

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

Tong Wang and Graeme Hirst
Computer Science, University of Toronto



Tong Wang

Supported by the Natural Sciences and
Engineering Research Council of Canada



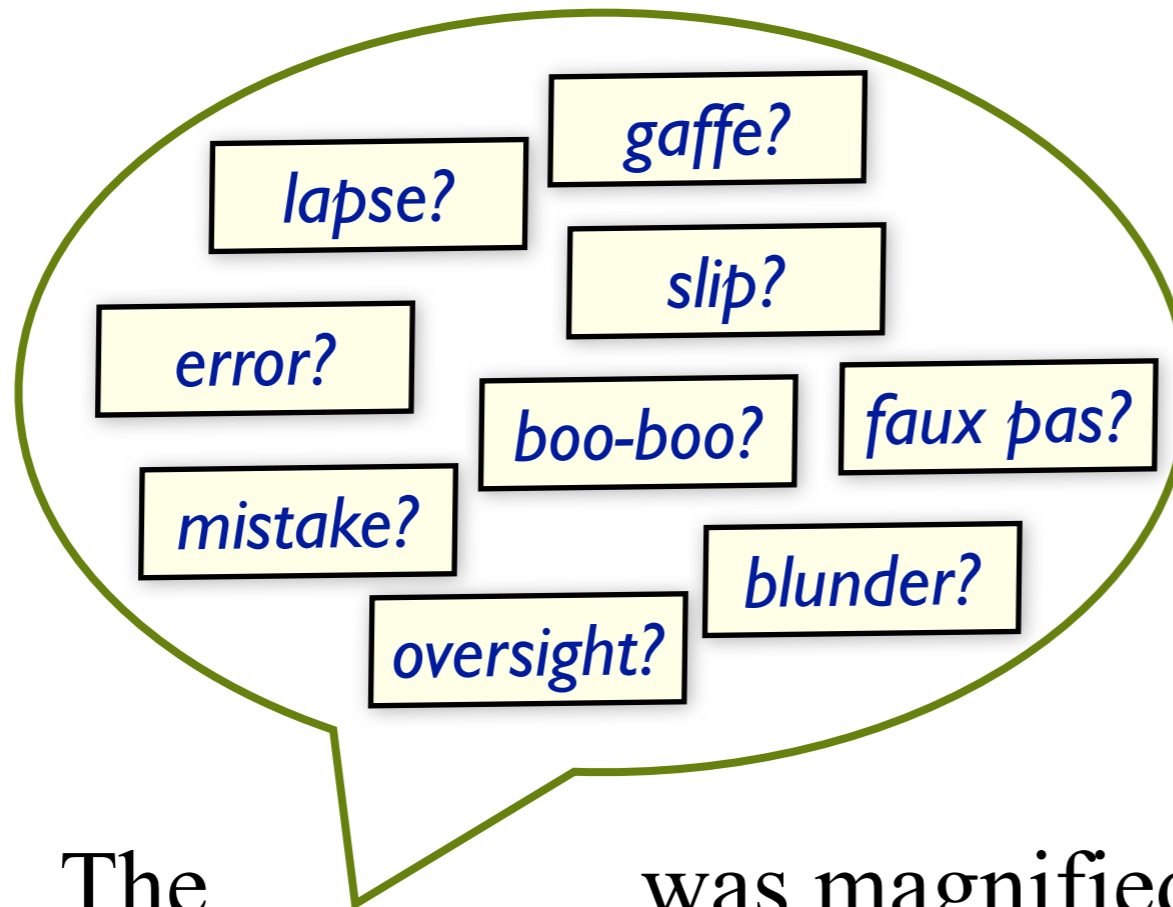
Outline

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?

Edmonds, Philip. 1997. Choosing the word most typical in context using a lexical co-occurrence network. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics*, 507–509, Madrid.

Fill in the blanks



However, such a move also would have dug us
deeply into the economic growth that has been
economic growth would be a big _____

error?

mistake?

oversight?

error?

mistake?

oversight?

The _____ was magnified when the Army failed to charge
the standard percentage rate for packing and handling.

