Automatic Evaluation and Simplification of Vocabulary with Applications in Course Materials Development

Laurence Anthony
Center for Language Education in Science and Engineering
Faculty of Science and Engineering, Waseda University
http://www.antlab.sci.waseda.ac.jp/
anthony@antlab.sci.waseda.ac.jp

JACET 2007 (Sept. 7th, 2007)

Outline

- Background
  - How many words does a learner need to know?
  - Implications for curriculum design
  - Lexical profiles and language simplification
- *AntVocabCheck*: A lexical profile and simplification tool
  - Features of *AntVC*
  - Getting started with *AntVC*
- Application of *AntVC* in the creation of course materials at Waseda University

How many words does a learner need to know?

- How many words are in English?
  - 988,968 different words *(Payack, 2006)*
  - Webster's 3rd International Dictionary ~54,000 word families *(Nation & Waring, 1997)*
- How many words does a native speaker know?
  - Native speakers typically learn 1000 word families each year of their life *(Nation, 2001)*
  - Typical university graduate ~20,000 word families *(Nation, 2001)*

How many words does a learner need to know?

- How many words does a non-native speaker know?
  - many adult learners of EFL know less than 5000 word families *(Nation & Waring, 1997)*
- Are all words equal?
  - Zipf's Law (1935)

Zipf's law stated that, in a corpus of natural language utterances, the frequency of any word is roughly inversely proportional to its rank in the frequency table.

A small number of words goes a long way!

How many words does a learner need to know?

- Result for the Brown Corpus (1 million running words)
  - 2000 basewords = 80% coverage
  - comprehension (without dictionary) is possible with 95% coverage *(Laufer, 1989)*
  - learners should **study** high frequency items
  - learners should **learn** low frequency items through incidental learning (e.g. extensive reading)

Vocab Size	Coverage (%)
1000	72.0
2000	79.7
3000	84.0
4000	86.8
5000	88.7
15,871	97.8

Francis and Kucera (1982)

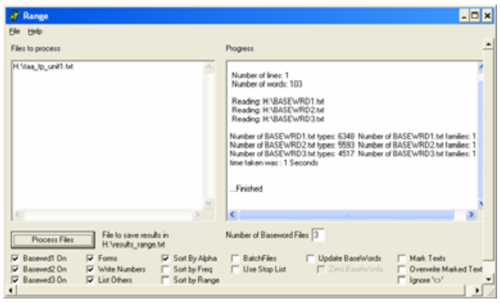
Implications for curriculum design

- How do we determine high frequency words?
  - General English
    - GSL: General Service List (West, 1953)
    - Brown Corpus (Kucera and Francis, 1967)
    - British National Corpus (eg. Leech et al. 2001)
    - JACET List of 8000 Basic Words (JACET, 2003)
  - Academic English
    - AWL: Academic Word List (Coxhead, 2000)
  - English for Specific Purposes
    - specialized corpora lists (via concordancing tools)
- The vocabulary of a task must be under the control of the learners → simplification

Implications for curriculum design

- Lexical profiles and language simplification
  - Current lexical profile tools
    - *Range* (Paul Nation)
    - *JACET 8000 Level Marker* (Shinichi Shimizu)
    - *VocabProfiler* (Tom Cobb)
  - AntVocabCheck (Laurence Anthony)
    - a modern, fast, easy-to-use, freeware vocabulary profiling tool with easy simplification editing tools

Range (I.S.P Nation)



