英語教科書の語彙分析

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、来年度からの実施に向けて、教養教育カリキュラムの大規模な刷新に取り組んでいる。その中で、英語については、所謂「言語の 4 技能」ごとに授業を設定する、なおかつレベル分けを行う、という考え方を柱として、カリキュラムの設計が行われている。

授業をレベル分けするための基準としては様々な考え方があろうが、今回のカリキュラム(新カリキュラムと呼ぶ)では、目安として語彙の数を用いる。例えば英語 I では 5000 語レベル、英語 II は 7000 語レベル、といった形である。

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道具として、テキストのレベル診断を行うものを一昨年の PC カンファレンスで発表した[1]。

今回は、実際に金沢大学の授業で使われている教科書について[1]と同様の分析を試みた。新カリキュラム下で個々の授業を計画する際の参考資料とするのが目的である。

2．手法について

英語の担当者に協力を要請し、昨年度および今年度に授業で使用した教科書 10 点を集め、個々のページをスキャナで取り込んだ後、OCR ソフト(Presto! OCR)を利用してデジタル化した。なお、デジタル化の対象は本文、教科書内のテスト等である。

その後の処理は[2]に準じたものである。まず、上でデジタル化したテキストを TreeTagger[3]で処理することにより、テキスト中の単語を見出し語の形に(lemmatize)する。

レベル分けのための語彙表はアルクの標準語彙水準 12000(Standard Vocabulary Level, SVL)[4]を使用した。12000 語の単語が 1000 語ごとに 12 にレベル分けされたものである。

TreeTagger で見出し語化したそれぞれの単語に対して単純に文字列比較をしてマッチングしている。

あとはレベルごとの出現回数をカウントすれば良いが、教科書という性格を考慮すると、重要なのはその単語が最初に出現したときであり、教科書としての難易度は単語の出現回数よりも、使われている語彙そのものの数に影響されると考えられるので、カウントの際は、同一の単語を重複して数えないよう加工することにした。同じ単語でも TreeTagger のタグが異なるものが現れることがあるが、これも同一と考える。この処理のため、重複した単語をまとめるフィルタを作成し、カウント前にそれを通している。フィルタを通さなければ、単純に出現回数をカウントすることも可能である。

3．結果と考察

10 点の教科書について分析を行った結果を図 1 に示した。A～J が個々の教科書、グラフの左側から SVL のレベル 1、2…に対応する。グラフ中で 13 となっている最も右側の部分は、SVL12000 語に含まれない単語である。

教科書によっては、SVL に含まれない語が 20%を越えている。SVL には固有名詞がほとんど含まれず、たとえば Japan や America、yen、dollar といったものも含まれていない。また、新しい表現(e-mail や BSE といったもの)も含まれない。ニュース英語を素材にした教科書(グラフ中の A)では、日本語をそのままローマ字表記したもの(amakudari)や企業名(Microsoft、Mitsubishi…)などが数多く出現しており、これらが SVL に含まれないもののパーセンテージを押し上げている。

こうした固有名詞や、個々のテキストが取り上げる分野ごとのテクニカルターム等は、教育の中で積極的に必要とされるものであり、これらのパーセンテージをそのまま 5000 語なり 7000 語なりのレベルを判断する材料に組み込むかどうかは、なお検討が必要である。

そこで、SVL に含まれない単語を外して計算したのが図 2 である。傾向としては図 1 と大きく変わらないが、図 2 では、たとえば 5000 語レベル(レベル 5)までで 90%近くを占めるもの(E・F・H・I)と、80%程度ないしそれ以下のもの(A・B・D・G・J)の相違が見て取れる。