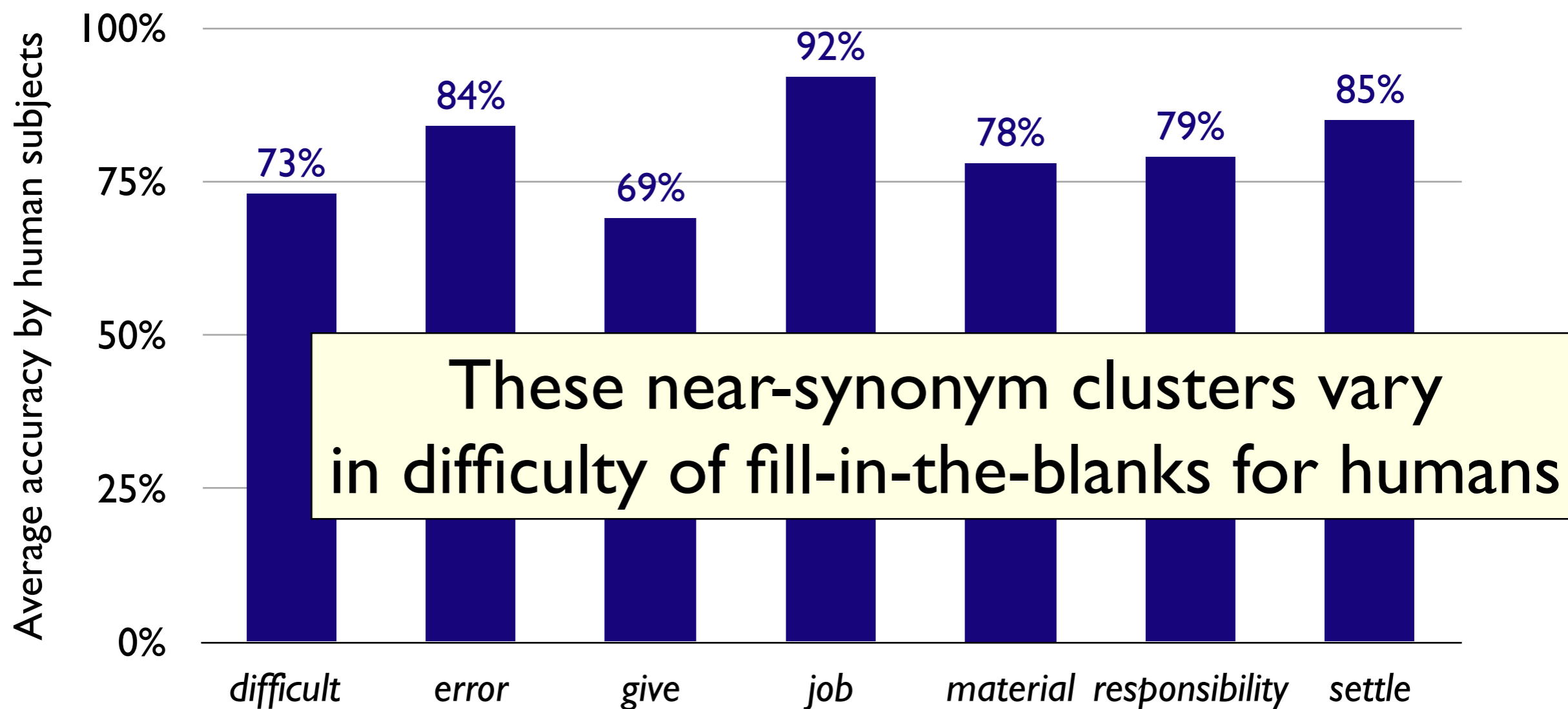
Fill in the blanks

- 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



Inkpen, Diana. 2007. A statistical model for near-synonym choice. *ACM Transactions on Speech and Language Processing*, 4, 1–17.

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

Near-synonyms **Dimension of variation**

drunk, inebriated

Formality

slender, skinny

Attitude

*error, mistake, blunder,
slip*

Abstractness and strength;
blameworthiness

seep, drip

Continuous / intermittent

enemy, foe

Emphasis on fighting or hatred

⋮

⋮

DiMarco, Chrysanne; Hirst, Graeme; and Stede, Manfred. 1993. The semantic and stylistic differentiation of synonyms and near-synonyms. *AAAI Spring Symposium on Building Lexicons for Machine Translation*, 114–121

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
 - ...

Stede, Manfred. 1993. Lexical choice criteria in language generation. *Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, 454–459.

Hovy, Eduard. 1988. *Generating Natural Language Under Pragmatic Constraints*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Not core

Core

— 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*.

