Mid-frequency Vocabulary: The Great Vocabulary Gap

Norbert Schmitt
University of Nottingham
Frequency

- Frequency has long been recognized as a key characteristic of vocabulary
Harold Palmer (1877-1949)
Edward Thorndike (1874 – 1949)
Michael West (1888 – 1973)
Word Lists
Graded Readers
Frequency

• Frequency is a useful tool to inform pedagogy

• Focus on the most frequent vocabulary

• High-frequency vocabulary
1960s – 1980s
Paul Nation
High-frequency Vocabulary

- Pedagogy as a cost/benefit analysis
- High-frequency vocabulary is essential for any language use
- Teach high-frequency vocabulary by whatever means necessary
Low-frequency Vocabulary

- Low-frequency vocabulary is so rare that it offers minimal utility to learners
- Low-frequency vocabulary does not warrant explicit teaching
Low-frequency Vocabulary

• Teach learners *strategies* to cope with this vocabulary on their own

• Learners need to deal with this vocabulary independently
Low-frequency Vocabulary

• What is low-frequency vocabulary?

• Not really defined

• Everything that is not high-frequency?
High / Low frequency Vocabulary

Hi-frequency vocabulary

Everything else = Low-frequency vocabulary?
High / Low frequency Vocabulary

Hi-frequency vocabulary

Everything else = Low-frequency vocabulary?

2,000 word families

3,000-5,000 families to read
Nation (2006)

• Most previous vocabulary size requirement figures were based on 95% lexical coverage
• More recent reading research seems to show that 98% or higher is a more realistic figure
• Nation (2006) used the 98% figure to recalculate vocabulary size requirements:

  ➢ 6,000 - 7,000 word families for spoken discourse
  ➢ 8,000 - 9,000 word families for written discourse
High / Low frequency Vocabulary

Hi-frequency vocabulary

Everything else = Low-frequency vocabulary?

2,000 word families

6,000-9,000 families to read
High / Low frequency Vocabulary

Hi-frequency vocabulary

2,000
word families

Low-frequency vocabulary

9,000+
word families
High / Low frequency Vocabulary

Hi-frequency vocabulary: 2,000 word families

Low-frequency vocabulary: 9,000+ word families
“Mid-frequency” Vocabulary

Hi-frequency vocabulary

Mid-frequency vocabulary

Low-frequency vocabulary
Mid-frequency Vocabulary

2,000?  
Hi-frequency vocabulary

? - ?  
Mid-frequency vocabulary

9,000+ ?  
Low-frequency vocabulary
Mid-frequency Vocabulary

2,000?  Hi-frequency vocabulary

? - ?  Mid-frequency vocabulary

9,000+?  Low-frequency vocabulary
Mid-frequency Vocabulary

2,000?
Hi-frequency vocabulary

? - ?
Mid-frequency vocabulary

9,000+?
Low-frequency vocabulary

Explicitly teach

Strategies
Mid-frequency Vocabulary

2,000?  
Hi-frequency vocabulary

? - ?  
Mid-frequency vocabulary

9,000+ ?  
Low-frequency vocabulary

Explicitly teach

?

Strategies
Defining High-frequency Vocabulary

What are the limits of high-frequency vocabulary?
Defining High-frequency Vocabulary

- Frequency evidence
- Frequency and Acquisition/Use
- Dictionary defining vocabulary
- Word Lists
- Vocabulary in Graded Readers
- Outside Norms
Frequency Evidence

• A small number of word types occur very frequently

• A very large number of types occur very rarely
Vocabulary Size and Text Coverage in the Brown Corpus

![Bar chart showing vocabulary size and text coverage percentage. The x-axis represents different coverage percentages (80%, 90%, 95%, 97%), and the y-axis represents vocabulary size from 0 to 16000. The chart shows an increase in vocabulary size as coverage increases, with the highest vocabulary size at 97% coverage.]
### Adolphs and Schmitt (2003)

