Assessing the Limits of the Distributional Hypothesis in Semantic Spaces

Trait-based Relational Knowledge and the Impact of Co-occurrences

Mark Anderson and Jose Camacho Collados

Trait-based Relational Knowledge

Traits/attributes

- Traits commonly associated with concepts.
- Split into types (colour, shape&size, material, components, tactile)
- E.g. The ubiquitous yellow banana.

Datasets

- McRae[1]
- Norms[2]
- McRae (Spanish)

---

[1] Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. *Semantic feature production norms for a large set of living and nonliving things.*Behavior Research Methods, 37:547–559.

[2]Barry Devereux, Lorraine K. Tyler, Jeroen Geertzen,and Billi Randall. 2014. *The centre for speech, language and the brain (CSLB) concept property norms.* Behavior Research Methods, 46:1119 – 1127

Example table — McRae

	trait type	N_C		N_T	
	colour	148		7	green (32), brown (32), black (24), white (21), red (16), yellow (13), orange (10)
	components	110		6	handle (39), legs (19), wheels (14), leaves (14), seeds (13), doors (11)
	materials	144		4	metal (79), wood (43), cotton (11), leather (11)
	size & shape	234		4	small (83), large (70), long (44), round (37)
	tactile	117		7	heavy (21), soft (19), furry (18), sharp (17), hard (16), juicy (16), slimy (10)

Co-occurences

- Sentence — e.g. *The **bananas** were ubiquitous in that town, and **yellow** jackets were the outerwear of choice for the inhabitants.*

- Window (±5 tokens of concept) — e.g. *He wore a [**yellow** t-shirt while eating a **banana** in the garage of a] kind stranger.*

- Syntactic — e.g. *If **bananas** aren't **yellow**, nobody will be interested in eating them, except for perhaps some questionable folk.*

Experimental Details 1

Corpora

- **UMBC** 3.4B Tokens (English)
- **Wikipedia** Dump 2.5B/0.6B Tokens (English/Spanish)
- **ES1B** 1.4B Tokens (Spanish)

Removing Co-occurrences

- Use 80% of corpus as training data nd keep 20% as reserves.
- Sentences with co-occurrences removed from training data.
- Replaced with random sentence from reserves.

Experimental Details 2

Vector Space

- **CBOW Word2Vec**
- Faster than skip-gram.
- Trained models with and without co-occurrences.
- For each separate trait type (e.g. colour)

Classifiers

- Multi-class (predict **yellow** given **banana** as input, as embedding from model).
- Binary-class (predict **related** or **not related** with $e_{concept}$-$e_{relation.}$ as input).
- Used SVMs.

Results?

