Vol.2012-CE-115 No.6
2012/7/7

適応型単語リストを用いた自律学習支援システムの構築

堀江郁美† 山口和紀†† 柏原賢二†† 飯島優雅†

現在，インターネット上には学生の自律学習の教材となり得るデータが，言語や分野を問わず多数存在している．しかし，単語の意味がわからなかったり，探しているデータを探し当てることができなかったりするため，学生はほとんどを教材として利用できずにいる．

そこで，本研究では，データに含まれる単語に着目し，学生の履修授業や文書の分野などをヒントに，データ間のコサイン類似度を計算し、類似しているデータに付属する辞書を用い単語の意味やデータを推薦し，インターネット上の資料を教材として利用できるようにする自律学習支援システムを提案する．

Improvement of Personalized Teaching Material Generator Based on Word Set

IKUMI HORIE† KENJI KASHIWABARA†† KAZUNORI YAMAGUCHI††
YUKA IIJIMA†

We propose the concept of a word set for constructing a teaching material generation system. The system suggests the meanings of words in reading materials for a student and generates a personalized glossary. We improved a previous prototype system based on this concept. The similarity of the reading materials is estimated by the cosine similarity in the proposed system. Then, some experiments verified that the system is promising.

1. はじめに

近来，コンピュータの性能や通信技術の発達により，ICT(Information and Communication Technology)ツールを教育に活用する試みが多くなされている[1,2]．実際，全国の国公私立大学，短期大学および高等専門学校で，多くの授業においてコンピュータが用いられ，学生はコンピュータを前にして授業を受けたり，レポートを作成したりしている．しかし，Web 技術がこれだけ身近なものになり，膨大な数の教材となり得るデータが Web 上のサービスとして提供されているにも関わらず，それらを活用できている教員や学生はほとんどいない．そこで，本研究では，Web 上のテキストデータの教材としての利用に焦点をあて，自律学習の支援システムを提案する．

Web 上のデータが教材として利用されない理由の一つに学生の知識・学習レベルが不均一・不透明であることがあげられる．入試や英語のレベルわけテストなどのおかげで，受講生のある程度のレベルを決めることのできる授業も存在するが，しかし，どの学生をとっても全く同じ知識レベルを持ち，同じ速度で学習するとは考え難い．学生の知識・学習レベルが不均一だと，例え教員が教材として Web サイトを紹介したとしても，その Web サイトが全ての学生に適しているかは不明である．また，学生が自律学習用に教材を探したとしても Web サイトがたくさんあり，どの Web サイトが自分に適しているかどうかがわからず見つけられないことが多々ある．これらのことから，学生にあわせた補助機能を持つシステムが望まれていることがわかる．そこで，本研究では，GSL(General Service List)[3,4]や AWL(Academic Word List)[5,6]を用いて，学生の申告するレベルにあわせて記事から重要と思われる単語を抜き出し，辞書から意味をひき表示するシステムを作成した[7,8]．しかし，辞書に関する問題などがあったため，本論文では，既存のプロトタイプシステムに，個々の学生や Web サイトの内容にあわせた単語の意味やデータを推薦する機能を持つように改良することにした．

関連研究としては，e-learning システムを採用した語学学習サイトやゲーム機器を用いた単語学習ゲームソフトなどが多々存在する．語学学習サイトには，英単語を覚えるための単語帳サイトや，Web 教材の英単語を抜き出すサイト[9]などがある．しかし，これらのサイトやソフトには掲示板や単語ゲームなど多数の機能を持つ非常に優れた語学学習サイトであるにも関わらず，利用者の知識レベルにあわせたコースを柔軟に選択できなかったり，教材の比較機能，推薦機能などがなかったりする．また，語学以外の分

† 獨協大学
Dokkyo University
†† 東京大学
The University of Tokyo

野の学習機能が存在しない．本研究で開発するシステムでは，利用者が作成する単語集を用いて利用者に適したコースが選べ，教材を比較できる上に，言語や分野を選ばず学習できる点が異なる．他に，学生に人気の気軽に利用できるサイトに，翻訳サイトがある[10,11]．これらの翻訳サイトは，語学の授業の予習などによく使われているが，本研究が目的としているのは翻訳ではなく，文章読解の学習支援であり，単語の意味推薦機能である．実際，学生は翻訳サイトのみを利用するのではなく，翻訳した結果を見て辞書の単語と照らし合わせて学習している．本研究では，自律学習には文章の翻訳ではなく，単語の意味を推薦する方が好ましいと考える．

