日本語 WWW 情報を用いた COCET3300

英単語学習支援に関する研究

情報・知能工学専攻

093750　渡邉　雄大

指導教員:河合 和久

平成25年1月18日

**平成24年度修士論文目次**

[第1章　序論 1](#_Toc346146195)

[第2章　COCET3300について 2](#_Toc346146196)

[第3章　研究内容 3](#_Toc346146197)

[3.1 先行研究 3](#_Toc346146198)

[3.2 システムの概要 4](#_Toc346146199)

[3.3 使用環境 5](#_Toc346146200)

[3.4 システムの実現 5](#_Toc346146201)

[3.4.1 システムの処理手順 5](#_Toc346146202)

[3.4.2 日本語文字列の抽出 6](#_Toc346146203)

[3.4.3 英単語と日本語文字列の登録 6](#_Toc346146204)

[3.4.4 類語の登録 7](#_Toc346146205)

[3.4.5 英単語の表示 7](#_Toc346146206)

[3.5 実行例 8](#_Toc346146207)

[3.5.1 ハイライトおよびポップアップ機能の実行例 8](#_Toc346146208)

[3.5.2 英単語最終表示ページへのジャンプ機能 9](#_Toc346146209)

[3.5.3 確認テスト機能 11](#_Toc346146210)

[3.5.4 COCET3300英単語リストの表示機能 13](#_Toc346146211)

[3.5.5 設定画面 15](#_Toc346146212)

[第4章　評価と検討 17](#_Toc346146213)

[4.1 評価内容 17](#_Toc346146214)

[4.2 検証結果 18](#_Toc346146215)

[4.3 考察 24](#_Toc346146216)

[第５章　まとめ 25](#_Toc346146217)

[参考文献 26](#_Toc346146218)

[謝辞 27](#_Toc346146219)

[付録A　プログラム・ソースリスト 28](#_Toc346146220)

[付録B　評価データ 28](#_Toc346146221)

# 第1章　序論

高等専門学校は中学校を卒業した生徒を受け入れ，五年間の一貫教育によって専門的知識，技術を習得させることを目的とした高等教育機関である．しかし一方で，高専生は高校から大学へ進学した学生と比べて，「高専生は英語力が低い」という通説が生まれている．[1][2]

　また，英単語学習において重要な点は継続である．継続して英単語に触れることによって，英単語を定着させることができる．しかし，学習には継続が必要だが，意識して学習を続けることは英語に苦手意識を持つ者にとって非常に難しい．もし英単語学習を意識せずに日常的に英単語に触れられたら，例え英語に苦手意識があっても英語能力は伸びるはずである．

　現在，多くの人がインターネットを活用し様々な情報を得ている．その間に，意識せずに英単語学習が行えたら，苦痛を感じることなく英語能力を伸ばすことができるはずである．

そこで本研究では日常的に使用するインターネットブラウザに着目し，英単語の学習支援を行うアドオンの開発に取り組んだ．

本研究では，本研究室の石川が開発したCOCET3300英単語学習システム[3]をもとに日本語WWW情報を用いた英単語学習支援システムの開発に取り組んだ．

# 第2章　COCET3300について

COCET3300とは高専生が習得すべき3300語からなる英単語リストと，それをテストするためのWebアプリケーションからなる．この語彙リストとWebアプリケーションはCOCETの会員である全国の高専の英語教員14名からなる研究グループにより編纂された高専に特化した英語教育の教材である．[4]

　COCET3300には，工学系の分野でよく用いられ高専生にとって必須とされる単語が収録されており，収録されている単語には意味だけでなく，高専生の興味に合わせた解説や主に専門教科で役立つと思われる関連語句がつけられている．またこのリストは，使用頻度の高い単語から順に収録されており，収録順に通し番号が振られている．COCET3300の内容の一部を図１に示す．