H:\taa_tp_unit1.txt

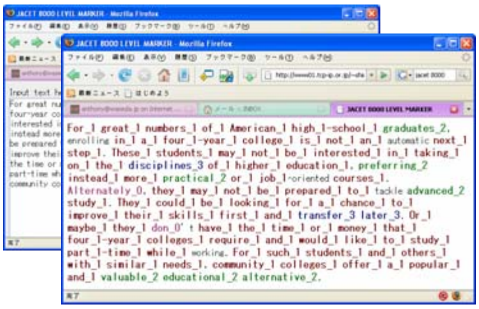
```
Number of lines: 1
Number of words: 103

Reading: H:\BASEWRD1.txt
Reading: H:\BASEWRD2.txt
Reading: H:\BASEWRD3.txt

Number of BASEWRD1.txt types: 6348  Number of BASEWRD1.txt families: 1
Number of BASEWRD2.txt types: 5593  Number of BASEWRD2.txt families: 1
Number of BASEWRD3.txt types: 4517  Number of BASEWRD3.txt families: 1
time taken was : 1 Seconds

...Finished
```

File to save results in H:\results_range.txt

Number of Baseword Files 3

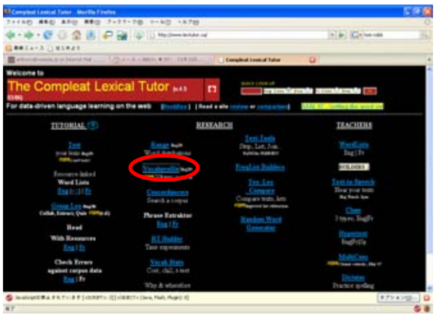
☑ Basewrd1 On ☑ Forms ☑ Sort By Alpha □ BatchFiles □ Update BaseWords □ Mark Texts
☑ Basewrd2 On ☑ Write Numbers □ Sort by Freq □ Use Stop List □ Zero BaseWords □ Overwrite Marked Text
☑ Basewrd3 On ☑ List Others □ Sort by Range □ Ignore '<>'

JACET 8000 Level Marker (Shinichi Shimizu)

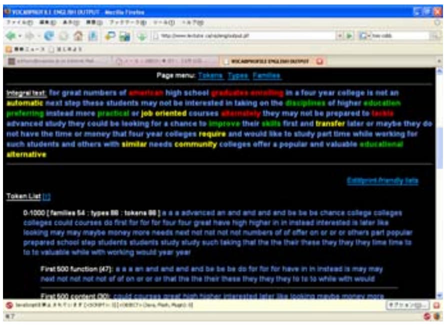

JACET 8000 LEVEL MARKER - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
JACET 8000 LEVEL MARKER
完了

For_1 great_1 numbers_1 of_1 American_1 high_1-school_1 graduates_2, enrolling in_1 a_1 four_1-year_1 college_1 is_1 not_1 an_1 automatic next_1 step_1. These_1 students_1 may_1 not_1 be_1 interested_1 in_1 taking_1 on_1 the_1 disciplines_3 of_1 higher_1 education_1, preferring_2 instead_1 more_1 practical_2 or_1 job_1-oriented courses_1. Alternately_0, they_1 may_1 not_1 be_1 prepared_1 to_1 tackle advanced_2 study_1. They_1 could_1 be_1 looking_1 for_1 a_1 chance_1 to_1 improve_1 their_1 skills_1 first_1 and_1 transfer_3 later_3. Or_1 maybe_1 they_1 don_0't have_1 the_1 time_1 or_1 money_1 that_1 four_1-year_1 colleges_1 require_1 and_1 would_1 like_1 to_1 study_1 part_1-time_1 while_1 working. For_1 such_1 students_1 and_1 others_1 with_1 similar_1 needs_1, community_1 colleges_1 offer_1 a_1 popular_1 and_1 valuable_2 educational_2 alternative_2.

Vocabprofile (T. Cobb)


The Compleat Lexical Tutor
For data-driven language learning on the web
TUTORIAL
RESEARCH
TEACHERS
Vocabprofile

Vocabprofile (T. Cobb)


VOCABPROFILE ENGLISH OUTPUT - Mozilla Firefox
Page menu: Tokens Types Families
Edit/print-friendly lists

Integral text: for great numbers of american high school graduates enrolling in a four year college is not an automatic next step these students may not be interested in taking on the disciplines of higher education preferring instead more practical or job oriented courses alternately they may not be prepared to tackle advanced study they could be looking for a chance to improve their skills first and transfer later or maybe they do not have the time or money that four year colleges require and would like to study part time while working for such students and others with similar needs community colleges offer a popular and valuable educational alternative

Token List [?]

