# TOEIC & TOEFL Vocabulary

## Secrets Revealed

JALT 2013 Kobe
Oct 26, 2013

Presenter: Guy Cihi

Lexxica R&D
2-7-8 Shibuya 5F
Shibuya-ku, Tokyo 150-0002
**info@lexxica.co.jp**

Copyright 2013 Lexxica

## Presentation Outline

1) What is “Coverage?”

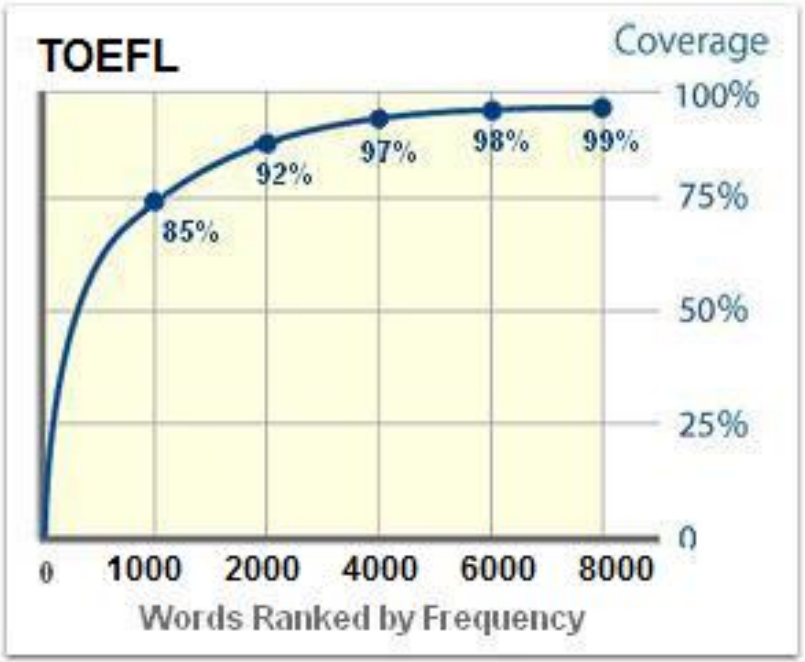
2) Corpus Analysis - TOEIC and TOEFL

3) Secrets of TOEIC and TOEFL vocabulary

4) How and why ETS uses esoteric vocabulary

5) How graded readers can best support TOEIC and TOEFL score increases

# Coverage




There are specific words that occur most frequently within a particular subject domain.

The most frequently occurring words provide the greatest amount of coverage for a domain.

Focusing on learning missing high frequency words is the fastest way to increase coverage of a domain.

Copyright 2013 Lexxica

lexxica

# We do our own corpus analysis work

We study exactly which words are required to master each subject area.