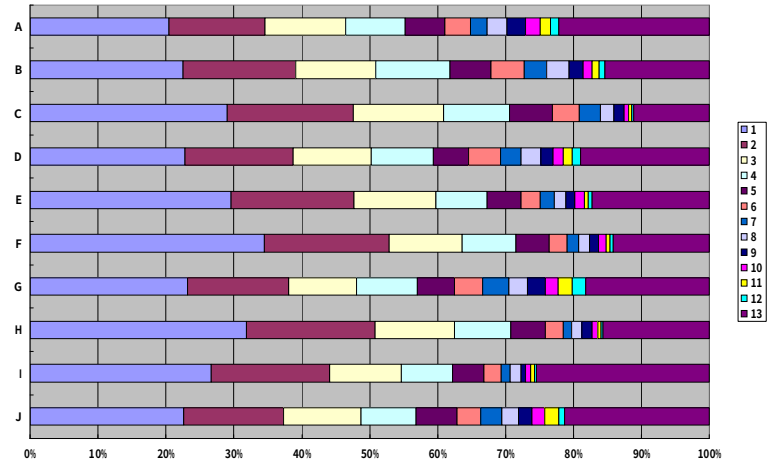
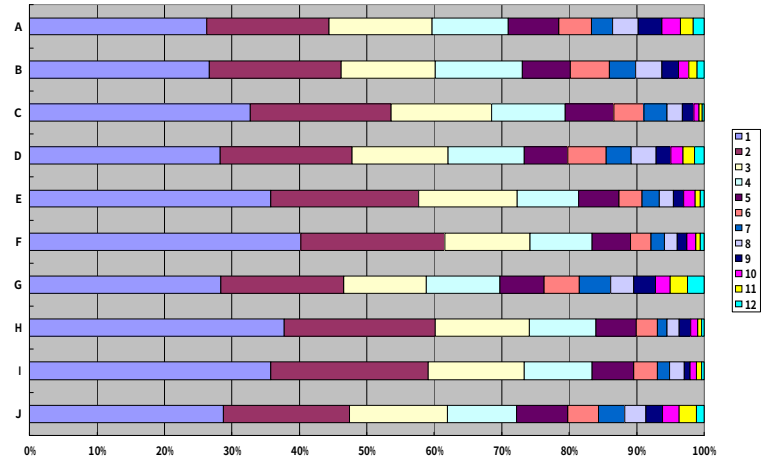


図 1．語彙レベルごとの比率(SVL に含まれない語も考慮)



教科書ごとの特徴もグラフに現れている。たとえば C は異文化コミュニケーションを題材としたもので、SVL に含まれない単語が少なく(11.2%)、図 2 でも 5000 語レベルまでの割合が高い(86.6%)。一方、G はネイチャーライティングのエッセイであり、もともと教科書として執筆されたものではないこともあって、5000 語レベルまでの割合は 76.3%と低くなっている。

今回の調査の目的としては、現在使われている教科書について、新カリキュラムで使用する際にどの程度使えるか、の目安を提供すること、特に、外国語の教員に数値だけでなく視覚的に相違を見せたい、ということがあったが、その目的には役に立つものと思われる。

4．課題

SVL には固有名詞や新しい単語が含まれない、ということは既に述べたが、これを補う仕組みが望まれる。現在は、中学・高校で履修する範囲の固有名詞と、個々のテキストの分野ごとの専門用語が、SVL に含まれない語という括りで一まとめにされてしまう。中学・高校の教科書を分析したものは既に市販されている[5]が、そういったものを活用することが考えられる。

今後は調査対象の教科書を増やし、金沢大学で使用されている教科書の中でどの位置にあるか等も視覚的に提示できるようにし、より授業支援の道具として役立つものを目指す。

最終的には、金沢大学の学生に英語の能力として何を身につけて欲しいのか、というところまで立ち戻った上で、独自の語彙表を作成するのが理想ではあるが、これは将来に向けての課題としておきたい。

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] 英語語彙指導のためのツール作り，西嶋愉一，2004PC カンファレンス論文集
[3] TreeTagger - a language independent part-of-speech tagger http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
[4] SVL 単語一覧 http://www.alc.co.jp/goi/svl_ichiran1.htm
[5] イー・キャスト 単語レベルチェック http://www.e-cast.jp/tangochk.htm