****

図1　COCET3300の一部

またCOCET3300のWebアプリケーションはメディア教育開発センターの主導のもと開発され，現在「理工系学生のための必修単語COCET3300」として，供用されている．図2にWebアプリケーションの実行画面を示す．このWebアプリケーションは，COCET3300英単語を各ユニット20単語ずつ，計165ユニットに分け各ユニットごとに四択和訳問題，リスニング問題，スペリング問題がある．それぞれのユニットの問題は一定の得点以上でクリアとなり，問題形式によって点数を得られる．その点数によってこのWebアプリケーションの利用者の中でのランキングが表示される．また，一定数ユニットをクリアすると過去にクリアしたユニットの中からランダムに過去問が出題され，過去問を間違うと，そのユニットのクリアが取り消される．

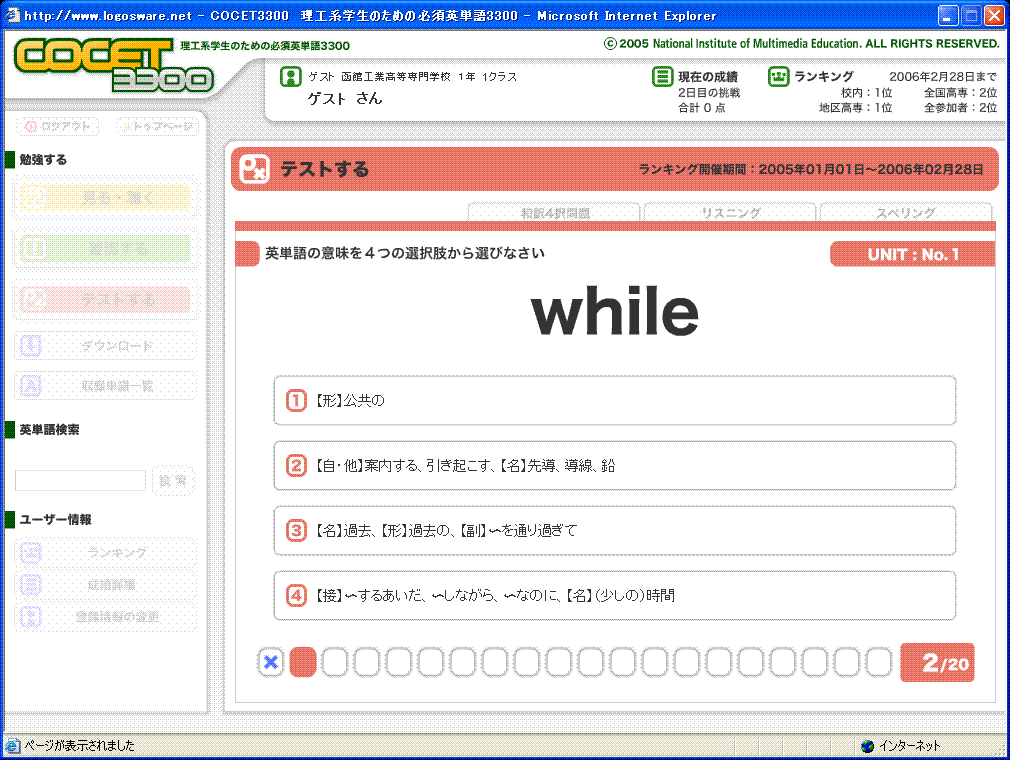


図2　COCET3300のWebアプリケーション

# 第3章　研究内容

## 3.1 先行研究

以下に前年度，本研究室所属石川が開発した英単語学習支援システムの概要を述べる．

このシステムはFirefoxブラウザ(http://www.mozilla.jp/firefox/)の拡張機能(アドオン)としてシステムが作られている．ブラウザであるFirefoxがサーバからWebページのソースコードであるHTMLデータを受けとった際，HTMLのテキスト部分を抽出し，英単語のハイライトを行う．

このシステムの特徴はHTMLデータ内の英単語部分を抽出し，該当する英単語をハイライト表示する．さらにハイライトした部分にマウスカーソルを乗せると日本語の意味がポップアップされる．図3に実行例を示す．