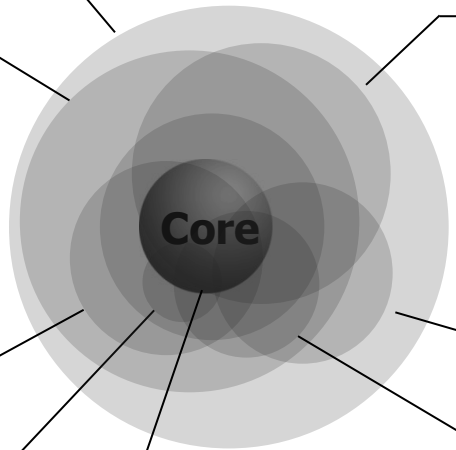
**All General English**
13,384 words

**Business English**
8,742 words

**College Entrance**
5,435 words

**High School**
3,552 words

**Elementary**
2,000 basic words



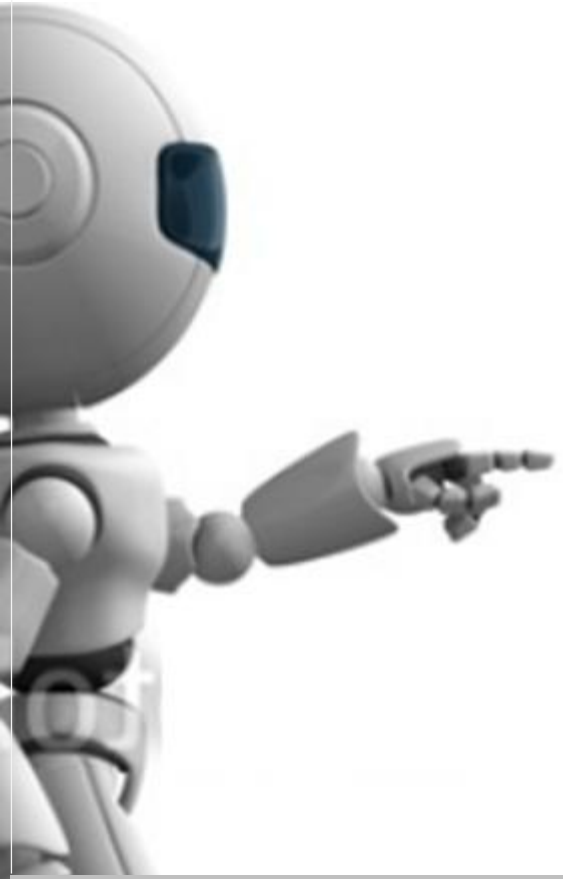

**TOEFL**
7,501 words

**TOEIC**
6,480 words

**IELTS**
5,870 words

Copyright 2013 Lexxica

# TOEIC Corpus Analysis



1,250,000 total words

## 14,652 different words

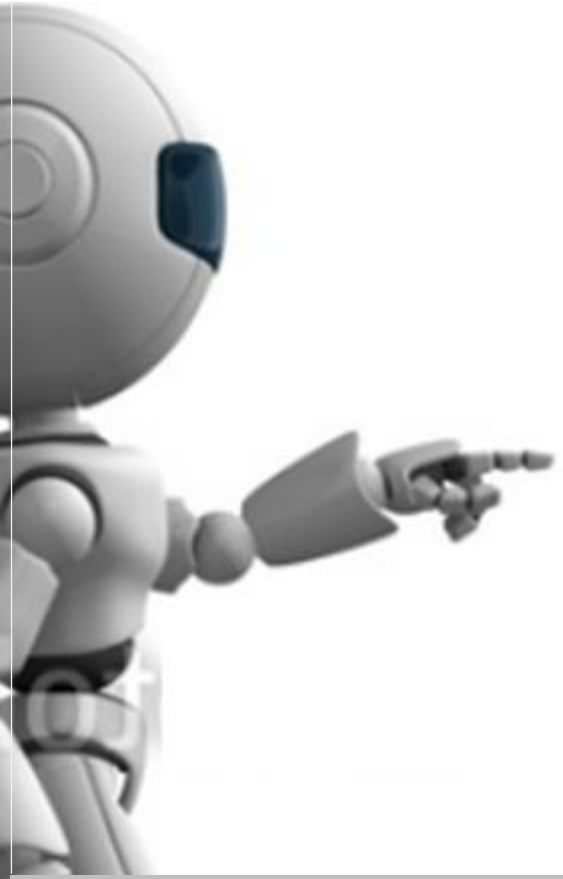
6,480 different words constitute
99% of all occurrences

982 different words constitute
90% of all occurrences.

These 982 are the absolutely essential
Super High Frequency words of TOEIC

Copyright 2013 Lexxica

# TOEFL Corpus Analysis



**1,250,000 total words**

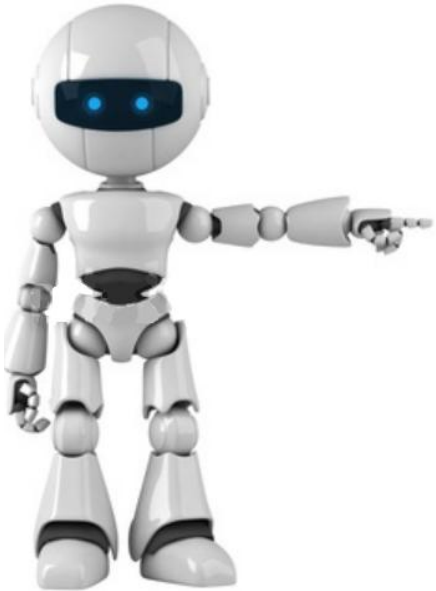
## 16,736 different words

**7,501 different words constitute 99% of all occurrences**

**1,513 different words constitute 90% of all occurrences.**

**These 1,513 are the absolutely essential Super High Frequency words of TOEFL**

Copyright 2013 Lexxica

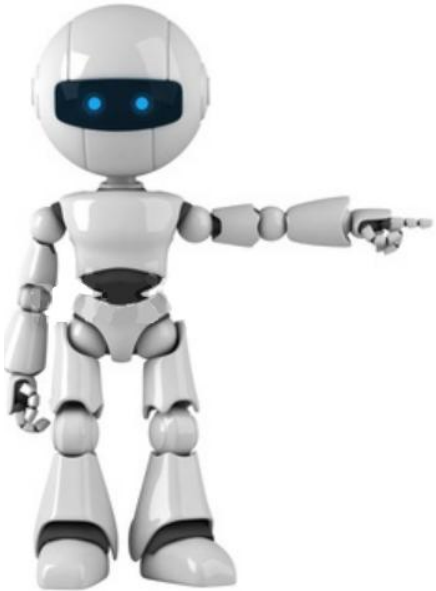


# Secret #1

TOEIC and TOEFL are Item Response Theory Proficiency Tests – not English ability diagnostic tests. **These tests are not designed to provide meaningful advice for improving English ability.**

Students are scored based on their correct responses to questions having known difficulty metrics. The difficulty metrics are established through statistical analysis of all prior uses of each question.