<table>
<thead>
<tr>
<th>Word families</th>
<th>Schonell <em>et al.</em></th>
<th>CANCODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>89</td>
<td>71.22</td>
<td>71.96</td>
</tr>
<tr>
<td>145</td>
<td>78.69</td>
<td>77.23</td>
</tr>
<tr>
<td>209</td>
<td>83.44</td>
<td>80.60</td>
</tr>
<tr>
<td>451</td>
<td>91.21</td>
<td>86.57</td>
</tr>
<tr>
<td>674</td>
<td>94.22</td>
<td>89.23</td>
</tr>
<tr>
<td>990</td>
<td>96.38</td>
<td>91.52</td>
</tr>
<tr>
<td>1,281</td>
<td>97.48</td>
<td>92.85</td>
</tr>
<tr>
<td>1,623</td>
<td>98.31</td>
<td>93.93</td>
</tr>
<tr>
<td>2,000</td>
<td>---</td>
<td>94.76</td>
</tr>
<tr>
<td>2,279</td>
<td>99.17</td>
<td>95.20</td>
</tr>
<tr>
<td>2,500</td>
<td>---</td>
<td>95.48</td>
</tr>
<tr>
<td>3,000</td>
<td>---</td>
<td>95.91</td>
</tr>
</tbody>
</table>
## Frequency and Coverage

<table>
<thead>
<tr>
<th>Levels</th>
<th>Approximate written coverage (%)</th>
<th>Approximate spoken coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st 1,000</td>
<td>78–81</td>
<td>81–84</td>
</tr>
<tr>
<td>2nd 1,000</td>
<td>8–9</td>
<td>5–6</td>
</tr>
<tr>
<td>3rd 1,000</td>
<td>3–5</td>
<td>2–3</td>
</tr>
<tr>
<td>4th–5th 1,000</td>
<td>3</td>
<td>1.5–3</td>
</tr>
<tr>
<td>6th–9th 1,000</td>
<td>2</td>
<td>0.75–1</td>
</tr>
<tr>
<td>10th–14th 1,000</td>
<td>&lt;1</td>
<td>0.5</td>
</tr>
<tr>
<td>Proper nouns</td>
<td>2–4</td>
<td>1–1.5</td>
</tr>
<tr>
<td>Not in the lists</td>
<td>1–3</td>
<td>1</td>
</tr>
</tbody>
</table>

Nation (2006)
## Frequency and Coverage

<table>
<thead>
<tr>
<th>Levels</th>
<th>Approximate written coverage (%)</th>
<th>Approximate spoken coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st 1,000</td>
<td>78–81</td>
<td>81–84</td>
</tr>
<tr>
<td>2nd 1,000</td>
<td>8–9</td>
<td>5–6</td>
</tr>
<tr>
<td>3rd 1,000</td>
<td>3–5</td>
<td>2–3</td>
</tr>
<tr>
<td>4th–5th 1,000</td>
<td>3</td>
<td>1.5–3</td>
</tr>
<tr>
<td>6th–9th 1,000</td>
<td>2</td>
<td>0.75–1</td>
</tr>
<tr>
<td>10th–14th 1,000</td>
<td>&lt;1</td>
<td>0.5</td>
</tr>
<tr>
<td>Proper nouns</td>
<td>2–4</td>
<td>1–1.5</td>
</tr>
<tr>
<td>Not in the lists</td>
<td>1–3</td>
<td>1</td>
</tr>
</tbody>
</table>

Nation (2006)
Frequency and Acquisition

- Cobb (2007)
- Learners need to meet words 6+ times for acquisition to occur
- “after the most frequent 1,000 items, family ranks tend to thin quite rapidly, and with them the learning opportunities”
- In natural reading settings, adequate repetition is unlikely to occur for words above the 2,000 word frequency level
Frequency and Use (Oral)

• Growing consensus that 98% lexical coverage is reasonable for good reading comprehension
• But listening requires less
• Van Zeeland and Schmitt (under review) found that L1 listeners can comprehend oral stories quite well with only 90% coverage
• L2 learners can as well, but with great variation
• At 95%, comprehension is good, but with much less variation
• 95% coverage is adequate for listening comprehension
Frequency and Use (Oral)