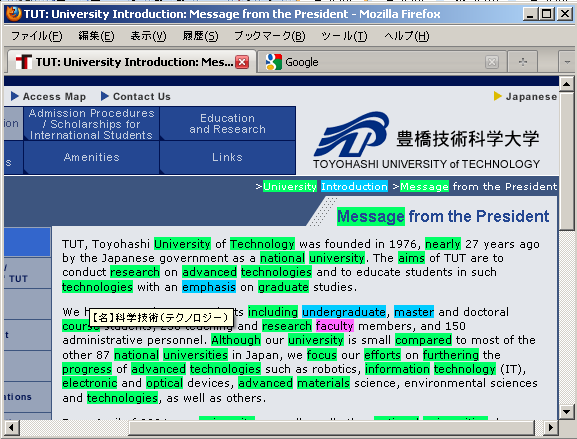


図3　石川学習支援システム実行例

　また，本システムを用いたCOCET3300英単語の出現頻度に関する研究[5]を本研究室の林が行った．この研究では，日本語WebページにおけるCOCET3300の英単語の出現頻度やWebページによって出現する英単語の偏りを調べ，COCET3300の英単語学習に日本語Webページがどれだけ役に立つか調べた．

## 3.2 システムの概要

本研究は前述した背景と先行研究をもとに，英単語学習支援用アドオンを開発した．まずはアドオンの機能について述べる．

日常的に英単語に触れることを大きな目的とするため，学習を意識せずとも英単語に目を触れさせなければならない．よって本研究では，Webページ上の日本語文字上にマウスカーソルを移動させることによってポップアップで英単語を表示させることにした．日本語文章ならば，英語の得意不得意に関わらず読むことができるため，英単語学習に苦手意識を持つ学生でも英単語に触れる機会を増やせると考えたためである．

また，ポップアップに類語を表示することにより，より多くの英単語に触れる機会を増やし，効率よく学習できるようにした．実際類語を追加したところCOCET3300英単語の日本語訳１つに対し平均で約1.9個の類語，英単語１つに対し約6.3個の類語が追加される．それにより，日本語訳1つに対し1つの英単語が対応すると考えると，もともとの日本語訳に対応する英単語に加え，類語によって表示される英単語が加わり約3個の英単語が表示されることとなる．よってもともと１つしか表示されなかった英単語が，3つ表示されることとなる．英単語の日本語訳の重複やWebページによる表示される日本語の偏り，日本語訳の持つ類語の偏りを考えなければ，単純に約3倍表示される英単語が増えると考えられる．

他にも，これまで表示された英単語の回数表示，学習の確認のためのテスト機能，英単語が最後に表示されたWebページの履歴，COCET3300の英単語リストを追加し，より英単語学習支援システムとして機能を充実させた．

## 3.3 使用環境

本研究では，Webページ閲覧に使用するブラウザをFirefoxとし，本システムをアドオンによる拡張機能として実現した．これはFirefoxがアドオンによる機能拡張のできるブラウザであり，またFirefoxの利用者の多くはアドオンを利用することが多いため，本学習支援システムを利用しやすいと考えたためである．

## 3.4 システムの実現

### 3.4.1 システムの処理手順

　図4に本システムの処理手順を示す．本システムはjavascriptとXULにより構成されている．

|  |
| --- |
|  |

図4　システムの処理手順

### 3.4.2 日本語文字列の抽出

　日本語文字列の抽出はHTMLデータを読み込み，対象とする日本語文字列の抽出処理について記述する．対象となる日本語文字列はプログラム内に英単語とセットで配列として書かれており，HTMLデータの中から該当する文字列の探索を行う．

　まず，HTMLデータから日本語文字列を単語に区切る処理を行う．これは分かち書きソフトウェアTinySegmenter[6]を利用している．TinySegmenterは辞書を使わない分かち書きシステムであり，文章から学習し，分かち書きを行っている．そのため，文の長さや内容によって分かち書き結果が変わることがある．実際に文章を区切った例を図5に示す．

|  |
| --- |
| ネットワークやマルチメディアに象徴される高度情報化社会を支える  ネットワーク | や | マルチメディア | に | 象徴 | さ | れる | 高度 | 情報 | 化 | 社会 | を | 支える |