Properties of core vocabulary

- Clear antonymy

— *fat* – *thin*; *laugh* – *cry* ← Core

— *emaciated* – ??; *guffaw* – ?? ← Not core

Examples from Carter (1998).

Properties of core vocabulary

- Many collocations (“collocability”).

	<i>bright ...</i>	<i>radiant ...</i>	<i>gaudy ...</i>
Core	<i>light</i>	<i>light</i>	* <i>light</i>
	<i>idea</i>	* <i>idea</i>	* <i>idea</i>
	<i>colours</i>	<i>colours</i>	<i>colours</i>
	<i>red</i>	<i>red</i>	* <i>red</i>
	<i>future</i>	* <i>future</i>	* <i>future</i>
	<i>child</i>	<i>child</i>	* <i>child</i>
	<i>sun</i>	? <i>sun</i>	* <i>sun</i>
	⋮	⋮	⋮

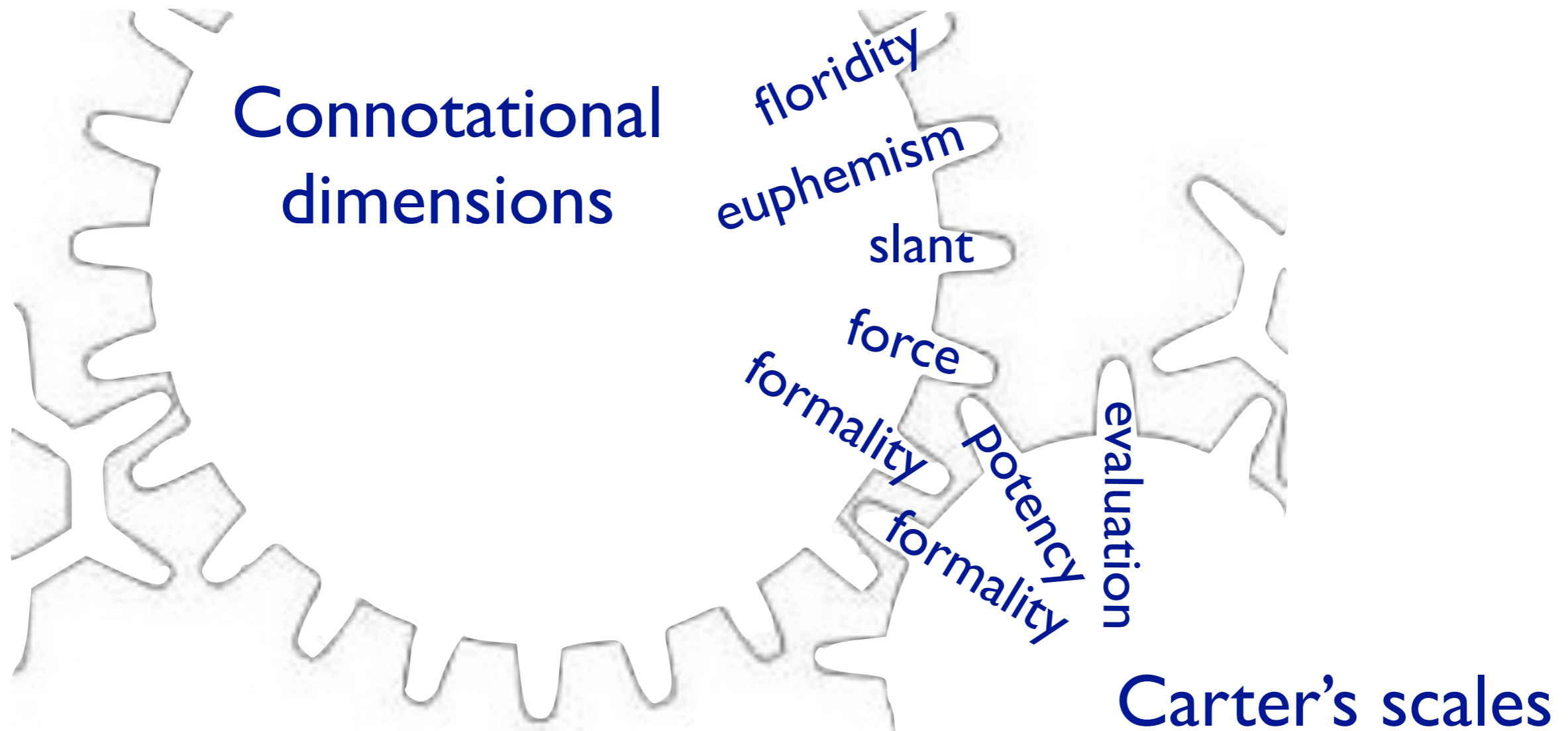
Examples from Carter (1998).