本論文では，2 章で既存のシステムの概要と課題，3 章で新たに追加する推薦機能と実験について説明し，4 章で結果をまとめる．

2. プロトタイプシステム

ここでは，既存のプロトタイプシステムについて説明する．まず，2.1 章において，単語の推薦の必要性について説明する．次に 2.2 章でプロトタイプシステムの概要，2.3 章でプロトタイプシステムの課題について述べる．

2.1 支援システムの必要性

本研究では，単語を学習の根底部分であると捉え，英語学習のみならず，どの分野においても単語取得は最重要テーマであると考えた．そこで，英語の教科書が提示する重要単語と学生の関連をアンケート調査した．実験詳細は，以下の通りである．

- 日時：2012 年 6 月 8 日
- 人数：20 名（うち女性 5 名）
- 学年：経済学部 3 年生
- 形式:「ハイスピード実現! TOEIC テスト リーディング速効ドリル」[12]から抜き出した英文を配布し，学生は質問にこたえ，解答をエクセルファイルに書き出し提出する．この書籍には，TOEIC のリーディング問題を強化するための問題が「シングルパッセージ」「ダブルパッセージ」にわかれ，それぞれに分野別に問題が載っている．ここでは，「シングルパッセージ」の章の「経済問題」を選択した．また，英語の文章と一緒に書籍に載っている，本文中の重要単語(8 単語)とその意味もあわせて配布した．

質問項目は以下の 3 問で，解答時間は 30 分とした．
1) 「英語」が好きか嫌いかを自由に述べよ．
2) 配布した経済に関する英文を読み，次の質問に答えよ．
- 提示される重要単語とその単語の意味をみて，未知の単語があれば列挙しなさい．
- 提示される重要単語以外で，文中に未知の単語があれば列挙しなさい．

1)の問いについては，学生の主観であり，「好きだができない」ので好きか嫌いか選べないなどと悩む学生もいた．そこで自由に記述して貰い「好き」か「嫌い」かを教師が判断し，グループわけした．最終的には空欄の学生がいたため，「英語好き」「英語嫌い」「記述なし」の 3 つのグループを作成した．この 3 つのグループを用い学生が列挙した単語数を表 1，図 1，図 2 にまとめた．

	人数	重要単語内の未習得単語の平均	重要単語以外の未習得単語の平均	合計
英語好き	7	4.4	3.3	7.7
英語嫌い	10	4.9	7.3	12.2
記述なし	3	4.7	4.3	9.0
合計	20	4.7	5.5	10.2

表 1 未習得の単語の平均
Table 1 The average of unknown words in an article

表 1 より，書籍内で重要単語として提示されている単語内では，学生の好き嫌いのグループわけではどれも差がほとんどないことがわかる．このことから，重要単語として提示されている単語が比較的難易度の高い単語であるため，どの学生にも未習得であることが想像できる．これに対し，重要単語以外で学生が未習得であると列挙した単語には好きと嫌いでは倍以上の差があった．これは，英語の学習レベルの違いと推測できる．

次に，図 1，図 2 を分析する．図 1，2 においては，横軸は学生番号であり，学生番号 1～7 までが英語好き，8～17 までが英語嫌い，18～20 が記述なしの学生とした．英語好きの学生たちは，図 2 においていずれも未習得単語数が平均以下であった．しかし，英語嫌いの学生たちは，図 1 において大きくばらつきがあることがわかった．これは，英語の基礎単語力に大きい差が出ていると推測できる．

これらの結果より，書籍が提示する重要単語は学生にとっては比較的難度の高い単語であり，学生が自律学習するためにはそれだけを提示するのでは不十分であることがわかった．また，特に英語嫌いの学生には基礎単語力にばらつきがあるため，どの学生にも支援するためには，個々の学生の単語習得レベルにあわせた支援が必要であることがわかった．そこで，本研究では，単語に焦点をあて，学生の実際の学習レベルにあわせて，個々の学生に対して適した重要単語を提示するシステムを作成することにした．