図5　日本語分かち書き例

　このように区切った文字列と英単語リストの文字列を比較し，一致した場合はハイライト表示するべき文字列として配列に入れられる．また処理速度の問題と分かち書きが文章の長さや内容によって変化するため動詞の語尾変化と複合語の処理は行わない．

### 3.4.3 英単語と日本語文字列の登録

　英単語の登録はすべて開発者が手作業でCOCET3300の日本語訳から選出し行っている．配列を英単語の数だけ用意し，その中にその英単語に関する情報を配列として登録する．図6に英単語「system」の記述例を示す．図6の例の場合，第一要素はCOCET3300における登録番号，第二要素はポップアップで表示する英単語，第三，四，五要素は実際にwebページ上の日本語文字列から探索する文字列となっている．これにそれぞれの日本語文字列に対する類語を追加し，ポップアップを表示する．

|  |
| --- |
| cocet3300[0] = new Array(  "1",  "system",  "システム",  "仕組み",  "体系",  " ",  "2",  "2",  "2",  "　",  "日本語で「○○システム」や「○○系」という場合には、おおむねこの語が使われると考えてよい。"); |

図6　英単語記述例

### 3.4.4 類語の登録

　本システムでは，COCET3300の英単語の日本語訳に対して類語の登録を行っている．

類語辞書には日本語WordNet[7][8]を用いた．日本語WordNetとは独立行政法人情報通信研究機構が開発している日本語の概念辞書である．個々の概念はそれぞれ「synset」という単位にまとめられており、それらが他のsynsetと意味的に結びついている．無料で利用できる点，元々が英語の概念辞書であったため英語との関連性が強い点，Webページ上で類語検索と表示ができ，本システムに組み込む際にCOCET3300の英単語の日本語訳に対応する類語が探しやすい点を考慮し本システムでは日本語WordNetを類語辞書として選んだ．

　類語の登録は関数set\_thesaurus内で行い，表示される類語を日本語訳と英単語からなる配列に加えることで行っている．

|  |
| --- |
| function set\_thesaurus(){  for(j=temp3300.length-1;j>0;j--){  if(-1 !== str.indexOf(" "+ "形" +" ")){  temp3300[j][4] = temp3300[j][4] + searchThesaurus("型") + searchThesaurus("形式") + searchThesaurus("像") + searchThesaurus("形態");  }  ．．．  }  } |

図7　類語の登録例

### 3.4.5 英単語の表示

前述の日本語文字列の抽出を行った後，該当する日本語文字列にハイライト表示を行う．さらにマウスカーソルがそのハイライト表示の上に乗った時に英単語のポップアップ表示を行う．ハイライト表示する文字列の部分に，ハイライトとポップアップ表示のためのタグを，HTML内の単語の両端に書き加える．

図8では，「仕組み」という日本語文字列に対し行うタグ付けを示す．

|  |
| --- |
| <span title="【 仕組み に対応する英単語】 mechanism：機械 仕組み メカニズム 　 system：システム 仕組み 体系 【類語】 プロセス：process 機構：organization 構造：construction,configuration,structure 原理：principle 環境：environment,setting 枠組み：frame 構想：concept,plot " style="background-color: #FF9966;border:gray solid 1px;" +="" onclick="alert(仕組み);">  <rb>仕組み</rb>  </span> |

図8　spanタグでのtitleの設定

これにより，その文字列の部分にマウスカーソルを載せることで，ポップアップメッセージによってその日本語文字列に対応する英単語を表示させた．

## 3.5 実行例

### 3.5.1 ハイライトおよびポップアップ機能の実行例

　図9，10にハイライトおよびポップアップ機能の実行例を示す．本システムを実行するには，まず英単語を表示させたいwebページを開き，右クリックメニューを開く．表示されたメニューの中の「英単語表示on」という項目をクリックすると，処理が実行されてページ内の英単語に対応する日本語文字列がハイライト表示される．