Copyright 2013 Lexxica

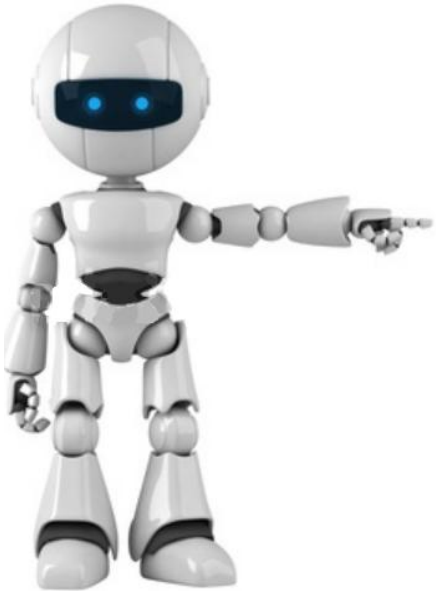


# Secret #2

Without a full range of questions from easy to difficult, Education Testing Service "ETS," would be unable to maintain its bell-curve and generate 'reliable' scores.

**It is impossible to write statistically difficult questions. Only field testing can identify the difficulty of questions.**

Copyright 2013 Lexxica

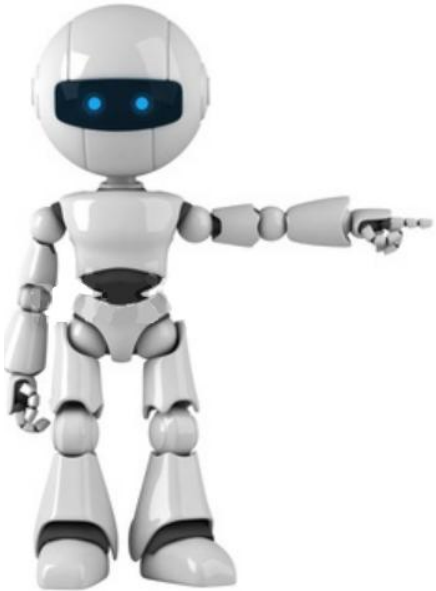


# Secret #3

95% of test questions are recycled.
5% are new questions that are in the process of being measured for difficulty.

**The 95% recycling requirement means that vocabulary on the tests can be accurately predicted.**

Copyright 2013 Lexxica

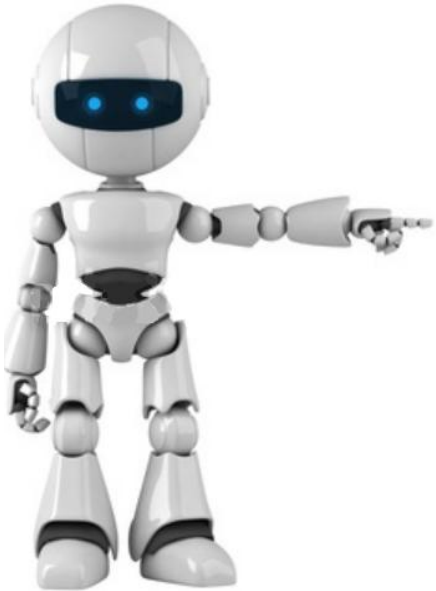


# Secret #4

**ETS has never, and likely will never issue a vocabulary guide for any of its major tests including: TOEIC, TOEFL, SAT and GRE.**

# Why?

Copyright 2013 Lexxica



# Secret #4

Because using difficult words, and irregular definitions, are the best way to create a wide variety of questions at all levels of difficulty.

**Publishing an official vocabulary guide would both expose a scoring system vulnerability** and defeat the purpose of their tests which is to measure familiarity and proficiency with authentic English.

Copyright 2013 Lexxica

# **TOEIC, TOEFL** (and IELTS) versus General English

**1/3 of the words in all parts of TOEIC and TOEFL are <u>not common</u>, high frequency words in General English.**

(¼ of the words in IELTS.)

Copyright 2013 Lexxica

# What kinds of words

Copyright 2013 Lexxica

# Top 2000 high frequency words of TOEIC and General English

| TOEIC | General |
|---|---|
| ability | ability |
| able | able |
| aboard | about |
| about | above |
| above | abroad |
| abroad | absence |
| absence | absolute |
| absent | absolutely |
| absolutely | absorb |
| abstract | abuse |
| accept | academic |
| | accept |

| TOEIC | General |
|---|---|
| gain | gain |
| gallery | gall |
| gallon | game |
| game | gap |
| garage | garage |
| garbage | garden |
| garden | gas |
| gardener | gate |
| gas | gather |
| gasoline | gaze |
| gate | gear |
| gather | gene |
| gender | general |
| general | |

Frequent only in the TOEIC corpus.
Frequent only in the General corpus.

Our general corpus contains 850 million words from all genres.

Copyright 2013 Lexxica

# What does this mean?

EFL students can't learn the words they need because they aren't in their study and reading materials.

**(Because study materials are simplified.)**

Copyright 2013 Lexxica

# I used to say:

Education Testing Service (ETS) purposefully uses difficult words and seldom used meanings of common words because otherwise their scoring system fails.

**(Then I talked to ETS authors and editors)**

Copyright 2013 Lexxica

# Now I say:

Education Testing Service (ETS) ~~purposefully~~ uses difficult words and seldom used meanings of common words because otherwise their scoring system fails.

