英語語彙指導のためのツール作り

金沢大学　外国語教育研究センター　西嶋　愉一
yuichi@ge.kanazawa-u.ac.jp

1. はじめに

金沢大学では、平成 18 年度からの実施に向けて、全学的に教養教育カリキュラムの刷新に取り組んでいる。その中で、英語については、レベルごとに段階を分けた授業を設定する、という考え方が柱となっている。

授業をレベル分けするための基準としては、使用する語彙の数を用いる。例えば初級は 3000 語レベル、中級は 5000 語レベル、といった形である。あらかじめ設定した語彙リストに従って、3000 語なら 3000 語の範囲で授業を行う。3000 語からはみ出ている部分については別途解説する、試験問題ではその部分に注をつける等の形で対応することになる。

こういった授業を行うためには、使用する教科書や、作成した試験問題などが、その授業で設定したレベルに合致しているかどうかが重要である。そのような作業を支援するためのツールとして、テキストのレベル診断を行うものを昨年の PC カンファレンスで発表した[1]。これは教科書等の素材を選ぶための判断の目安にはなるが、授業支援のための道具としては、任意のテキストに対し、設定した語彙レベルにどの程度おさまっているか、逸脱しているのは具体的にどの単語か、といった情報をわかりやすく提示するものが求められる。今回はそのようなツールを目指し、試験的にプロトタイプを作成した。

2. ツールの構成

入力となる英語のテキストは、あらかじめ OCR 等の手段を用いてデジタルデータ化しておく。今回作成したツールでは、対象はプレーンテキストのみである。

入力されたテキストを TreeTagger[2]で処理し、辞書引きと品詞タグ付けを行う。これにより、テキスト中に出現する単語が見出し語の形になるので、あとはレベル分けした語彙表とマッチングさせればよい。

語彙表は暫定的に株式会社アルクの標準語彙水準 12000(Standard Vocabulary Level, SVL)[3]を使用した。SVL は日本人の学習者を対象とした 12000 語の語彙レベル表で、1 レベル 1000 語ずつ、12 レベルで 12000 語をカバーしている。将来的にはオリジナルの語彙表を作成することも考慮しているが、今回は日本人の学習者にとってレベル分けが適切な SVL を使用することにした。

TreeTagger の出力には品詞タグが含まれ、また、SVL についても[4]には品詞が記載されているため、品詞まで含めたマッチングも可能と考えられるが、品詞つき SVL のデジタルデータがないため、今回は字面だけでマッチングすることにした。

SVL とマッチングした結果について､SVL のレベルごとのパーセンテージ(単語の出現頻度を考慮した場合と考慮しない場合それぞれについて計算する)を計算させる。さらに、表示用に HTML タグを付加し、1) SVL のレベルごとに該当する単語を強調表示する、2) SVL のあるレベルまでに含まれない単語、例えばレベル 1 から 3 の 3000 語から逸脱する語を強調表示する、3) 個々の単語にマウスカーソルを置いてクリックすると、SVL のレベルをポップアップ表示する、といった機能を持たせている。

ツールと便宜上呼んできたが、今回のツールは試験的に作成したもので、単独の完結したソフトウェアとして動作するわけではなく、複数のソフトウェアの組み合わせである。TreeTagger は Solaris 用と Linux 用(Windows 用はデモ版のみ存在する)のバイナリで提供されているため、今回は Solaris 上で動作させている。TreeTagger の出力を SVL とマッチングし、その後、画面表示用に HTML タグを付加する部分については、同じく Solaris 上の awk で処理させている。

現在はテキスト入力用の Web インターフェイス等を持たないので、ツールは Solaris のシェル上で実行した上、出力した HTML ファイルを Windows の動作する PC にダウンロードして表示させる形を取っている。

3. ツールの出力

英語学習者向け雑誌 English Zone の記事を OCR でデジタル化したものについて、本ツールを適用した結果の例が図 1 である。1～12 のボタンは SVL の各レベルに対応している。「指定したレベルのみを

図 1. ツールの HTML 出力を表示

表示」を選ぶと、クリックされたボタンに応じて、そのレベルの単語のみを強調表示する(それ以外の単語は薄く表示する)。「指定したレベルからの逸脱を表示」を選んだ場合は、例えばレベル 3 の 3000 語に含まれない単語を強調表示する。さらに、表示されている単語(図 1 では”assignments”を指している)をクリックすると、その単語のレベルをポップアップ表示する。

4. 問題点と課題

ツールそのものの問題ではないが、今回の評価に使用したテキストは OCR でデジタル化したものであるため、単語中にハイフンが入っているケースや、OCR の認識誤りによるスペルミスも散見される。こういった誤りがあると、SVL にはマッチするものがないので、本来はレベル 1 に含まれるべきものがレベル 12 までの 12000 語にもあてはまらない、ということになる。テキスト入力用のインターフェイスを設計する際には、こうした誤りの修正を支援する機能が必要であろう。

現在は組み込んでいないが、授業支援のためには、設定したレベルを逸脱した単語については自動的に辞書引きを行い、単語をクリックするとレベルだけでなく語義も表示する機能が望まれる。

今回のツールは Solaris 上で作成したが、作成した部分は awk のスクリプトのみであり、他の Unix 互換 OS 上でも動作可能である。TreeTagger も含めて、より一般的な環境で動作する方が好ましいので、現在、Linux への移行を計画している。

テキスト入力用に Web インターフェイスを作成し、シェルの知識がなくても容易に使えるものとすることも必要である。さらに、ツールをプロキシサーバに組み込み、英語の Web を表示させると今回のツールのような形で HTML タグを付加する、といった方向も考えられる。本文中に記した品詞タグのマッチングによる精度向上等も含め、今後、一層の機能強化を行い、授業に役立つツールを目指す予定である。

References

[1] 英語テキストのレベル診断, 西嶋愉一, 2003PC カンファレンス論文集
[2] TreeTagger - a language independent part-of-speech tagger http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
[3] SVL 単語一覧 http://www.alc.co.jp/goi/svl_ichiran1.htm
[4] 最強のボキャブラリー, アルク, 2000