図9　右クリック実行例



図10　ハイライトおよびポップアップ実行例

### 3.5.2 英単語最終表示ページへのジャンプ機能

　右クリックメニューから「画面下表示切り替え」を選択するとウィンドウの下部に英単語最終表示ページへのジャンプ機能などの機能が実行できるボタンが表示される．

ジャンプ機能では上記のハイライト機能を実行すると，Webページ内にポップアップされる英単語を英単語ごとに記録し，最後に英単語が表示されたWebページを開くことができる．ウィンドウ下にあるメニューから英単語を選択し，最終ページを新しいタブで表示する．図11，12に実行例を示す．実行例では，「isotope」を選択し，最後に「isotope」が表示されたページを表示している．

### 

図11　英単語の選択

### 

図12　最終表示ページへジャンプ

### 3.5.3 確認テスト機能

　図13，14に確認テスト機能の実行例を示す．確認テストはウィンドウ下部から行え，Webページ内で出現するCOCET3300英単語の確認を目的としている．確認テストは，そのページ内でハイライト表示を実行した後，そのWebページ内でポップアップ表示される英単語の中からランダムで出題する．問題は英単語記述式で，回答により正解と不正解の画面が表示される．英単語ごとに出題数と正解数が記録され後述するCOCET3300英単語リストの表示機能で正答率を確認することができる．



図13　テスト回答画面



図14　答えの表示

### 3.5.4 COCET3300英単語リストの表示機能

　COCET3300英単語リストの表示機能の実行例を図15，16に示す．この機能では，COCET3300の英単語について英単語の意味や簡単な解説を表示し，より深く英単語学習をしたい利用者を支援するために作成した機能である．テキストボックスから英単語または日本語文字列で検索ができ，表示される英単語と日本語訳からリンクもされている．図15は「system」で検索した結果を表示している．

　日本語文字列で検索するか，表示された日本語訳をクリックすることで新たな英単語が表示される．また，日本語での検索や日本語からのリンクでは，その日本語に応じた類語も表示される．この類語にもリンクが張られていて，クリックすることで新たな英単語が表示される．図16は図17の日本語訳の「仕組み」をクリックした際の実行結果である．英単語一覧には「仕組み」が日本語訳として登録されている英単語一覧が表示され，類語には「仕組み」に対する類語が表示される．