Copyright 2013 Lexxica

# To create new test questions:

Authors are told to search through authentic materials to find texts and dialogs to adapt for the different types of test questions.

Copyright 2013 Lexxica

# To evaluate new test questions:

When finished, the authors and editors do not know how difficult their new questions are.

The only way to find out is for ETS to put them into actual tests alongside questions for which they do know the difficulty.

Copyright 2013 Lexxica

# Testing the test questions:

On every TOEIC and TOEFL test 5% of the questions are new questions that have no affect on scoring.

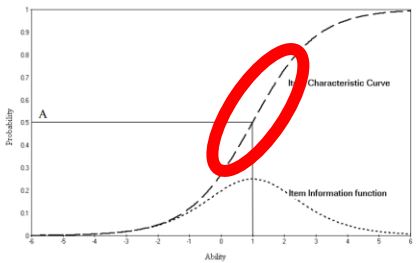
95% are recycled questions that have known and reliable difficulties that can be used for scoring.

Copyright 2013 Lexxica

# ETS's Primary Concern

ETS's primary concern is the **consistency** with which their test scores reflect each respondent's **relative proficiency** with **authentic** English.

Figure 1. Characteristic Curve and Information Function of the Item



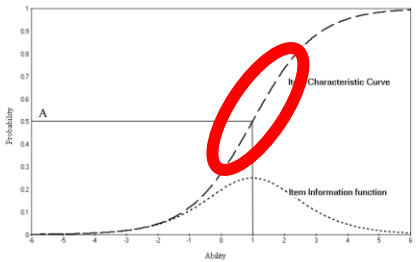

Copyright 2013 Lexxica

# From corpus analysis we confirm:

1/3 of the words on TOEIC and TOEFL tests are low frequency ‘authentic’ vocabulary words.

Vocabulary is the primary reason that one test question is more or less difficult than another.

Figure 1. Characteristic Curve and Information Function of the Item




Copyright 2013 Lexxica

# Note that many of the 1/3 low frequency words have multiple meanings

| TOEIC | General |
|---|---|
| ability | ability |
| able | able |
| aboard | about |
| about | above |
| above | abroad |
| abroad | absence |
| absence | absolute |
| absent | absolutely |
| absolutely | absorb |
| abstract | abuse |
| accept | academic |
| | accept |

| TOEIC | General |
|---|---|
| gain | gain |
| gallery | gall |
| gallon | game |
| game | gap |
| garage | garage |
| garbage | garden |
| garden | gas |
| gardener | gate |
| gas | gather |
| gasoline | gaze |
| gate | gear |
| gather | gene |
| gender | general |
| general | |

Frequent only in the TOEIC corpus.

Frequent only in the General corpus.

Our general corpus contains 850 million words from all genres.

Copyright 2013 Lexxica

## Typical low frequency definition: crack

A line along which something has split without breaking into separate parts: “a crack in the surface.”

An illegal street drug: “possession of crack."

Very good, esp. at a specified activity: “He’s a crack shot.”

To open something after making a concerted effort: “to crack a safe.”

Copyright 2013 Lexxica

## Typical low frequency definition: crack

A line along which something has split without breaking into separate parts: “a crack in the surface.”

An illegal street drug: “possession of crack."

Very good, esp. at a specified activity: “He’s a crack shot.”

**ETS used this:**

“…it took several years for Apple to _____ the market.”

A: crack  B: break open  C: secure  D: invert

Copyright 2013 Lexxica

# Why use low frequency definitions?

They are difficult and they are **authentic**.

**(ETS doesn't promise practical English.)**

Copyright 2013 Lexxica

**ETS's advice for scoring higher on TOEIC and TOEFL is to read authentic texts.**

(Graded readers can't help because the vocabulary is simplified)

Copyright 2013 Lexxica

# How much authentic text?

**Based on incidence of occurrence research by Rob Waring, they'll need to read 6,250 hours of authentic text in order to meet the lower frequency test words often enough to learn them.**

A statistical analysis of the number of English words you need to meet (at given recurrence rates) to 'learn' that number of words

