Extensive Reading and Viewing for Vocabulary Growth: Some Insights from Corpus Linguistics

Dr Clarence Green

School of Education, Federation University
Australia

c.green@federation.edu.au

Room 130A, Building 901, Berwick,
Melbourne, 100 Clyde Road, 3806
+61343137955
Outline

• Extensive Reading and Extensive Viewing
• Incidental Learning
• Corpus Modelling: How and Why?
• Corpus modelling ER/EV as a source for academic vocabulary
• Corpus modelling ER as a pedagogy to acquire a 9000 word vocabulary
About me

• Senior Lecturer, School of Education, Federation University Australia.
• PhD and MA Tesol (University of Melbourne)
• Corpus linguist: vocabulary, first and second language acquisition, literacy, psycholinguistics, and English grammar.
• Particularly interested in corpus research k-12 contexts and teacher training.
Vocabulary

- Fundamentally important to SLA/ FLA (Nation, 2006)
- Correlates with comprehension (Schmitt et al., 2017)
- Academic performance? (Milton, 2022)
- In younger children (L1/L2), benefits behaviour, social skills

- How much vocabulary do we need?
- What type of vocabulary?
- What are effective pedagogies?
Extensive Reading/ER and Extensive Viewing/EV

**ER:** vocabulary pedagogy (Krashen, 2016; Nation & Waring, 2019).
  - Self-select, read for pleasure.
  - Large amount of comprehensible input, prolonged, focus on meaning.
  - Text difficulty slightly > proficiency (Renandya & Jacobs, 2016).

**EV:** multimodal input (visual, auditory etc).
Extensive Reading/ER and Extensive Viewing/EV

- Pedagogies to develop vocabulary incidentally.
- How much and what kind?

- Gallagher et al. (2019): “consideration of whether instruction is explicit or incidental is important given the time constraints teachers face” (p. 100).
Incidental and Intentional Learning

- **Incidental learning**: vocabulary acquired within a meaning-focused task, e.g. reading comprehension.
- Repetition (Nation, 2014), salience, engagement.
- No number = acquisition
- From 2-20 is a safe bet in comprehensible input (e.g., 95-98% surrounding vocab known) (Webb, 2020)

- **Intentional learning**: direct instruction, off word-lists, teacher-selected (Beck et al., 2002)
How much and what type of vocabulary?

• General Vocab: foundation for grammar, phonology, comprehension (Nation, 2016)
• Academic Vocab = foundation for academic literacy (Coxhead, 2018).
  • Discipline-specific = conceptual knowledge (Green & Lambert, 2018)
Corpus Modelling ER: Academic Vocabulary?

- McQuillan (2019): *Harry Potter and the Prisoners of Vocabulary Instruction: Acquiring Academic Language at Hogwarts*

  - 22 novels e.g. Harry Potter, Twilight, Hunger Games etc.
  - 1,025,943 words
  - How many AWL targets?
  - How many repeated 10, 12, 25 times? (rates associated with incidental learning)
  - How far would ER of 1 hour a day (150 wpm) get you?
Corpus Modelling ER: Academic Vocabulary?

Corpus model results:
• 484 of 570 (85%) the AWL occurs in ER corpus
• < one year = 113 (20%) to 213 (37%) of AWL at rates associated with incidental learning, depending on N repetitions targeted
Corpus Modelling EV: Academic Vocabulary?

- Dang (2019) medical vocabulary in EV of medical TV?
- Extracted vocab from lectures
- Modelled at 5-20 repetitions
- Corpus = 11,036,771 words, 2,073 TV episodes
- All episodes viewed, 99% > 20 times
- A few seasons: 76% x 5 times, 66% x 10, 60% x 15, 56% x 20 times.
Corpus Modelling: ER/EV (Green, 2022)

• AWL & Secondary Vocab Lists: which items through ER/EV? (Green & Lambert, 2018)
• Not only disciplinary-related EV or juvenile fiction, but through general ER/EV.
• How much input required (6, 12, 20 encounters) & how long?
• Model: Multiple reading rates, 100, 260, 350 wpm
  Viewing/speech rates, 80, 140, 200 wpm (Romero-Fresco, 2016)
Corpus Modelling: ER/EV

- ER Corpus: COCA general fiction, 21,000,000 words e.g. The Da Vinci Code, Diary of a South Beach Party Girl.
- EV Corpus: COCA Movies + TV, 33,045,952 words, e.g. Buffy the Vampire Slayer, Mission Impossible, Fox News, CNN Showbiz.