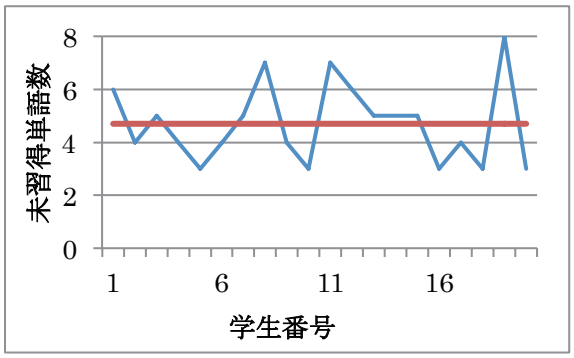

図 1 重要単語内の未習得単語数

Figure 1　The number of unknown words in important word lists.

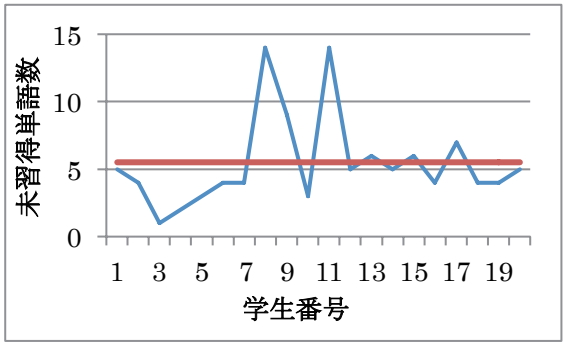

図 2 重要単語以外の未習得単語数

Figure 2　The number of unknown words in an article.

2.2 概要

ここでは，既存のシステムについて概要を説明する．既存のシステムでは，図 3 のように，フィルターと文章を用い検出された単語に対し，必要に応じて辞書をひき単語集を作成している．次に，文章，フィルター，辞書について順に詳しく説明する．

(1) 文章　個人が学習用に用意したものであり，この文章から単語を抜き出し単語の分析を行う．ここで用いられる文章は，分野や言語を問わず，電子テキスト状態であれば他の条件はない．英語の場合，この文章から WordNet[13]を用い単語の原形だけを抜き出し順序情報を削除した単語集合を作成する．一度作成された単語集合は次のフィルターとしても利用できる．

(2) フィルター　(1)で用意された文章から，単語を抜き出すためのフィルターとして用いる．フィルター以下の 3 種類が存在する．

(ア) 標準単語集合　この単語集合を用いて，学術的に学習者のレベルを測定することができる．標準単語集合は多々存在するが，General Service List(GSL)[3,4]，Academic Word List(AWL)[5,6]，JACET8000[14] の3つを導入した．GSLには，英語学習者がまず取得すべき最重要単語約2000語を示す単語が含まれている．AWLは英語圏の大学で使われる様々な分野の教科書・学術論文等で使用される単語を頻度別に分類整理した結果，まとめられた単語リストである．JACET8000 は日本大学教育学会(JACET)が作成した日本の高等教育機関で利用されている最重要単語8000語の単語リストである．GSL，AWL，JACET8000はどれも非常に信頼度の高い英単語リストとして英語教育研究の分野で利用されている[15]ものである．

(イ) 個人用単語集合　学生所有の単語集合であり，プロトタイプシステムを用いて各々学生が用途にあわせてグループ化した単語の集合である．授業で学習した単語をグループ化し，授業で学んだ単語集合を作成したり，記憶済みの単語をグループ化し，自分の語彙を示す単語集合作成したりすることができる．

(ウ) その他の単語集合　(1)で紹介した第二外国語学習者のための標準単語集合と異なり，「経済分野」や「情報科学分野」など分野別や「TOEIC 頻出単語」など用途別の単語の集合である．このシステムを利用し作成することもできる上に，インターネット上で気に入った単語集合をダウンロードし単語集合として利用することもできる．