| A | B | C (= 100 / B) | D (= x times C) | | | | E (= D / Book length) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Word rank | Percentage of general English that this word covers | Number of running words needed to be met to meet all these words once | Volume of text you need to read to meet the words at these recurrence rates | | | | Number of books to cover this volume given these recurrence rates | | | | |
| | | | 5 times | 10 times | 20 times | 50 times | Book length | 5 times | 10 times | 20 times | 50 times |
| 1st most frequent (*the*) | 5.83898% | 17 (1) | 86 | 171 | 343 | 856 | 4,500 | 0.0 | 0.0 | 0.1 | 0.2 |
| 2nd most frequent (*be*) | 5.12332% | 20 | 98 | 195 | 390 | 976 | 4,500 | 0.0 | 0.0 | 0.1 | 0.2 |
| 25th (*as*) | 0.44382% | 225 | 1,127 | 2,253 | 4,506 | 11,266 | 4,500 | 0.3 | 0.5 | 1.0 | 2.5 |
| 50th (*like*) | 0.24109% | 415 | 2,074 | 4,148 | 8,296 | 20,739 | 4,500 | 0.5 | 0.9 | 1.8 | 4.6 |
| 100th (*hear*) | 0.10505% | 952 | 4,759 | 9,519 | 19,038 | 47,595 | 4,500 | 1.1 | 2.1 | 4.2 | 10.6 |
| 500th (*present*) | 0.02477% | 4,037 | 20,183 | 40,366 | 80,732 (4) | 201,829 | 4,500 | 4.5 | 9.0 | 17.9 | 44.9 |
| 1000th (*blood*) | 0.01172% | 8,533 (3) | 42,665 | 85,329 | 170,658 | 426,645 | 10,000 | 4.3 | 8.5 | 17.1 | 42.7 |
| 1500th (*intent*) | 0.00677% | 14,773 | 73,864 | 147,727 | 295,455 | 738,636 | 15,000 | 4.9 | 9.8 | 19.7 | 49.2 |
| 2000th (*stumble*) | 0.00432% (2) | 23,103 | 115,625 | 231,250 | 462,500 | 1,156,250 | 20,000 | 5.8 | 11.6 | 23.1 | 57.8 |
| 3000th (*sergeant*) | 0.00211% | 47,343 | 236,713 | 473,425 | 946,850 | 2,367,126 | 30,000 | 7.9 | 15.8 | 31.6 | 78.9 |
| 5000th (*satellite*) | 0.00076% | 132,143 | 660,714 | 1,321,429 | 2,642,857 | 6,607,143 | 80,000 | 8.3 | 16.5 | 33.0 | 82.6 |
| 10,000th (*relativity*) | 0.00016% | 632,000 | 3,164,474 | 6,328,947 | 12,657,895 | 31,644,733 | 80,000 | 39.6 | 79.1 (5) | 158.2 | 395.6 |

Examples:
(1) The most frequent word in English (*the*) covers 5.839% of any general English text (i.e. it occurs once in every 17 words).
(2) The 2000th most frequent word in English covers 0.00432% of any general English text (and occurs once every 23,103 words).
(3) To meet all the 1000 most frequent words in English once, you'd need to read 8,533 words.
(4) To meet all the 500 most frequent words in English 20 times, you'd need to read 80,732 words.
(5) To meet all the 10,000 most frequent words in English 10 times, you'd need to read 79.1 books that are 80,000 words long.