0-1000 [families 54 : types 88 : tokens 88] a a a advanced an and and and and be be be chance college colleges colleges could courses do first for for for four four great have high higher in in instead interested is later like looking may may maybe money more needs next not not not not numbers of of offer on or or or others part popular prepared school step students students study study such taking that the the their these they they they time time to to to valuable while with working would year year

First 500 function (47): a a a an and and and and be be be do for for for have in in instead is may may next not not not not of of on or or or that the the their these they they they to to to while with would

First 500 content (30): could courses great high higher interested later like looking maybe money more

AntVocabCheck (L. Anthony)

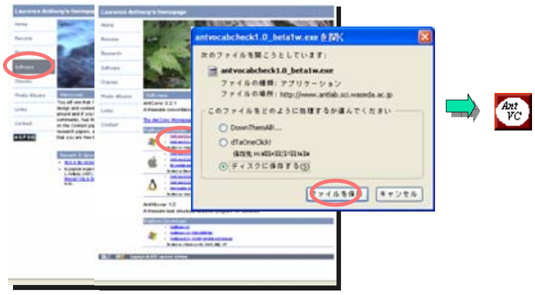
- Features of *AntVC*
  - Freeware
  - Standalone (no Internet connection required)
    - can handle **large** target files and/or vocab level lists
  - Multiplatform:
    - Windows (98+, NT, XP, Vista), Macintosh OS X, Linux
  - Flexible design
    - Can be used with any vocabulary level lists and target files
  - Fast processing (1 million words in ~30 sec.)
  - Export results to a plain text file or Excel file
  - Teacher friendly graphical user interface (GUI)

Getting started with *AntVC*

- **Step 1:** Download the software (*AntVC*)
- **Step 2:** Import a baseword family/lemma list (to convert word inflections to base forms)
  - e.g. Nation's basewords (pre-installed)
- **Step 3:** Import one or more level lists
  - e.g. Jacet 8000, GSL, AWL, ...
- **Step 4:** Analyze a target file

Step 1: Download the software

- Access homepage ⇨ Download ⇨ Double click ⇨ Start!



antvocabcheck1.0_beta1w.exe を開く

次のファイルを開こうとしています:

antvocabcheck1.0_beta1w.exe

ファイルの種類: アプリケーション

ファイルの場所: http://www.antlab.sci.waseda.ac.jp

このファイルをどのように処理するか選んでください

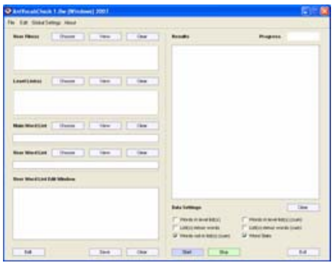
○ DownThemAll!...

○ dTaOneClick!

● ディスクに保存する(S)

Step 1: Download the software

- Access homepage ⇨ Download ⇨ Double click ⇨ Start!



http://www.antlab.sci.waseda.ac.jp/software/

Step 2: Import a baseword family/lemma list

...
aerial->aerial, aerials, aerially, aerialist, aerialists
aerobatic->aerobatic, aerobatics
aerobic->aerobic, aerobically
aerobics->aerobics
aerodrome->aerodrome, aerodromes
aerodynamic->aerodynamic, aerodynamically
aerofoil->aerofoil, aerofoils
aeronautical->aeronautical, aeronautically
aeronautics->aeronautics
aeroplane->aeroplane, aeroplanes, aero
aerosol->aerosol, aerosols
...

Step 3: Import one or more level lists

J 1000	J 2000	J 3000	J 4000
the	determine	disabled	republican
and	article	hey	used
to	cultural	significant	senator
of	direct	slight	clip
a	technique	file	according
in	soldier	invent	chapter
I	female	error	means
that	evidence	increasingly	investment
it	unless	arrival	review
you	recognize	agricultural	fucking
...	...	...	...