- 95% coverage requires a vocabulary size of between 2,000-3,000 families

- based on:
  - CANCODE (Adolphs & Schmitt, 2003)
  - Wellington Corpus of Spoken English (Nation 2006)
  - 88 television programs (Webb & Rodgers, 2009a)
  - 318 film scripts (Webb & Rodgers, 2009b)
Defining Vocabulary in Learner Dictionaries

- Longman ≈ 2,000
- Oxford ≈ 3,000
- Macmillan ≈ 2,500
Word Lists

- General Service List ≈ 2,000
- Longman Communication 3,000
- Oxford 3,000
- Bare Naked Lexis (BNL) (Steve Neufeld/Ali Billuroğlu) ≈ 2,700
Vocabulary in Graded Readers

- Graded Reader Series: 200 – 3600 headwords

- Most series top out by 3,000 headwords

- Cambridge: 400 – 3,800 headwords

- Macmillan: 300 – 2,200 headwords

- Penguin: 200 – 3,000 headwords
Defining High-frequency Vocabulary

• The traditional boundary of high-frequency (2,000) seems too low
Defining High-frequency Vocabulary

• The traditional boundary of high-frequency (2,000) seems too low

• On balance, it seems that 3,000 word families is a more pedagogically-useful definition
Defining Low-frequency Vocabulary

Where does low-frequency vocabulary begin?
Defining Low-frequency Vocabulary

• Paul Nation suggests that 5,000 word families is the end of ‘general’ English vocabulary

• Beyond this, vocabulary tends to become more context-specific

• 5,000+ = Low-frequency vocabulary?
Defining Low-frequency Vocabulary

- High-frequency + technical vocabulary?
- Fraser (2011)
  - 360,000 word pharmacology corpus
Defining Low-frequency Vocabulary

• Low-frequency vocabulary has traditionally been thought to start at around the 10,000 frequency level
Defining Low-frequency Vocabulary

• Low-frequency vocabulary has traditionally been thought to start at around the 10,000 frequency level

• ‘Round number’ effect
Defining Low-frequency Vocabulary

- 10,000 is the top level of the *Vocabulary Levels Test*, as popularized by Nation’s (1990) book
Defining Low-frequency Vocabulary

- Hazenberg and Hulstijn (1996)
- **10,000** base words to study at university in Dutch
- Transferred to English
Defining Low-frequency Vocabulary

- Nation (2006)
  - 6,000 - 7,000 word families for spoken discourse
  - 8,000 - 9,000 word families for written discourse
Defining Low-frequency Vocabulary

- Nation (2006)

  - 6,000 - 7,000 word families for spoken discourse

  (Van Zeeland & Schmitt: 2,000 - 3,000)
Defining Low-frequency Vocabulary

- Nation (2006)
  - 8,000 - 9,000 word families for written discourse
Defining Low-frequency Vocabulary

- Nation (2006)

- 8,000 - 9,000 word families for written discourse

- 9,000 allows widespread engagement with English
Defining Low-frequency Vocabulary

• If 9,000 word families allow widespread engagement with English, then should everything beyond this be considered low-frequency vocabulary?
Defining Low-frequency Vocabulary

• Cost/benefit: Low-frequency vocabulary is so rare that it offers minimal utility to learners

• Low-frequency = 9,000+
High / Mid / Low frequency Vocabulary

3,000 Hi-frequency vocabulary
3,000 – 9,000 Mid-frequency vocabulary
9,000+ Low-frequency vocabulary
Pedagogical Implications

- High-frequency vocabulary (3,000) should be taught with focused attention both in the classroom and in independent study.
- Low-frequency vocabulary (9,000+) should be handled independently by learners, by using strategies which have been taught and practiced.
Pedagogical Problem

• But what about **mid-frequency** vocabulary?
• Learners need to know much of this vocabulary, especially in written contexts
• The field only has a solid idea of how to teach high-frequency vocabulary
• This leaves up to **6,000** word families of mid-frequency vocabulary items which learners need but which are not dealt with in any principled way
Pedagogical Problem

• How can we best help learners to master this large range of mid-frequency vocabulary?
Pedagogical Problem

• One of the ‘big questions’ for vocabulary research and pedagogy in the next decade