McRae (English) — Multi-class

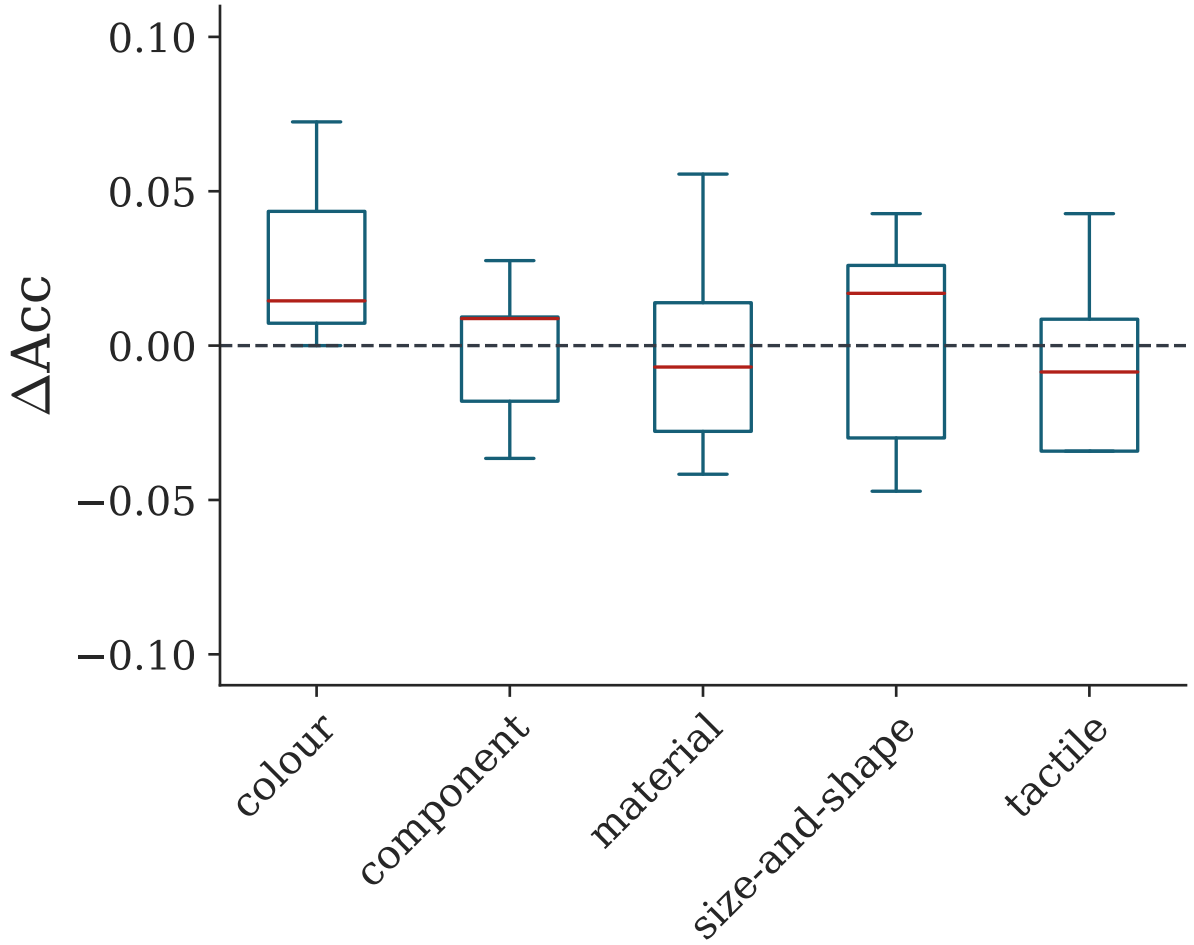

Figure 1: Distribution of multi-class accuracy differences between models trained with and without co-occurrences separated based on trait type. For McRae dataset and English corpora.