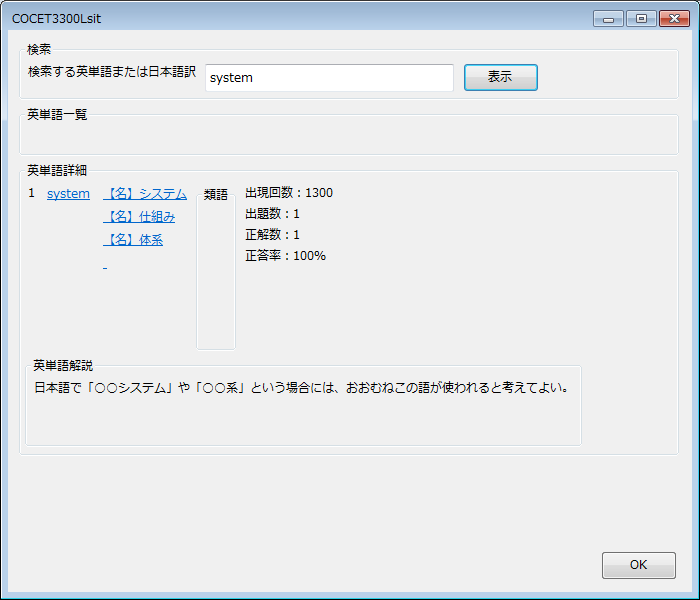


図15　「system」の検索結果

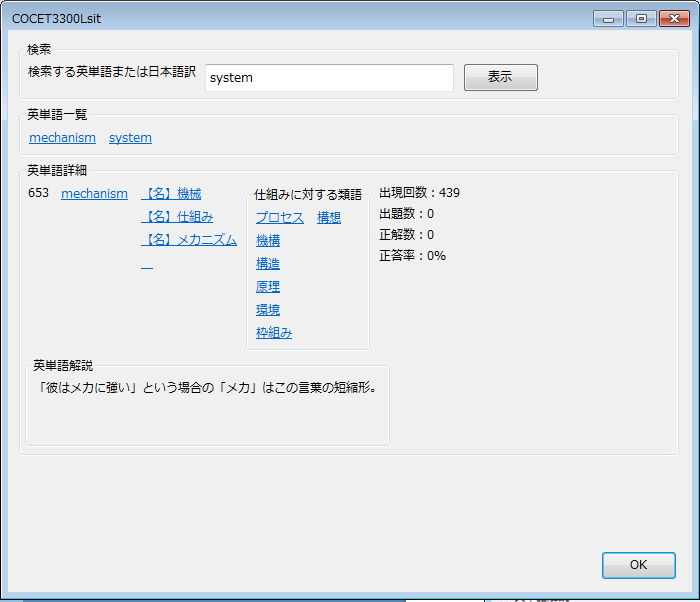


図16　「仕組み」をクリックした際の画面

### 3.5.5 設定画面

　図17，18，19に設定画面を示す．利用者が学習しやすいよう機能をカスタマイズすることで，利用者にあった学習を行うことができ，継続的な学習に取り組みやすくするために作成した機能である．図17は英単語の色分けに関する設定画面，図18はこれまで表示された英単語の回数一覧，図19はその他の設定画面である．

　英単語一覧にはポップアップされた日本語に対応する英単語と類語に対応する英単語がすべてカウントされる．



図17　ハイライト色設定画面

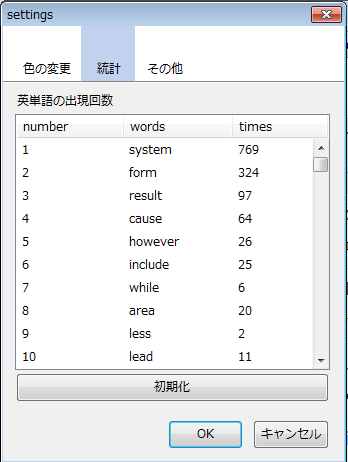


図18　英単語表示回数一覧

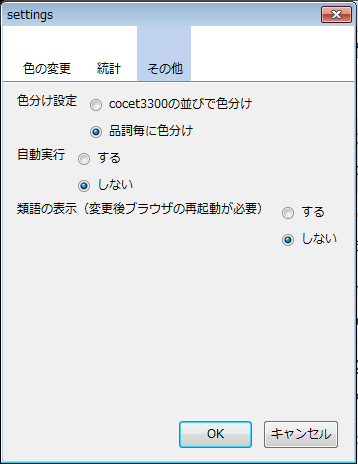


図19　その他の設定画面

# 第4章　評価と検討

## 4.1 評価内容

　類語をCOCET3300に組み込むことで英単語学習にどのような効果が期待できるか検証を行う．検証するシステムは，類語を表示するシステム（システムA）と，類語を表示しないシステム（システムB），前年度本研究室の林が検証を行ったCOCET3300の英単語が900番目まで登録されているシステム（システムC）で，COCET3300の単語リストの追加と類語の追加により英単語出現頻度を検証する．Webページ内で本システムを実行し，単語リストや類語によってどれだけ英単語の出現度と出現する英単語の種類が変わるか検証を行う．

検証に利用するWebページは「初心者のためのポイント学習Ｃ言語（トップページとメインコンテンツ）」「豊橋技術科学大学（大学概要）」「日本生命保険相互会社（保険の基礎知識）」である．専門的な語句が多いWebページと一般的な語句と専門的な語句が両方あるWebページ，工学系の語句が少ないWebページとで英単語の出現頻度を検証する．

## 4.2 検証結果

　以下に各システムの総単語に対するCOCET3300の英単語の出現頻度と総出現英単語数，英単語の種類をまとめる．なおグラフの各項目の実数値は付録に掲載する．

表2　「初心者のためのポイント学習Ｃ言語」における出現頻度

|  |  |  |  |
| --- | --- | --- | --- |
|  | 総出現COCT3300  英単語数 | WebページにおけるCOCET3300単語の割合 | 英単語種類 |
| システムA | 22786 | 12.64％ | 934 |
| システムB | 5563 | 12.64％ | 457 |
| システムC | 2747 | 7.67％ | 227 |