© R. Waring 2007

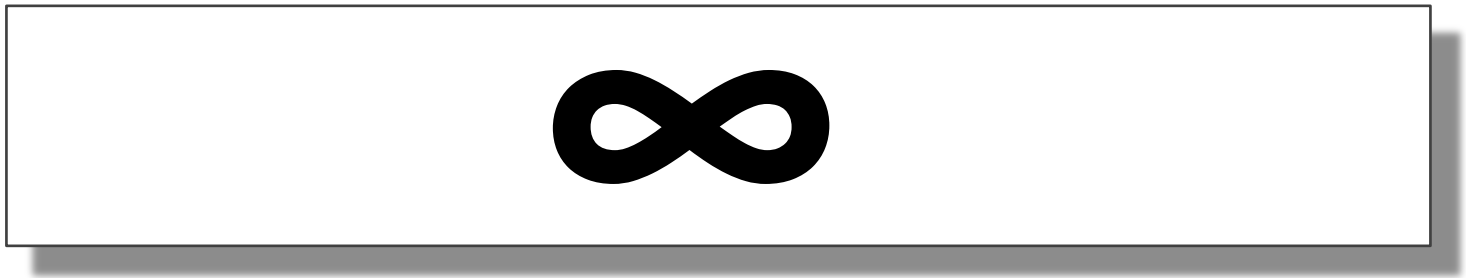
Copyright 2013 Lexxica

**Reading at 70 authentic words per minute…**

**2 hours each day for 8.5 years**

Copyright 2013 Lexxica

**Reading at 70 graded words per minute…**



Copyright 2013 Lexxica

# Graded readers are general English

**All General English**

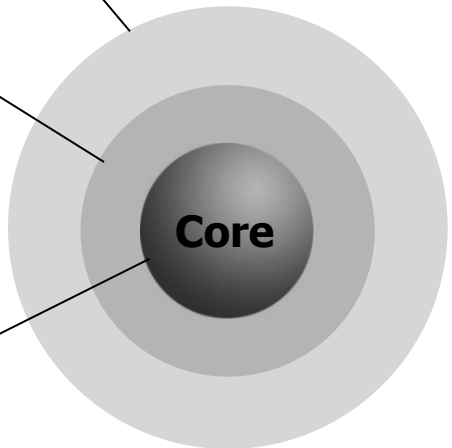
18,000 semantemes

**Advanced Graded Readers**

9,000 semantemes

**99% of Graded Readers**

4,000 semantemes




Copyright 2013 Lexxica

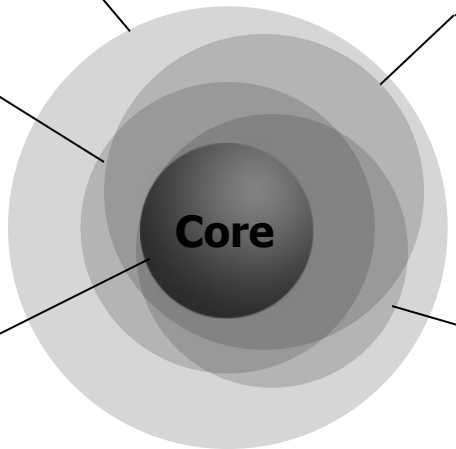
lexxica

# TOEIC and TOEFL are not general English

**All General English**
18,000 semantemes

**Advanced Graded Readers**
9,000 semantemes

**99% of Graded Readers**
4,000 semantemes




**TOEFL**
9,000 semantemes

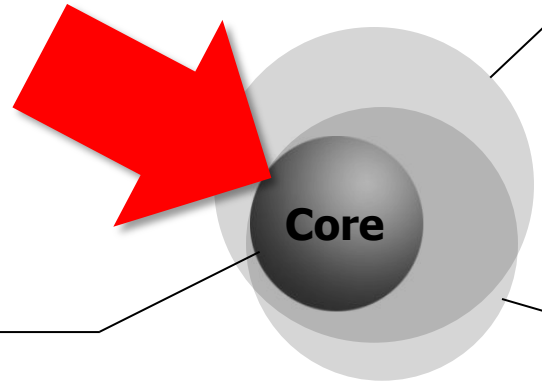
**TOEIC**
8,000 semantemes

Copyright 2013 Lexxica

lexxica

# TOEIC and TOEFL are not general English

**EFL students are here…**




**TOEFL**

9,000 semantemes

**TOEIC**

8,000 semantemes

**99% of Graded Readers**

4,000 semantemes

Copyright 2013 Lexxica

# How can graded reading help EFL students prepare for TOEIC and TOEFL?

Copyright 2013 Lexxica

90% of the words that occur in beginner and intermediate level graded readers are also super high frequency words in the TOEIC and TOEFL domains.  Because the tests are timed, students who can process the Super High Frequency words faster enjoy a huge scoring advantage. Graded readers can't teach vocabulary they don't contain but, they can help students develop automaticity (instant recognition) for the Super High Frequency words occurring in every TOEIC and TOEFL.