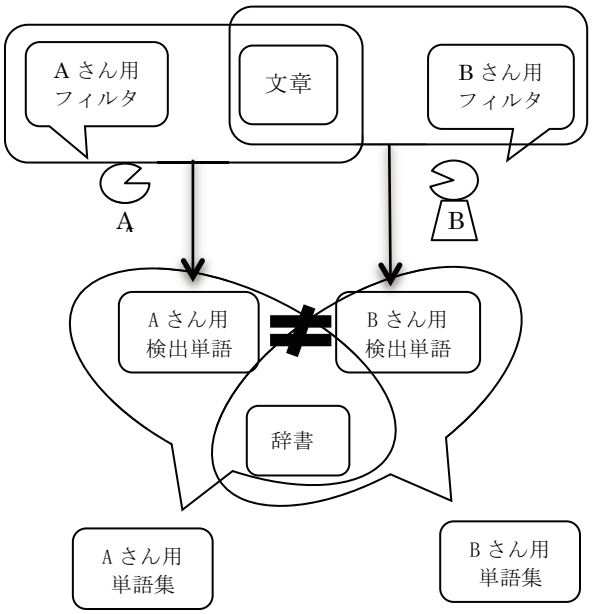

図 3 プロトタイプシステム：文章から各々に適した単語集が作成される例

Figure 3　The previous prototype system: overview

(3) 辞書　辞書としてフリーの gene 辞書[16]と WordNet を用い，学習者がどちらかを選択する．gene 辞書は，NiftyServe の英会話フォーラム(FENG)で公開されていた辞書である．利用の歴史があり現在も多くのシステムで用いられている．単行本・雑誌・新聞・パンフレット･広告などから文書や単語が抜き出されており，生きた英語を学ぶ自律学習用に適しているといえる．WordNet は，Princeton 大学で開発された歴史ある信頼おける辞書である．WordNet は複数の機能を持ち，意味を調べる以外で，本システムでは単語の原形を調べるのにも用いている．

2.3 既存システムの課題

既存のプロトタイプシステムを用いて，利用実験を行った結果，以下，辞書と著作権の 2 つの大きい問題があることがわかった．

(1) 辞書

辞書については,以下 2 つの問題があることがわかった．一つ目は，商用辞書の契約問題であった．語学教員よりフリーではない，信頼と実績ある出版社より発売された辞書の導入を希望されたことにより，商用辞書の導入に向け 3 社と協議した．しかし，いずれも費用や契約上の問題で導入に至る事はなかった．商用辞書は，既に他のオンラインシステムと独占使用契約をしてしまっている場合や，厳しい人数制限をかけている場合などがあり，今後も導入が難しいことがわかった．

二つ目は，語学の学習初心者からの要望で，知らない単語を自動で抽出し辞書をひいてくれるのはいいが，辞書には複数の意味が載っているため，どの意味なのかわかりにくいという指摘があった．単語には分野によって意味が異なる単語などもあり，初心者が学習に利用するにはまだ利用が難しいようであった．そこで，辞書をひき意味を列挙するのではなく，読んでいる文章にあわせた意味の推薦が必要なことがわかった．

以上のことより，商用辞書を利用せず，信頼のある意味をひいたり，文章にあわせた単語の意味を推薦したりする機能が必要であることがわかった．

(2) 著作権問題

2010 年の著作権改訂によって,プロトタイプシステムのサーバ側にインターネット上で公開されている文章を保存することができなくなり，文章から順序情報を抜いた単語と出現頻度のみがサーバ側で保存できることがわかった．また，学習者の個人利用に関しては順序情報付きの文章の利用が認められているため,学習者は自分の USB や HDD，オンラインストレージなどに順序付きの文章を保存して，再利用が可能であることもわかった．そこで，プロトタイプシステムの仕様を一部変更することで解決できることがわかった．

3. 提案するシステム

本研究では，2 章で述べた既存のプロトタイプシステムの辞書の課題を解決するために，学生のレベルや文書の分野などにあわせて単語の意味を推定するシステムを考えた．また，複数の辞書を使い，重み付けを行うことによって，文章にあった信頼できる辞書を推薦する機能を追加する．3.1 章では概要を，3.2 章では辞書に関する変更点について，3.3 章で推薦システムのアイデアについて，3.4 章で実験について述べる．

