコーパスを利用する必要がある．

これらの例については，単語の字面だけをみる処理では対処するのが困難であるが，本実験で作成された教材については，この問題は，それほど多くは生じていないようであった．

15 ただし，本稿執筆の段階(2004年11月)では，作成教材は，まだ利用の途中であるので，教材の有効性について，確定的なことを述べることはできない．そのため，本節においては，授業における学生の態度などについて述べる．

5.3　関連研究

音声言語処理技術を教育に利用しようという試みは盛んである．たとえば，最近のワークショップとして，(Burstein and Leacock 2003)には，英語のエッセイの自動採点とか，文法とか発音のチェックとか，テスト問題の自動作成とかに，音声言語処理技術を利用する研究がある．また，(内山, 佐野, 菅谷, 宮田, 中條, 西垣, 原田 2003)においても，コーパスや音声言語処理技術の言語教育への応用が論じられている．

これらの研究と本稿での研究の主要な相異点は，本稿での研究が，たとえば，大学の半期や通年の授業，もしくは，それに付随する自習教材として使えるようなコースウェアとしての教材を自動的に作成することを目的としているのに対して，これらの研究では，コースウェアというよりは，ある授業計画のなかでの個々の技能に関する授業や自習における自動化を目的としていることである．そのため，我々の研究は，これまでの研究と相補的なものであると言える．

次に，本稿においては，学習対象の語彙とコーパスとが与えられていることを前提としたが，学習対象の語彙については，その選定を補助するために，学習対象のコーパスから特徴的な単語を抽出する研究がある (中條・内山 2004; 内山, 中條, 山本, 井佐原 2004)．つまり，学習対象のコーパスがあれば，そこから，語彙は抽出可能である．したがって，これらの研究と本稿での研究とを組み合わせることにより，学習対象のコーパスがあれば，比較的容易に，語彙と読解教材とを作成できることが期待できる．

6　おわりに

本稿では，与えられた学習対象語彙とコーパスとから，その語彙を効率的に獲得できるような読解教材を作成することを目的とする算法を提案し，それにより作成された教材について，その性質および利用例を述べた．作成された教材について，そこに含まれる学習対象語彙の数を調べたところ，それは，無作為抽出された記事集合に比べて，統計的に極めて有意に大きく，作成された教材の有効性が示された．今後の課題は，作成された教材の実際の授業における有効性を検証することである．

参考文献

Burstein, J. and Leacock, C. (Eds.) (2003). *HLT-NAACL 2003 Workshop on Building Educational Applications Using Natural Language Processing.*

中條清美 (2003). “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.

中條清美 内山将夫 (2004). “統計的指標を利用した特徴語抽出に関する研究.” 関東甲信越英語教育学会研究紀要第18号 pp. 99–108.

影浦峡 (2000). 計量情報学. 丸善株式会社.

Robertson, S. E. and Walker, S. (2000). “Okapi/Keenbow at TREC-8.” In *Proc. of TREC 8*, pp. 151–162.

内山将夫, 佐野洋, 菅谷史昭, 宮田高志, 中條清美, 西垣知佳子, 原田康也 (2003). “パネル討論:言語教育・言語学習と知的情報処理研究.” 電子情報通信学会　思考と言語研究会 (TL)2003年12月.

内山将夫 井佐原均 (2003). “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, **10** (4), 201–220.

内山将夫, 中條清美, 山本英子, 井佐原均 (2004). “英語教育のための分野特徴単語の選定尺度の比較.” 自然言語処理, **11** (3), 1–33.

ウィリアムスH.P. (1995). 数理計画モデルの作成法. 産業図書.

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 (2004). “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, **37, 29–43.**

略歴

内山 将夫: 1997年筑波大学大学院工学研究科博士課程修了. 博士（工学）. 1997年信州大学工学部電気電子工学科助手．1999年郵政省通信総合研究所非常勤職員. 現在，独立行政法人情報通信研究機構主任研究員．ACL，ACM, 言語処理学会，情報処理学会，人工知能学会，大学英語教育学会，英語コーパス学会，外国語教育メディア学会，各会員.

谷村 緑: 2004年大阪外国語大学大学院言語社会研究科博士後期課程修了．博士（言語文化学）. 現在，独立行政法人情報通信研究機構短期専攻研究員．コーパス研究、応用言語学研究に従事. 英語コーパス学会，外国語教育メディア学会，全国英語教育学会, 日本第二言語習得学会, 大学英語教育学会，各会員.

井佐原 均: 1980年京都大学大学院修士課程修了．博士（工学）．同年通商産業省電子技術総合研究所入所．1995年郵政省通信総合研究所．現在，独立行政法人情報通信研究機構けいはんな情報通信融合研究センター自然言語グループリーダーおよびタイ自然言語ラボラトリー長．自然言語処理，語彙意味論の研究に従事．言語処理学会，情報処理学会，人工知能学会，日本認知科学会，ACL，各会員.