Properties of core vocabulary

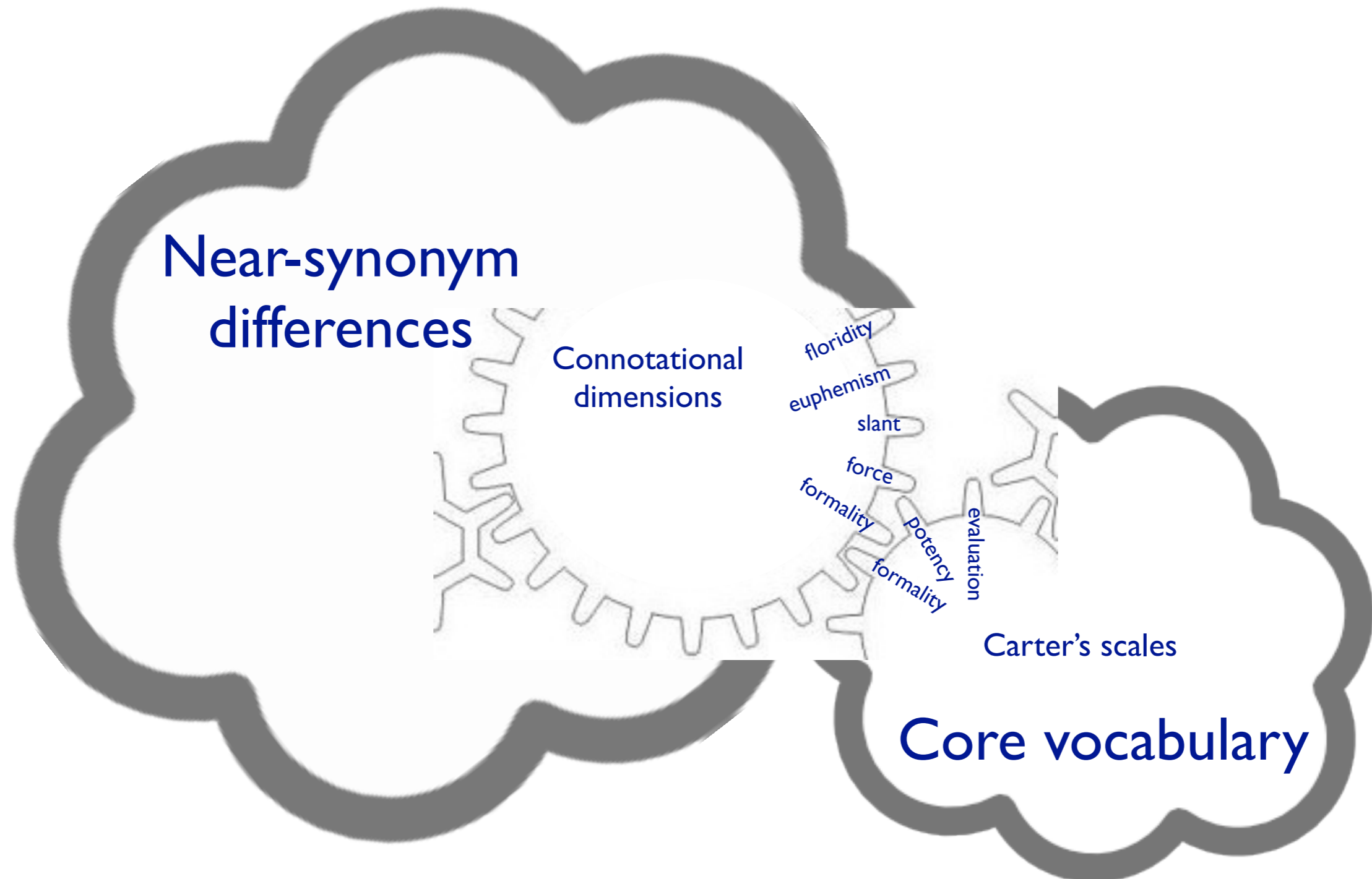
- Neutrality on Osgood scales:
 - Evaluation: positive ↔ negative ← ≡ Slant
 - Potency: strong ↔ weak ← ≡ Force
- Neutrality of tenor:
 - Formality: formal ↔ informal

Osgood, Charles E.; Suci, George; and Tannenbaum, Percy. 1957. *The Measurement of Meaning*. University of Illinois Press.