図22　「初心者のためのポイント学習Ｃ言語」におけるシステムAの出現英単語

図23　「初心者のためのポイント学習Ｃ言語」におけるシステムBの出現英単語

図24　「初心者のためのポイント学習Ｃ言語」におけるシステムCの出現英単語

表2　「豊橋技術科学大学」における出現頻度

|  |  |  |  |
| --- | --- | --- | --- |
|  | 総出現COCT3300  英単語数 | WebページにおけるCOCET3300単語の割合 | 英単語種類 |
| システムA | 15686 | 20.49％ | 731 |
| システムB | 3184 | 20.49％ | 348 |
| システムC | 1995 | 16.55％ | 189 |

図25　「豊橋技術科学大学」におけるシステムAの出現英単語

図26　「豊橋技術科学大学」におけるシステムBの出現英単語

図26　「豊橋技術科学大学」におけるシステムCの出現英単語

表3　「日本生命保険相互会社」における出現頻度

|  |  |  |  |
| --- | --- | --- | --- |
|  | 総出現COCT3300  英単語数 | WebページにおけるCOCET3300単語の割合 | 英単語種類 |
| システムA | 8232 | 18.73％ | 688 |
| システムB | 2152 | 18.73％ | 304 |
| システムC | 1125 | 9.97％ | 167 |

図27　「日本生命保険相互会社」におけるシステムAの出現英単語

図28　「日本生命保険相互会社」におけるシステムBの出現英単語

図29　「日本生命保険相互会社」におけるシステムCの出現英単語

## 4.3 考察

　まずシステムBとシステムCを比較すると，どのWebページにおいても英単語の出現回数と種類がシステムBのほうが多いのがわかる．しかし登録英単語においてシステムBはシステムCより2400語も多いにも関わらず，登録されている英単語数の差ほど出現単語数と種類の差は開かなかった．これはWebページによる出現単語の偏りとCOCET3300の英単語が高専生にとって重要となる英単語の順に並んでいるためと考えられる．

　次にシステムAとシステムBを比較すると，出現単語数が初心者のためのポイント学習Ｃ言語のページでは約4.1倍，豊橋技術科学大学のページでは約4.93倍，日本生命保険相互会社のページでは約3.82倍となっている．また出現する単語の種類も初心者のためのポイント学習Ｃ言語のページでは約2.04倍，豊橋技術科学大学のページでは約2.1倍，日本生命保険相互会社のページでは約2.26倍となっている．グラフからもCOCET3300の後半部分の英単語がシステムAのほうがシステムBより多く出現していることがわかる．

　1つの日本語訳につき平均で類語が約1.9個付けられていることから，3.2で述べた出現英単語数の増加の単純な予測値は約3倍であった．しかし，実際は1つの日本語訳に複数の英単語が対応しているもの，日本語訳によっては10個以上の類語が追加されているものもある．そういった理由から予測値を上回る結果が出たと考えられる．

　また，総単語数を英単語の種類で割ると，1つの英単語が出てきた平均の回数となる．初心者のためのポイント学習Ｃ言語のページではシステムAが約24.40回，システムBが約12.17回，システムCが約12.10回となる．豊橋技術科学大学のページではシステムAが約21.46回，システムBが約9.15回，システムCが約10.56回となる．日本生命保険相互会社のページではシステムAが約11.97回，システムBが約7.08回，システムCが約6.74回となる．このことからシステムAはシステムBやシステムCより同じ英単語が何度も出現しているということがいえる．同じ英単語が多く出現するため，英単語の定着効果が期待できる．