Copyright 2013 Lexxica

# What is the best way to use existing graded readers to improve reading and listening?

Copyright 2013 Lexxica

# Repeated timed aural readings.

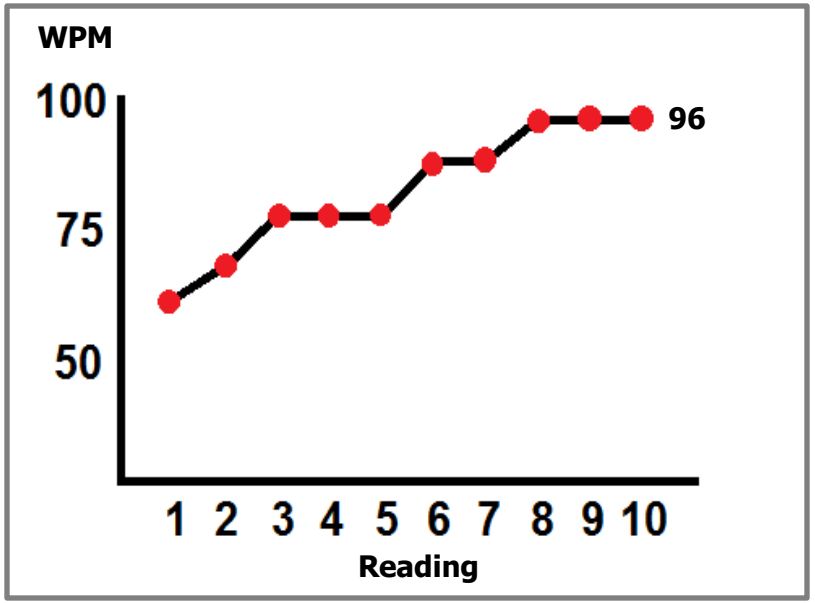
Copyright 2013 Lexxica

# Example of a repeated, timed, spoken reading approach. This method is highly effective!

**Spoken Reading Speed**

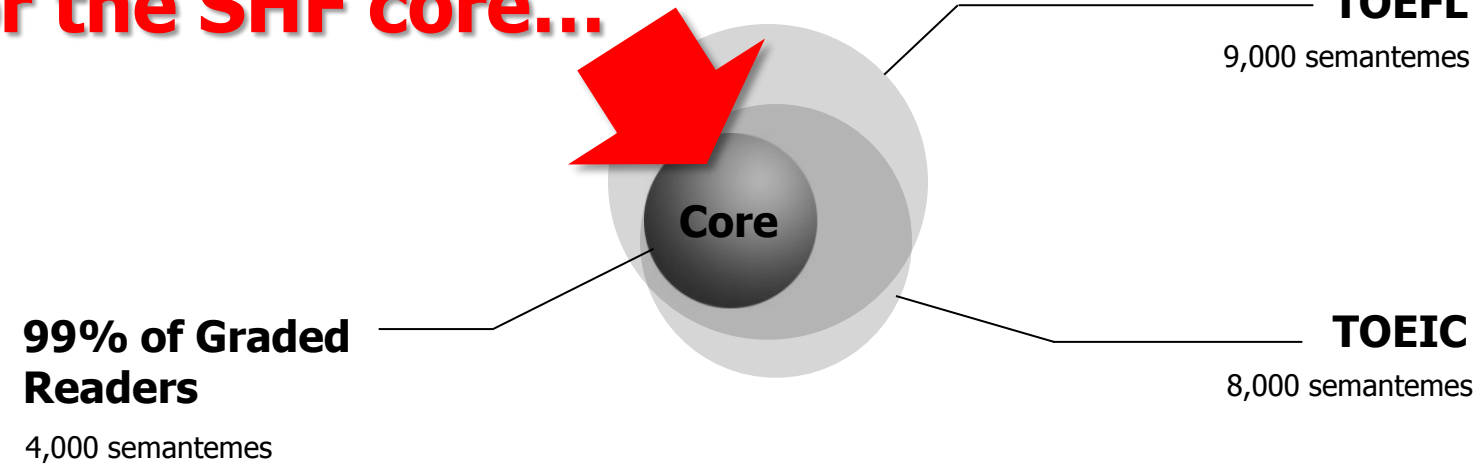
**Goal: 90**

| Title; Headwords | | Minutes | Total words | Words per min. |
|---|---|---|---|---|
| Good Dog, Bad Dog; 75 | 1 | 10 | 622 | 62 |
| Good Dog, Bad Dog; 75 | 2 | 9 | 622 | 69 |
| Good Dog, Bad Dog; 75 | 3 | 8 | 622 | 78 |
| Good Dog, Bad Dog; 75 | 4 | 8 | 622 | 78 |
| Good Dog, Bad Dog; 75 | 5 | 8 | 622 | 78 |
| Good Dog, Bad Dog; 75 | 6 | 7 | 622 | 89 |
| Good Dog, Bad Dog; 75 | 7 | 7 | 622 | 89 |
| Good Dog, Bad Dog; 75 | 8 | 6.5 | 622 | 96 |
| Good Dog, Bad Dog; 75 | 9 | 6.5 | 622 | 96 |
| Good Dog, Bad Dog; 75 | 10 | 6.5 | 622 | 96 |




Copyright 2013 Lexxica

# Implemented properly, a graded speed-reading program can help develop automaticity for the SHF core…



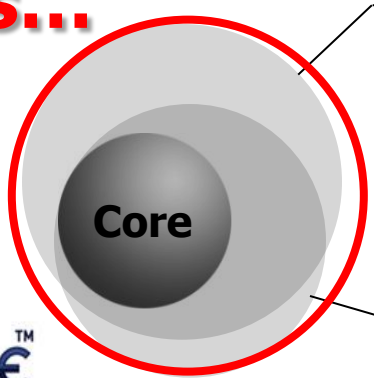

Copyright 2013 Lexxica

# The WordEngine high speed vocabulary system has been proven to develop automaticity for all of the words...

WordEngine™
ワードエンジン

High Speed Vocabulary Learning System

Learn more than 100 new words per day!




