Associating Difficulty in Near-Synonymy Choice with Types of Nuance using Core Vocabulary

Tong Wang and Graeme Hirst
Computer Science, University of Toronto
Supported by the Natural Sciences and Engineering Research Council of Canada
Outline

**Associating Difficulty in Near-Synonymy Choice with Types of Nuance using Core Vocabulary**
Outline

Associating Difficulty in Near-Synonymy Choice with Types of Nuance using Core Vocabulary
Choosing between near-synonyms

The ________ was magnified when the Army failed to charge the standard percentage rate for packing and handling.
Evaluating near-synonym choice

• How to evaluate a lexical choice process for near-synonyms?

• Edmonds 1997: The fill-in-the-blanks task:
  — Does the system’s choice match that of the original human author?

The error was magnified when the Army failed to charge the standard percentage rate for packing and handling. However, such a move also would have run the risk of cutting deeply into U.S. economic growth, which is why some economists think it would be a big mistake. The oversight was magnified when the Army failed to charge the standard percentage rate for packing and handling.
Fill in the blanks

- Edmonds: Seven near-synonym clusters, set of *Wall Street Journal* sentences for each.
  - *difficult, hard, tough* (6665 sentences)
  - *error, mistake oversight* (1030 sentences)
  - *job, task, duty* (5402 sentences)
  - *responsibility, burden, obligation, commitment* (3138 sentences)
  - *material, stuff, substance* (1828 sentences)
  - *give, provide, offer* (10204 sentences)
  - *settle, resolve* (1568 sentences)

- Subsequently used by a number of researchers.
Interpreting performance

- Ideally, system always matches writer’s original choice.
- But humans cannot perform that well.
- Difficulty depends on
  - Relative absence of syntactic or collocational constraints
  - Number of alternatives
  - Closeness of meaning among alternatives
Human performance on FITB test

These near-synonym clusters vary in difficulty of fill-in-the-blanks for humans

Outline

Associating Difficulty in Near-Synonymy Choice with Types of Nuance using Core Vocabulary
Near-synonymy

- Any pair of near-synonyms differs on one or more dimension.
## Near-synonymy

<table>
<thead>
<tr>
<th>Near-synonyms</th>
<th>Dimension of variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>drunk, inebriated</td>
<td>Formality</td>
</tr>
<tr>
<td>slender, skinny</td>
<td>Attitude</td>
</tr>
<tr>
<td>error, mistake, blunder, slip</td>
<td>Abstractness and strength; blameworthiness</td>
</tr>
<tr>
<td>seep, drip</td>
<td>Continuous / intermittent</td>
</tr>
<tr>
<td>enemy, foe</td>
<td>Emphasis on fighting or hatred</td>
</tr>
</tbody>
</table>

Near-synonymy — Denotational

• Set of dimensions of differentiation is open-ended (infinite); includes arbitrary aspects of denotation.
  — Blameworthiness
  — Enmity

• But many denotational dimensions recur:
  — Magnitude or strength
  — Continuous / intermittent
  — Intentional / accidental
  — …
Near-synonymy — Connotational

• Connotational / pragmatic dimensions relate to style and evaluation.
  — Formality
  — Floridity
  — Euphemism
  — Abstractness
  — Force
  — Slant
  — …


Outline

Associating Difficulty in Near-Synonymy Choice with Types of Nuance using Core Vocabulary
Core vocabulary

• Intuitively: The most basic or central words of a language.

• Carter 1998: Ten lexical properties that indicate or correlate with ‘coreness’.

Properties of core vocabulary

- Acts as defining vocabulary for other words.
  - e.g., *dine* can be defined in terms of *eat*, but not *eat* in terms of *dine*.

- Can substitute for other words as more-general terms.
  - e.g., *eat* can substitute for *dine* (but *ingest* cannot).

Examples from Carter (1998) except *ingest*. 

Example: *dine* can be defined in terms of *eat*, but not *eat* in terms of *dine*.

Example: *eat* can substitute for *dine* (but *ingest* cannot).
Clear antonymy

- fat – thin; laugh – cry

- emaciated – ??; guffaw – ??

Properties of core vocabulary

- Many collocations ("collocability").

<table>
<thead>
<tr>
<th>Core</th>
<th>radiant ...</th>
<th>gaudy ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>bright</td>
<td>light</td>
<td>*light</td>
</tr>
<tr>
<td>light</td>
<td>light</td>
<td>*light</td>
</tr>
<tr>
<td>idea</td>
<td>*idea</td>
<td>*idea</td>
</tr>
<tr>
<td>colours</td>
<td>colours</td>
<td>colours</td>
</tr>
<tr>
<td>red</td>
<td>red</td>
<td>*red</td>
</tr>
<tr>
<td>future</td>
<td>*future</td>
<td>*future</td>
</tr>
<tr>
<td>child</td>
<td>child</td>
<td>*child</td>
</tr>
<tr>
<td>sun</td>
<td>?sun</td>
<td>*sun</td>
</tr>
</tbody>
</table>