　システムAとシステムBの総出現英単語の差は，類語によって表示される英単語の差である．類語で表示される英単語の中で同一コンテンツ内の日本語も表示されている割合を考える．初心者のためのポイント学習Ｃ言語のページでは，類語で表示される英単語が716個，その中の239個がWebページ内の日本語文からも表示されていた．457個ある日本語文から表示される英単語のうち，239個は類語でも表示されることとなる．つまり日本語から表示された英単語のうち，約52.30％の単語が類語として出現することになる．同じように豊橋技術科学大学のページでは約66.95％，日本生命保険相互会社のページでは約48.68％の割合で出現する．つまり日本語から直接表示された英単語のうち約半数が類語として別の場所で出現することとなる．このことから，日本語から直接表示された英単語が再び別の場所で類語として目に触れる機会が増え，継続して学習する中で繰り返し学習になり，より英単語を定着させやすくなると考えられる．

　以上のことから，COCET3300の英単語を類語で関連付けることによって，継続して学習するうえで有用な効果が得られると考えられる．

# 第５章　まとめ

　本研究では工業高等専門学校生向けのCOCET3300英単語学習支援システムの開発，評価を行った．Webページの日本語文章に注目し，閲覧中に自然と英単語が目に入ることにより，日常的に多くの英単語に触れることができるシステムを開発した．また，確認テストやCOCET3300英単語の辞書など学習に必要な機能を実装することにより，学習者に便利なシステムとなった．さらに，類語を用いることにより，より多くの英単語が繰り返し出現し，継続学習するうえで有用なシステムとなった．

本システムは全国高等専門学校英語教育学会(COCET)が開発したCOCET3300のオンライン教材などの英語学習と併用し，英単語学習を支援するシステムとして利用されることを想定している．COCETのオンライン教材などに苦手意識を持った学生でも，Web閲覧中に英単語が学習できるシステムの開発という本研究の目的は達成できたと考えられる．

今後の課題・展望としては，テスト機能の充実や学生に使用してもらっての検証，タブレット端末などへの応用が挙げられる．

本研究では工業高等専門学校生向けのためCOCET3300の英単語のみの表示であったが，利用者の対象を変えてもこのシステムは応用できるため，英語学習全般に応用できればさらによいシステムになるだろうと考える．

# 参考文献

[1] 亀山 太一，全国高等専門学校英語教育学会(COCET)について, 日本高専学会誌 : journal of the Japan Association for College of Technology 15(2), 3-4, 2010-07-30

[2] 大谷　浩，高専とTOEIC, 文部科学教育通信 シリーズ・高専における英語教育のいま，<http://cocet.org/monkyotusin.htm>

[3] 石川 哲郎，Web環境を用いたCOCET3300英単語学習支援システムの開発，平成21年度豊橋技術科学大学卒業論文発表会要旨集,p.24,2009

[4]独立行政法人メディア教育開発センター：理工系学生のための必修英単語3300，<https://cocet.code.ouj.ac.jp/>

[5] 林 佑樹，日本語WWW情報におけるCOCET3300英単語の出現頻度に関する研究，平成23年度豊橋技術科学大学卒業論文発表会要旨集,p.45,2011

[6]工藤　拓：TinySegmenter，<http://chasen.org/~taku/software/TinySegmenter/>

[7]独立行政法人情報通信研究機構：日本語WordNet，http://nlpwww.nict.go.jp/wn-ja/

[8] Francis Bond, Timothy Baldwin, Richard Fothergill and Kiyotaka Uchimoto (2012),[Japanese SemCor: A Sense-tagged Corpus of Japanese](http://nlpwww.nict.go.jp/wn-ja/pubs/2012-gwc-jsemcor.pdf) in The 6th International Conference of the Global WordNet Association (GWC-2012), Matsue. Kow Kuroda, Francis Bond and Kentaro Torisawa (2010)

# 謝辞

本論文の執筆、及び、研究を進めるにあたり、親切にご指導頂きました本学情報・知能工学系の河合和久先生、副査をしてくださった情報・知能工学系の梅村恭司先生に深く感謝いたします。また、多くのご指導を頂きました本研究室の皆様に感謝いたします。

# 付録A　プログラム・ソースリスト

|  |
| --- |
|  |

# 付録B　評価データ