Core vocabulary and connotation



Core vocabulary and near-synonyms



Quantifying coreness

- As defining vocabulary
 - $defvoc(w)$ = frequency of w in the defining vocabulary
- Counted word frequencies in all definitions in the *Macquarie Dictionary*.

Quantifying collocability

- Collocability defined as entropy H .

$$\begin{aligned} \text{colent}(w) &= H(\text{distribution of } w\text{'s collocates}) \\ &= \sum_{w_i \in \mathcal{V}} p(w_i) \log \frac{1}{p(w_i)} \end{aligned}$$

- 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:

bright ...

light – .20

smile – .20

color – .20

future – .20

child – .20

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

radiant ...

light – .33

smile – .33

color – .33

future – .00

child – .00

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

gaudy ...

light – .10

smile – .10

color – .80

future – .00

child – .00

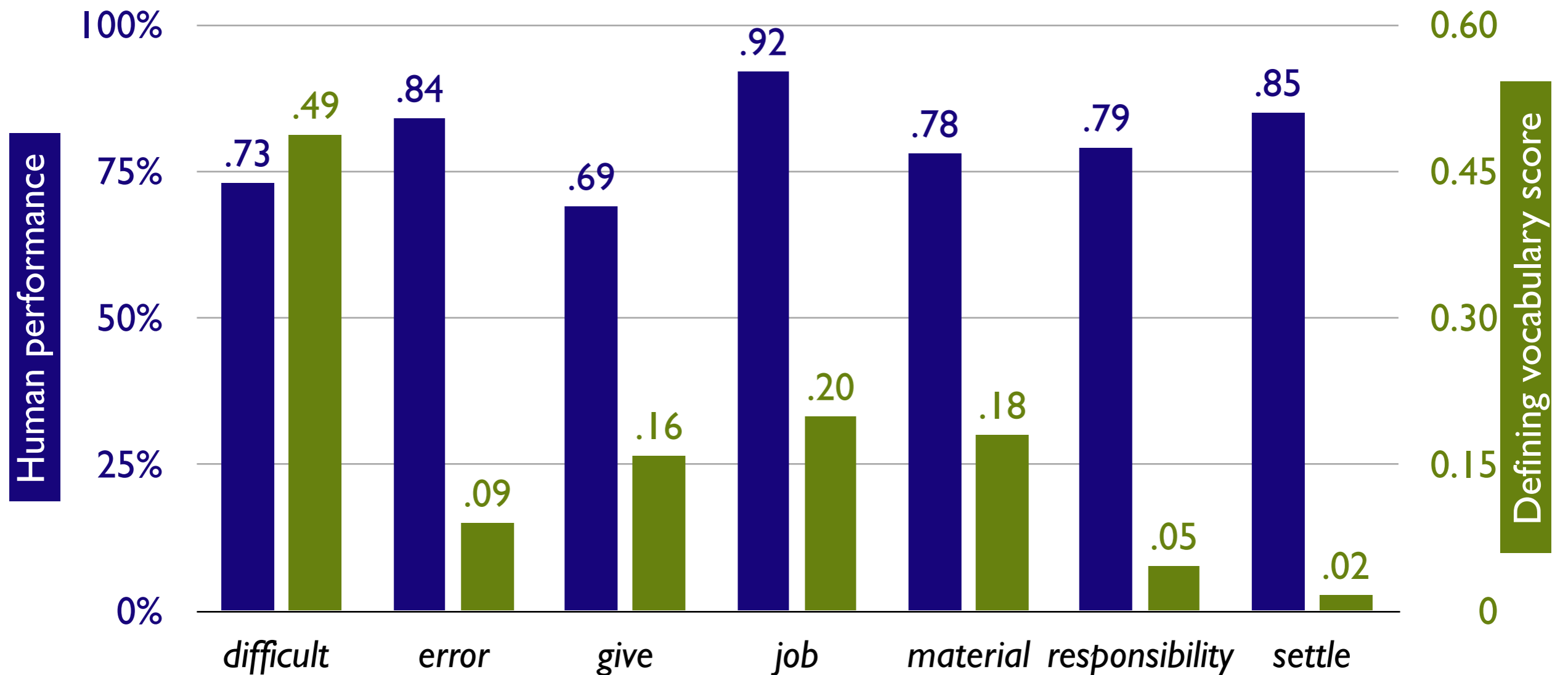
$$H_3 = 2 \times 0.1 \times \log \frac{1}{0.1}$$
$$+ 0.8 \times \log \frac{1}{0.8}$$
$$\approx 0.2 + 0.0775 = \mathbf{0.2775}$$