3.1 概要

ここでは，既存のシステムの改良案の概要を説明する．今回は，2.3 章で述べた様に，商用の辞書を使わずに，できるだけ信頼性を出し，また，読んでいる文書にあわせた単語の意味を自動で抽出できることを目的とし，それ以外の部分は，既存のシステムを使い回すこととする．

新システムでは，図のように，フィルター，文章を用い各々ユーザにあった単語を検出するまでは，既存のシステムと同じである．既存システムでは，学習者が用意されている二つの辞書のどちらを利用するかを選択できたが，新システムでは学習者の選択以外に，辞書の推薦を行う機能を持つ．

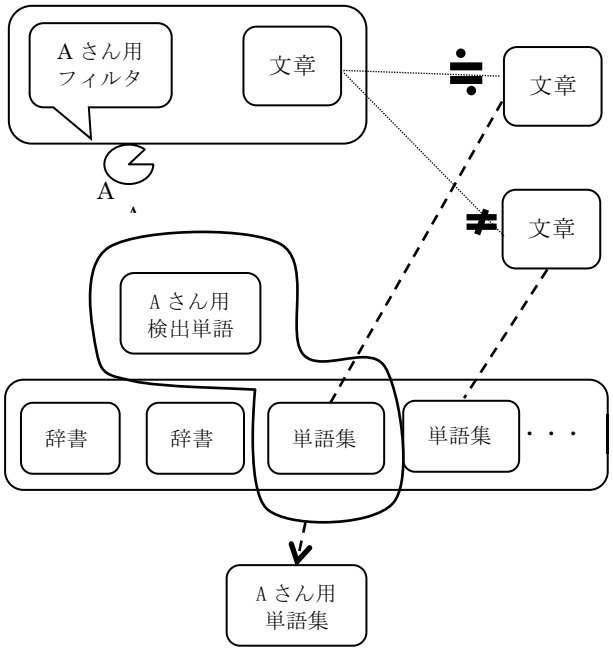

図 4　辞書推薦システムの概要
Figure 4 The proposed system：overview.

新システムでは，作成された単語集を辞書として用いることができる．単語集には，単語一つに対し，一つの意味をペアとして格納されている．また，単語集は作成された元の文書とも対応付けられている．新たに単語集を作成し

ようとしている文章と似ている文章から作成された単語集は，作成する単語集と似ているとみなし，辞書として単語集を推薦する．

3.2 利用できる辞書

ここでは，以下 3 種類の辞書を用いる予定である．

（ア） WordNet，日本語 WordNet
（イ） Gene95
（ウ） 単語集

WordNet，Gene95 辞書は既存システムと同じものを用いている．WordNet，日本語 WordNet, gene95 はそれぞれ特有のフォーマットを持っているが，本研究のシステムでは，単語一つに対し一つの意味を組とし保存する．(ウ)の単語集は本研究のシステムを用いて作成することができるもので，辞書として用いることができる．ユーザは辞書を分類して複数所有することができる．また，単語集は作成された元の文章の単語集合と対応付けされている．

単語一つに対し，意味を一つだけ対応させることによって，英語の文章から日本語の文章を探すよう拡張できる．

3.3 推薦機能

ここでは，辞書の推薦機能について説明する．まず，文書の類似度を測定するためのコサイン尺度について定義し，次に実験結果についてまとめる．

(1) コサイン尺度

文書検索においてよく利用されている，文書と検索質問のベクトル間の類似度を測定するためのコサイン尺度[17]を以下の用に定義する．ベクトル空間において，文書中に存在する全単語の出現頻度を要素とするベクトルで文書を表現する．類似度判定を行う文書を$d_1, d_2, \cdots, d_n$とし，これら文書集合全体を通して全部でm個の単語$w_1, w_2, \cdots, w_m$があるとする．このとき，文書 d_jは次のようなベクトルで表現される．

$$d_j = \begin{bmatrix} d_{1j} \\ d_{2j} \\ \vdots \\ d_{mj} \end{bmatrix}$$

ここで，d_{ij}は文書中の単語 w_iの文書D_j内での出現頻度とする．

検索質問も，文書と同様に出現頻度を要素とするベクトルで表現することができ，検索質問文に含まれる単語 w_iの文書D_j内での出現頻度をq_iとすると，検索質問ベクトルqは次のように表される．