**TOEFL**

9,000 semantemes

**TOEIC**

8,000 semantemes

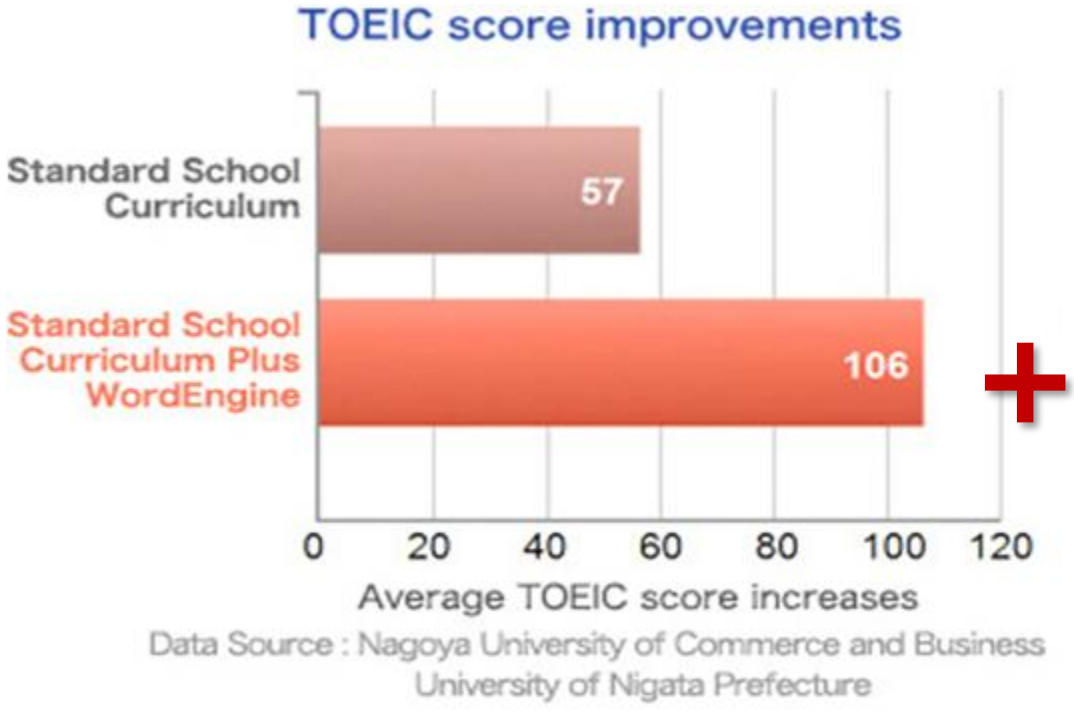
Copyright 2013 Lexxica

Word Engine

**When improved outcomes are important, professionals trust WordEngine to get results!**

Word Engine

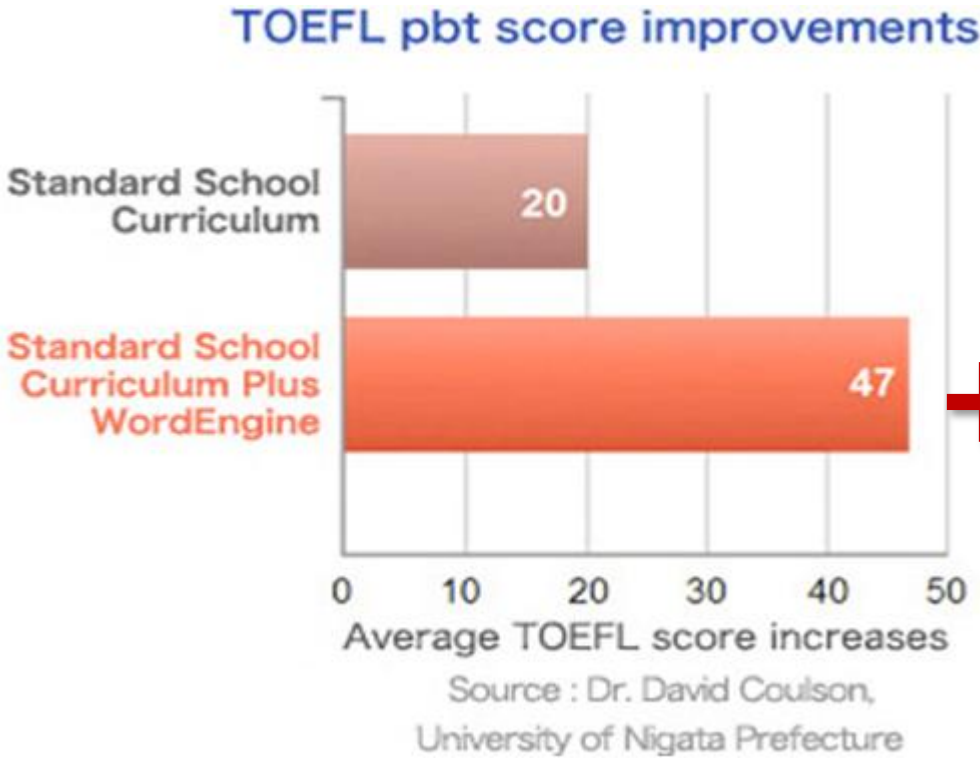
## Average TOEIC score increases




**+86%**

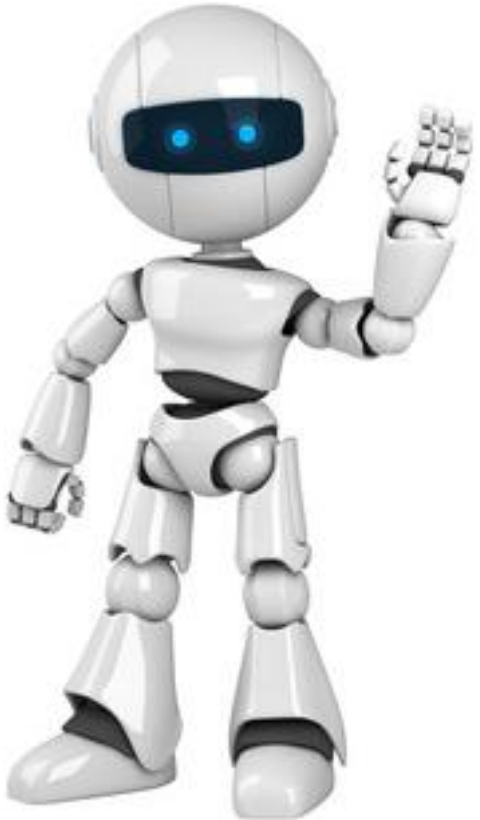
Word Engine

## Average TOEFL score increases




**+135%**

# Contact Lexxica to start a trial program at your school.



Lexxica R&D
2-7-8 Shibuya 5F
Shibuya-ku, Tokyo 150-0002
**info@lexxica.co.jp**

Copyright 2013 Lexxica