<table>
<thead>
<tr>
<th>AWL: ER</th>
<th>12 x Input Required (Words)</th>
<th>12 x @ 100 wpm Reading Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPARENT</td>
<td>136287</td>
<td>22.71</td>
</tr>
<tr>
<td>INVESTIGATE</td>
<td>137985</td>
<td>23.00</td>
</tr>
<tr>
<td>REVEAL</td>
<td>142861</td>
<td>23.81</td>
</tr>
<tr>
<td>EVIDENT</td>
<td>145389</td>
<td>24.23</td>
</tr>
<tr>
<td>SVL (Biology): ER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>INTENSITY</td>
<td>740469</td>
<td>123.41</td>
</tr>
<tr>
<td>HEARTBEAT</td>
<td>742627</td>
<td>123.77</td>
</tr>
<tr>
<td>MUSCULAR</td>
<td>751390</td>
<td>125.23</td>
</tr>
<tr>
<td>INSECT</td>
<td>758099</td>
<td>126.35</td>
</tr>
</tbody>
</table>
Corpus Modelling: ER/EV

- AWL, all 570 targets in ER & EV.
- < 3 years ER, 45 mins a day, at 260 wpm, 98.5% > 12 times.
  100 wpm, 93% > 12, 93%.
- EV takes longer but largely similar
- 40%-60% disciplinary academic vocab sufficient encounters for incidental learning within 1-3 years ER/EV (approx. 45 min a day).
Corpus Modelling: What about an entire functional vocabulary?

- Past 15 years, series of influential corpus modelling studies.
- General English proficiency + could ER “do the entire job” or most of it (Krashen, 1989, p. 448)
- 9000 word families = 95%-98% coverage = Comprehension Coverage Model (CCM)
  “cornerstone of research and pedagogy concerning the importance of vocabulary to reading” (McLean, 2021).
To model if incidental learning to 9000 words through ER is possible:

• Books with 95%-98% vocabulary coverage need to exist for vocabulary sizes from beginner to proficient (Schmitt et al., 2017).
• Enough for self-selection “at the right level” (Nation, 2014, p. 7).
• Rates associated with incidental learning (Uchihara et al. 2019).
• Input & timeframes ‘reasonable’ (McQuillan, 2016b).

• If modelling cannot show above, claims ER might “do the entire job” or most (probably) falsified without real-world testing (Cobb, 2016).

- Threshold of 6 encounters for incidental learning
- Random sample from BNC lists: 1000, 2000, 3000.
- Computed repetitions in Brown (fiction, press, academic) for word families.
- Results: 1000 level, all but 1-2 words at threshold, no 3000 and one 2000-level.
- 300,000 words from seven Jack London novels.
- Results: 3000 level, only 469 repeated 6 or more times

- Cobb (2007): “extreme unlikelihood of developing an adequate L2 reading lexicon through reading alone, even in highly favourable circumstances”
Corpus Modelling: Major Studies. Nation (2014)

- 25 novels, e.g. *Animal Farm, Great Gatsby*
- BNC-COCA lists > 1000 - 9000 word families
- Threshold of 12 encounters, at 150 and 200 wpm
- Results: “if learners read a total of 3 million tokens, then they would meet the 1st 9,000 words often enough to have a chance of learning them” (p. 7)
- Nation (2016) “1000 words a year through reading” (p. 305)
- But few books for mid-range vocab sizes: “poor conditions for reading and incidental vocabulary learning” (p. 13).

Nation (2014) “largely positive study of the opportunities for learning through input” (p. 13).
Corpus Modelling: Major Studies. McQuillan (2016)

- Corpus of Juvenile Fiction: 14 novels, *Twilight*, *Hunger Games* etc.
- Results: many readable (95-95%) with mid-freq. vocab size, e.g. *Hunger Games* 95% at 4000.
- Re-modelling Nation’s (2014) figures at 1 hour a day, 12 repetitions:

  “sufficient input each step of the way, all at Nation’s recommended 98% vocabulary coverage…a little over three years of reading takes readers all the way to the 9,000-word family level” (McQuillan, 2016: p. 72).
Corpus Modelling: Green (forthcoming)