Outline

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

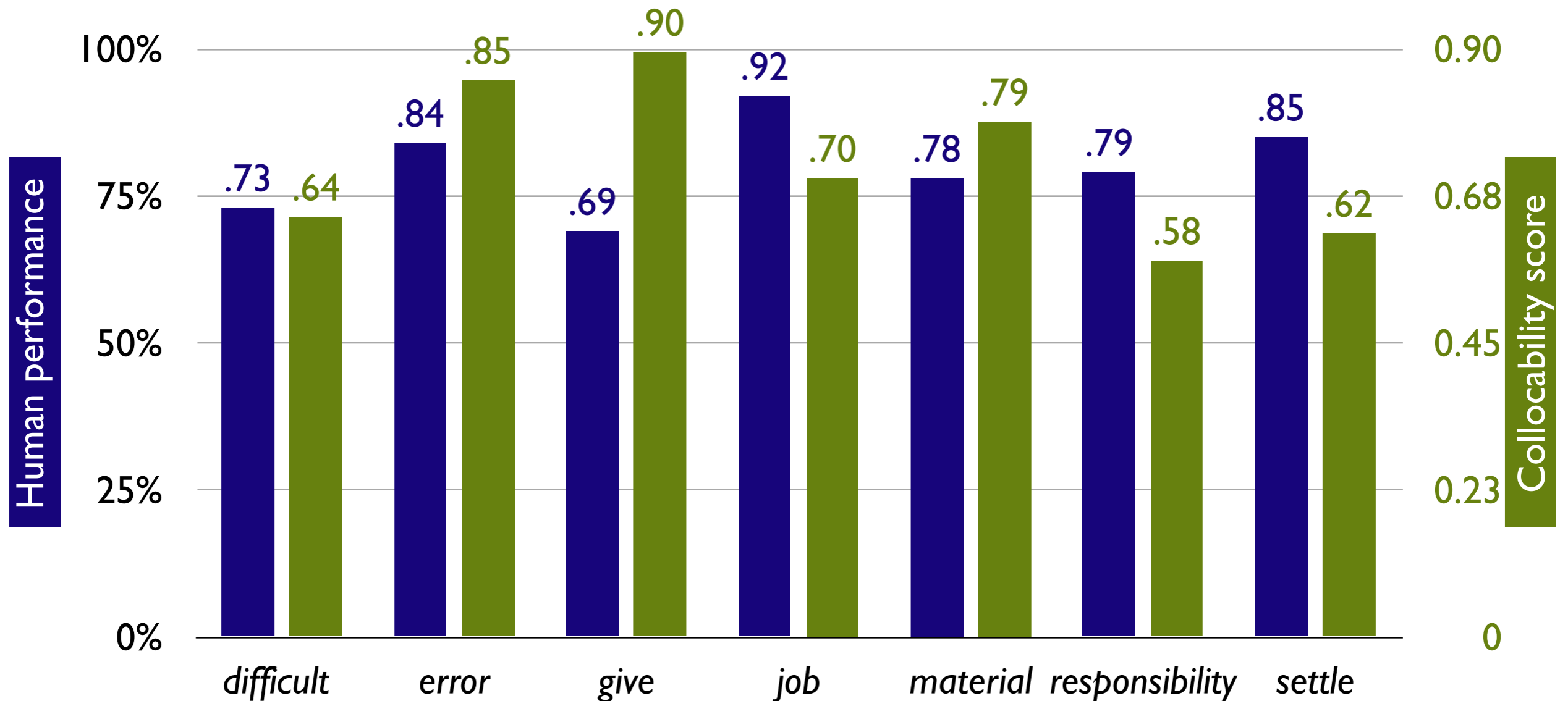
Results

Correlation between average human accuracy and max defining vocabulary score, $\rho = -.68$.



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.

Wang, Tong and Hirst, Graeme. 2010. Near-synonym lexical choice in latent semantic space. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010)*, Beijing.

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*