Step 4: Analyze a target file with *AntVC*

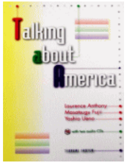
For great numbers of American high-school graduates, enrolling in a four-year college is not an automatic next step. These students may not be interested in taking on the disciplines of higher education, preferring instead more practical or job-oriented courses. Alternately, they may not be prepared to tackle advanced study. They could be looking for a chance to improve their skills first and transfer later. Or maybe they don't have the time or money that four-year colleges require and would like to study part-time while working. For such students and others with similar needs, community colleges offer a popular and valuable educational alternative.

Application of *AntVC* at School of Sci. & Eng., Waseda University

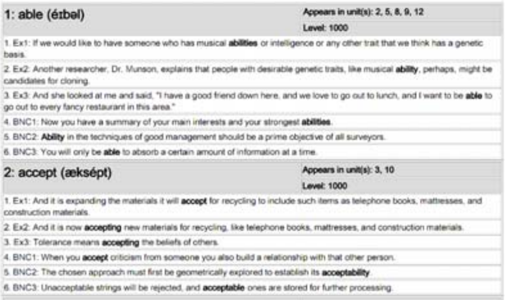
- Courses offered:
  - **1st year**: Communication Strategies, Academic Lecture Comprehension
  - **2nd year**: Academic Reading, Concept Building and Discussion
  - **3rd/4th year**: Technical Writing, Technical Presentation, Special Topics in Functional English
  - **Graduate School**: Technical Writing, Technical Presentation, Technical Communication
- Classes per week: 2 x 90 minutes
- Number of students: ~1800 (1st year)

Application of *AntVC* at School of Sci. & Eng., Waseda University

- Communication Strategies (1st year)
  - Classes per week: 1 x 90 minutes
  - Number of students per class: ~25
- Textbook: *Talking about America*
  - *AntVC* used for vocabulary selection, and lexical profiling (identification of ‘easy’ and ‘difficult’ vocabulary)



Application of *AntVC* at School of Sci. & Eng., Waseda University



1: able (éɪbəl)	Appears in unit(s): 2, 5, 8, 9, 12 Level: 1000
1. Ex1: If we would like to have someone who has musical **abilities** or intelligence or any other trait that we think has a genetic basis.	
2. Ex2: Another researcher, Dr. Munson, explains that people with desirable genetic traits, like musical **ability**, perhaps, might be candidates for cloning.	
3. Ex3: And she looked at me and said, "I have a good friend down here, and we love to go out to lunch, and I want to be **able** to go out to every fancy restaurant in this area."	
4. BNC1: Now you have a summary of your main interests and your strongest **abilities**.	
5. BNC2: **Ability** in the techniques of good management should be a prime objective of all surveyors.	
6. BNC3: You will only be **able** to absorb a certain amount of information at a time.	
2: accept (æksépt)	**Appears in unit(s): 3, 10 Level: 1000**
1. Ex1: And it is expanding the materials it will **accept** for recycling to include such items as telephone books, mattresses, and construction materials.	
2. Ex2: And it is now **accepting** new materials for recycling, like telephone books, mattresses, and construction materials.	
3. Ex3: Tolerance means **accepting** the beliefs of others.	
4. BNC1: When you **accept** criticism from someone you also build a relationship with that other person.	
5. BNC2: The chosen approach must first be geometrically explored to establish its **acceptability**.	
6. BNC3: Unacceptable strings will be rejected, and **acceptable** ones are stored for further processing.	

Summary

- Evaluation and simplification of vocabulary are essential components of language program design
- *AntVocabCheck (AntVC)* is a standalone, multiplatform, flexible, fast, and user friendly tool for lexical profiling and simplification
- *AntVC* has already proved to be an effective aid in the development of course materials at School of Science and Engineering, Waseda University