- 11m words, fiction (4989 books) from COCA (Davies, 2011)
- AntWord Profiler (Anthony, 2014): how many readable +95% at each BNC-COCA list (Laufer, 2020).
- Created subcorpora ‘at the right’ level for different vocabulary sizes.
- Modelled time to reach 12 encounters of words from subsequent levels.
- Modelling at: 150 wpm, 60 minutes per day (Nation, 2014; McQuillan, 2016a).
## Corpus Modelling: Green (forthcoming)

<table>
<thead>
<tr>
<th>Vocabulary size</th>
<th># books with 95% coverage</th>
<th># words</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000 level</td>
<td>5</td>
<td>53,617</td>
</tr>
<tr>
<td>2000 level</td>
<td>393</td>
<td>1.8m</td>
</tr>
<tr>
<td>3000 level</td>
<td>1990</td>
<td>9.4m</td>
</tr>
<tr>
<td>4000 level</td>
<td>3470</td>
<td>16.6m</td>
</tr>
<tr>
<td>5000 level</td>
<td>5600</td>
<td>26.5m</td>
</tr>
<tr>
<td>6000 level</td>
<td>6713</td>
<td>30.8m</td>
</tr>
<tr>
<td>7000 level</td>
<td>7111</td>
<td>32.2m</td>
</tr>
<tr>
<td>8000 level</td>
<td>7270</td>
<td>32.7m</td>
</tr>
<tr>
<td>9000 level</td>
<td>7349</td>
<td>32.9m</td>
</tr>
</tbody>
</table>
## Corpus Modelling: Green (forthcoming)

<table>
<thead>
<tr>
<th>Reading at 95%</th>
<th># Words needed</th>
<th>Reading time @ 150 wpm</th>
<th>3K 12+</th>
<th>4K 12+</th>
<th>5K 12+</th>
<th>6K 12+</th>
<th>7K 12+</th>
<th>8K 12+</th>
<th>9K 12+</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000 k</td>
<td>4.5m</td>
<td>1 y, 4.5 months</td>
<td>800</td>
<td>683</td>
<td>573</td>
<td>461</td>
<td>314</td>
<td>226</td>
<td>167</td>
</tr>
<tr>
<td>3000 k</td>
<td>1.7m</td>
<td>6 months</td>
<td>800</td>
<td>727</td>
<td>602</td>
<td>490</td>
<td>353</td>
<td>295</td>
<td></td>
</tr>
<tr>
<td>4000 k</td>
<td>1.4m</td>
<td>5 months</td>
<td>800</td>
<td>688</td>
<td>597</td>
<td>470</td>
<td>383</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000 k</td>
<td>3.0m</td>
<td>11 months</td>
<td>800</td>
<td>738</td>
<td>638</td>
<td>583</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6000 k</td>
<td>2.8m</td>
<td>10 months</td>
<td>800</td>
<td>710</td>
<td>659</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7000 k</td>
<td>4.1m</td>
<td>1 y, 3 months</td>
<td>800</td>
<td>756</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8000 k</td>
<td>3.0</td>
<td>11 months</td>
<td>800</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Time estimates for +12, reading at right level
What are the take-home points?

• My model: 6.3 years from reading at 2k level to 9k level if at one hour a day.
• Less than Nation’s (2014) approx. 9(7?) years
• More than McQuillan’s (2016) 3 years, but at 2 hours a day, 3.5 years ok
• Like McQuillan (2016) evidence of a pathway through comprehensible input
• Unlike Cobb (2007), ER repetition at incidental learning rates in mid-freq ranges
Conclusions: what are the insights from Corpus Linguistics?

- Corpus approaches to ER ad EV provide strong support for the value of ER/EV
- Academic vocabulary likely from general pleasure reading and viewing
- EV appears to be rich input along with ER.
- With access to books, comprehensible input at the right level is available
- More advanced vocab occurs in Comprehensible Input
- Can ER do ‘most of the job’ (i.e. 9th words)? Corpus models don’t rule it out.

- But.. Modelling has limitations.

Green, C. (forthcoming, 2022). Extensive Reading of General Fiction and Incidentally Building a 9000 Word Vocabulary: Does Corpus Modelling Suggest it is Possible? *Reading in a Foreign Language*


Nation, P. (2014). How much input do you need to learn the most frequent 9,000 words?. *Reading in a Foreign Language, 26*(2), 1-16.