$$q = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_m \end{bmatrix}$$

ここで，文書D_jと検索質問qの類似度をコサイン尺度で表すと次のようになる．

$$\cos(d_j \cdot \mathrm{q}) = \frac{d_j \cdot \mathrm{q}}{\| d_j \| \| \mathrm{q} \|} = \frac{\sum_{j=1}^{m} d_{ij} \cdot q_i}{\sqrt{\sum_{i=1}^{m} d_{ij}^2} \sqrt{\sum_{i=1}^{m} q_i^2}}$$

一般に，0 に近い場合，文書D_jと検索質問qは似ておらず，1 に近づくと，文書D_jと検索質問qは似ているとみなされる．

(2) コサイン類似度の実験

本研究ではコサイン尺度を用い，6 つの文書間のコサイン類似度を求める実験を行った．用意した文章は以下の 6 つである．

- d_1：「ハイスピード実現! TOEIC テストリーディング速効ドリル[12]」より「経済問題」の文章
- d_2：「はじめての理系英語リーディング[18]」の「QWERTY keyboard」の文章
- d_3：「はじめての理系英語リーディング[18]」の「Dvorak keyboard」の文章
- d_4：The economist の「The wait is over」[19] ギリシャのデフォルトの記事の一段落目のみ
- d_5：「Super User」の「Ubuntu におけるギリシャフォンを変更する」文章[20]
- d_6：About.com の「キーボードショートカットやキーをリセットする」文章[21]

6 つの文章の，それぞれの単語数と分野，共通する単語は表 2 の通りである．それぞれ「default」,「Greek/Greece」「keyboard」の単語を含む場合は○をつけた．また，分野に関しては，経済の場合「経」，コンピュータの場合「コ」，経済にもコンピュータの分野にも入る場合「経コ」と略して記した．

文章	単語数	「Default」	「Greek／Greece」	「keyboard」	分野
d_1	54				経
d_2	37			○	経コ
d_3	29			○	経コ
d_4	49	○	○		経
d_5	42	○	○	○	コ
d_6	31	○		○	コ

表 2　6 つの文章と単語数，分野
Table 2 The number of words and areas in 6 articles.

d_2, d_3は同一書籍の「製品の生き残り戦略(Product Survival Strategy)」からそれぞれ抜き出した．d_4, d_5, d_6にはそれぞれ「default」の単語が含まれているが，d_4は「債務不履行」の意味で，d_5, d_6は「初期設定」の意味で利用されている．また，d_4, d_5はギリシャに関する記事だが，d_4は

ギリシャの金融状況に関する記事であり，d_5はギリシャフォントに関する記事である．経済分野の記事はd_1, d_4であり，コンピュータ分野の記事は d_5, d_6，経済分野にもコンピュータ分野にも含まれる記事は d_2, d_3である．

これらのコサイン類似度は表 3 のとおりである．類似度の最大は$\cos(d_2, d_3) = 0.367$であった．これは，同じ書籍の同じ章の同じテーマの文章であり，共通のキーワードも存在するため，予想通りの結果となった．次に高いのは，$\cos(d_3, d_5) = 0.115$と$\cos(d_3, d_6) = 0.190$であった．これはd_3, d_5, d_6ともに同じ分野であり，キーボードに関するものであったため，高くなったと思われる．

また，d_4は，d_2, d_3, d_5, d_6との距離がいずれも 0 であった．これは，d_4はd_1, d_2, d_3とは分野が同じではあるが，内容が違うことに加え，d_1, d_2, d_3は初心者用テキストからの抜粋で比較的難易度の低い単語が使用されているのに比べ，d_4はエコノミストのオンライン記事からの抜粋のため，比較的難易度の高い英単語が使われていることが理由にあげられる．

この実験結果により，簡単なコサイン尺度を調べるだけでも，文章間の類似度が推測できることがわかった．

	d_2	d_3	d_4	d_5	d_6
d_1	0.010	0.027	0.011	0	0.015
d_2		0.367	0	0.050	0.039
d_3			0	0.115	0.190
d_4				0	0
d_5					0.067

表 3　文章間のコサイン類似度

Table3　The cosine similarity of 6 articles.