Properties of core vocabulary

- Neutrality on Osgood scales:
  - Evaluation: positive ↔ negative
  - Potency: strong ↔ weak

- Neutrality of tenor:
  - Formality: formal ↔ informal

Core vocabulary and connotation

Connotational dimensions

floridity
euphemism
slant
force
formality
potency
evaluation

Carter’s scales
Core vocabulary and near-synonyms

Near-synonym differences

Connotational dimensions

floridity

euphemism

slant

force

formality

potency

evaluation

Carter's scales

Core vocabulary
Quantifying coreness

• As defining vocabulary

  \[ \text{defvoc}(w) = \text{frequency of } w \text{ in the defining vocabulary} \]

• Counted word frequencies in all definitions in the *Macquarie Dictionary*. 

Quantifying collocability

- Collocability defined as entropy $H$.

$$colent(w) = H(\text{distribution of } w\text{'s collocations})$$

$$= \sum_{w_i \in \mathcal{V}} p(w_i) \log \frac{1}{p(w_i)}$$

- Greater entropy means
  - greater collocability (more collocates);
  - more-even distribution among collocates.
Quantifying collocability

• Collocation distribution calculated from bigram collocates of $w$ in the British National Corpus.
Quantifying collocability

Simplified example of collocability as entropy:

<table>
<thead>
<tr>
<th></th>
<th>bright ...</th>
<th>radiant ...</th>
<th>gaudy ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>light</td>
<td>light – .20</td>
<td>light – .33</td>
<td>light – .10</td>
</tr>
<tr>
<td>smile</td>
<td>smile – .20</td>
<td>smile – .33</td>
<td>smile – .10</td>
</tr>
<tr>
<td>color</td>
<td>color – .20</td>
<td>color – .33</td>
<td>color – .80</td>
</tr>
<tr>
<td>future</td>
<td>future – .20</td>
<td>future – .00</td>
<td>future – .00</td>
</tr>
<tr>
<td>child</td>
<td>child – .20</td>
<td>child – .00</td>
<td>child – .00</td>
</tr>
</tbody>
</table>

\[
H_1 = \sum_{i=1}^{5} 0.2 \times \log \frac{1}{0.2} = \log 5 \approx 0.6990
\]

\[
H_2 = \sum_{i=1}^{3} 0.33 \times \log \frac{1}{0.33} = \log 3 \approx 0.4771
\]

\[
H_3 = 2 \times 0.1 \times \log \frac{1}{0.1} + 0.8 \times \log \frac{1}{0.8} \approx 0.2 + 0.0775 = 0.2775
\]
Associating Difficulty in Near-Synonymy Choice with Types of Nuance using Core Vocabulary
Results

Correlation between average human accuracy and max defining vocabulary score, $\rho = -0.68$. 

Human performance

<table>
<thead>
<tr>
<th></th>
<th>difficult</th>
<th>error</th>
<th>give</th>
<th>job</th>
<th>material</th>
<th>responsibility</th>
<th>settle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.73</td>
<td>0.09</td>
<td>0.16</td>
<td>0.20</td>
<td>0.18</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>Defining</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vocabulary score</td>
<td>0.49</td>
<td>0.84</td>
<td>0.69</td>
<td>0.92</td>
<td>0.78</td>
<td>0.79</td>
<td>0.85</td>
</tr>
</tbody>
</table>
Results

Correlation between average human accuracy and max collocability score, $\rho = -.54$
Results

• Near-synonym clusters differ in coreness (by two key measures of coreness).

• Coreness and difficulty in choice are correlated.
Moreover, …

• Core-vocabulary words are neutral in style.

• Near-synonyms with stylistic variations are easier to differentiate.

• Coreness and difficulty in choice are positively related.
Conclusions

• Near-synonym clusters are more difficult to differentiate if they contain words that are more-core.

• Core vocabulary is a promising concept in characterizing near-synonym differences.
Concurrent work

- Characterizing difficulty of near-synonym lexical choice.
  - Relating difficulty to latent semantic space dimensionality (Wang and Hirst 2010).
  - Characterization of subtlety of near-synonym differentiating nuances ($\propto$ difficulty) as those residing in the higher dimensions of the latent semantic space.

Future work

- Characterizing difficulty of near-synonym lexical choice.
  - Obtaining human judgement on difficulty / subtlety among near-synonym sets.
  - Building larger data sets for both automated system and human judgement.
Future work

• Characterizing core vocabulary:
  — Summary: Statistical measures of word frequency in titles, abstracts, and opening and closing sentences of paragraphs.
  — Antonymy: Use Mohammad and Hirst’s antonym lists.
Future work

- Characterizing core vocabulary:
  - Meaning extensibility:
    
    a bright spark; bright and early; brighten up
  
  - Syntactic variation:
    
    give the book to them / give them the book
    donate the book to them / *donate them the book