3.4 今後の課題

3.3 章で説明したコサイン類似度を用いた拡張について説明する．今回は，文書と文書を比較し，コサイン類似度を求めた．しかし，本研究のシステムでは，辞書を作成し複数持つことができるため，関係のある辞書を複数推薦した上で，重みづけを行い，より文章にあった意味を推薦できるように改良する予定である．これにより，学生が間違って作成した辞書などは利用するうちに重みが低くなり利用されなくなるはずである．こうして，最終的には信頼ある辞書のみが利用されるようになるはずである．

また，辞書の書式を単語と意味の組に変更したおかげで，英単語から日本語の単語，日本語の単語から英単語への変換が簡単になり，日本語の文章と英語の文章の類似度も調べることが可能になった．今後は，辞書の推薦だけでなく，分野や言語を問わない文章の推薦も行っていきたい．

4. おわりに

本研究では，単語に焦点をあてた自律学習を支援するシステムの構築を行っている．本論文では，既存のプロトタイプシステムの問題点のうちの辞書に関する問題を解決するために，新たな辞書推薦システムの提案と実験を行った．今後は，辞書や教材となり得る文章の推薦を行う予定である。

謝辞　本研究は，獨協大学研究奨励費と獨協大学情報学研究所の助成を受けている．

参考文献

1)　特定非営利活動法人日本イーラーニングコンソシアム：“e ラーニング白書 2008/2009 年版”，東京電機大学出版局，2008
2)　e ラーニング専門家のためのインストラクショナルデザイン，玉木欽也監修，齋藤裕他 5 名著，東京電機大学出版局，2006
3)　J. Bauman, and B. Culligan, About the General Service List, http://jbauman.com/gsl.html, 1995
4)　A General Service List of English Words, West, M., Longman, 1953
5)　A. Coxhead, A new academic word list, TESOL Quarterly, 34, pp.213-238, 2000
6)　The Academic Word List, http://www.victoria.ac.nz/lals/resources/academicwordlist/
7)　堀江郁美，飯島優雅，”学生の自律学習を支援する適応型単語リスト作成ツールの開発”，獨協大学情報科学研究，第 27 号，p59-66，2010
8)　Ikumi HORIE, Kenji KASHIWABARA, Kazunori YAMAGUCHI, Yuka IIJIMA, "Personalized Teaching Material Generator Based on Word Set," Information Technology Based Higher Education and Training (ITHET), 2010, pp. 343-348.
9)　ライフサイエンス辞書プロジェクト, http://lsd.pharm.kyoto-u.ac.jp/ja/index.html
10)　Excite 翻訳，http://www.excite.co.jp/world/
11)　Yahoo 翻訳，http://honyaku.yahoo.co.jp/
12)　細井 京子，山本 千鶴，“ハイスピード実現! TOEIC テストリーディング速効ドリル”，コスモピア，2006
13)　WordNet at Princeton University, http://www.wordnet.prinston.edu
14)　The Japan Association of College English Teachers, JACET List of 8000，Basic Words(JACET8000)[in Japanese], JACET, 2003
15)　“A Word List Generator Program for Using Authentic Texts in an Academic English Reading Class”, Iijima, Yuka, Horie, Ikumi, Information Technology Based Higher Education and Training, p. 407-412, 2010
16)　gene95 辞書, http://www.namazu.org/~tsuchiya/sdic/data/gene.html
17)　北研二，津田和彦，獅々堀正幹，”情報検索アルゴリズム”，共立出版，2002
18)　佐藤洋一，”理系たまごシリーズ（2）はじめての理系英語リーディング “，アルク，2007
19)　The economist, “The wait is over”, http://www.economist.com/node/21550271
20)　Super user, “In Ubuntu, how do I change my default Greek font? “, http://superuser.com/questions/71350/in-ubuntu-how-do-i-change-my-default-greek-font
21)　All About, “Resetting Keyboard Shortcuts and Keys”, http://wordprocessing.about.com/cs/quicktips/qt/Resetkey.htm