Norms (English) — Multi-class

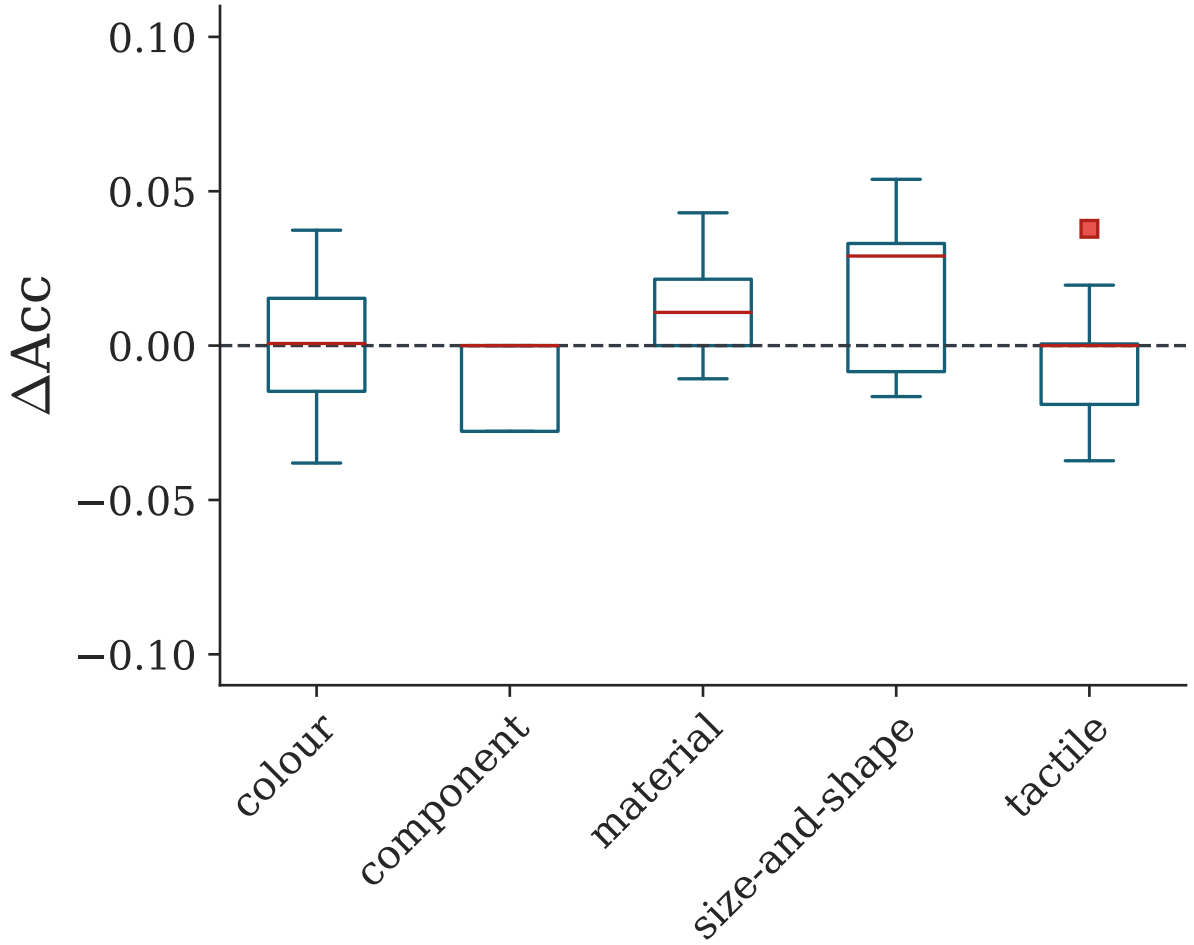

Figure 1: Distribution of multi-class accuracy differences between models trained with and without co-occurrences separated based on trait type. For Norms dataset and English corpora.

McRae (Spanish) — Multi-class

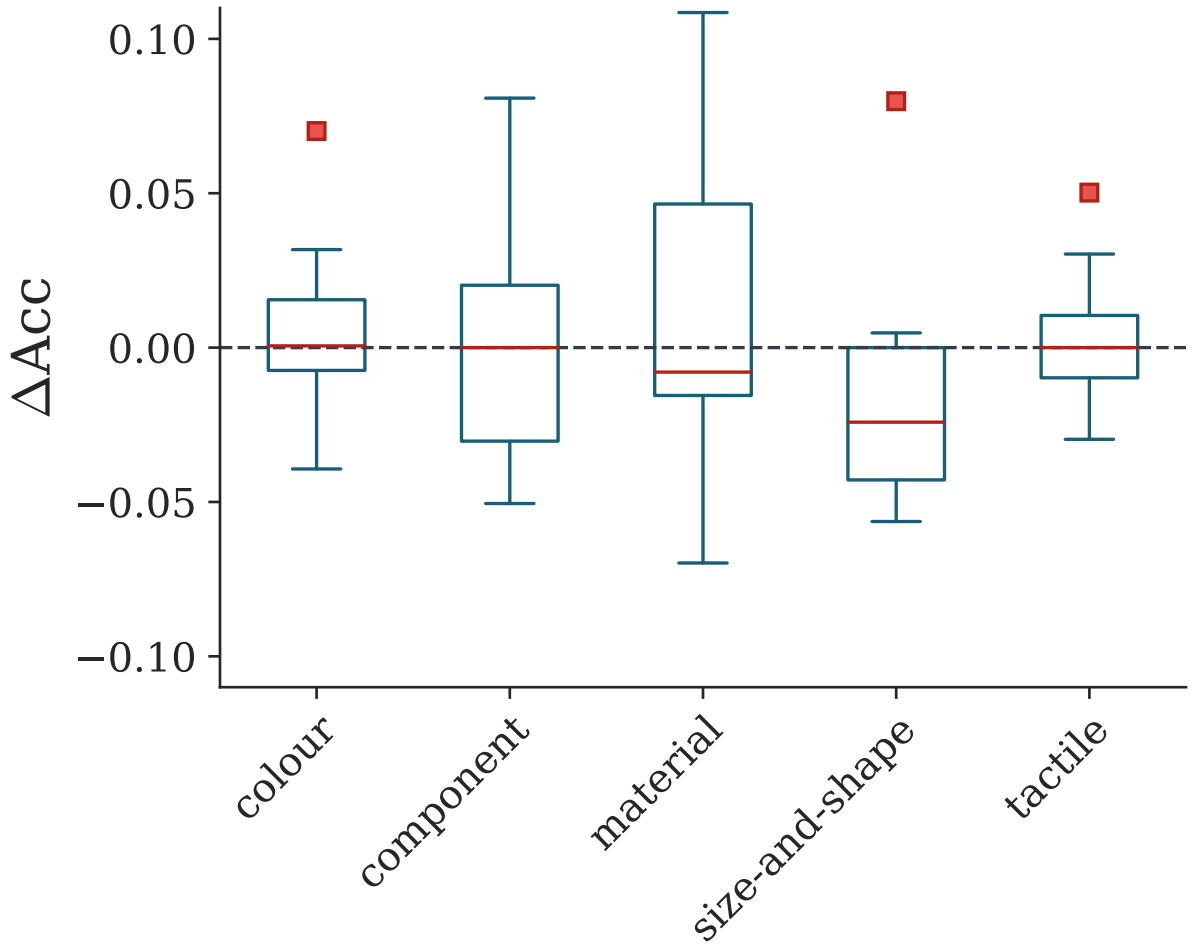

Figure 1: Distribution of multi-class accuracy differences between models trained with and without co-occurrences separated based on trait type. For McRae dataset and Spanish corpora.

McRae (English) — Multi-class

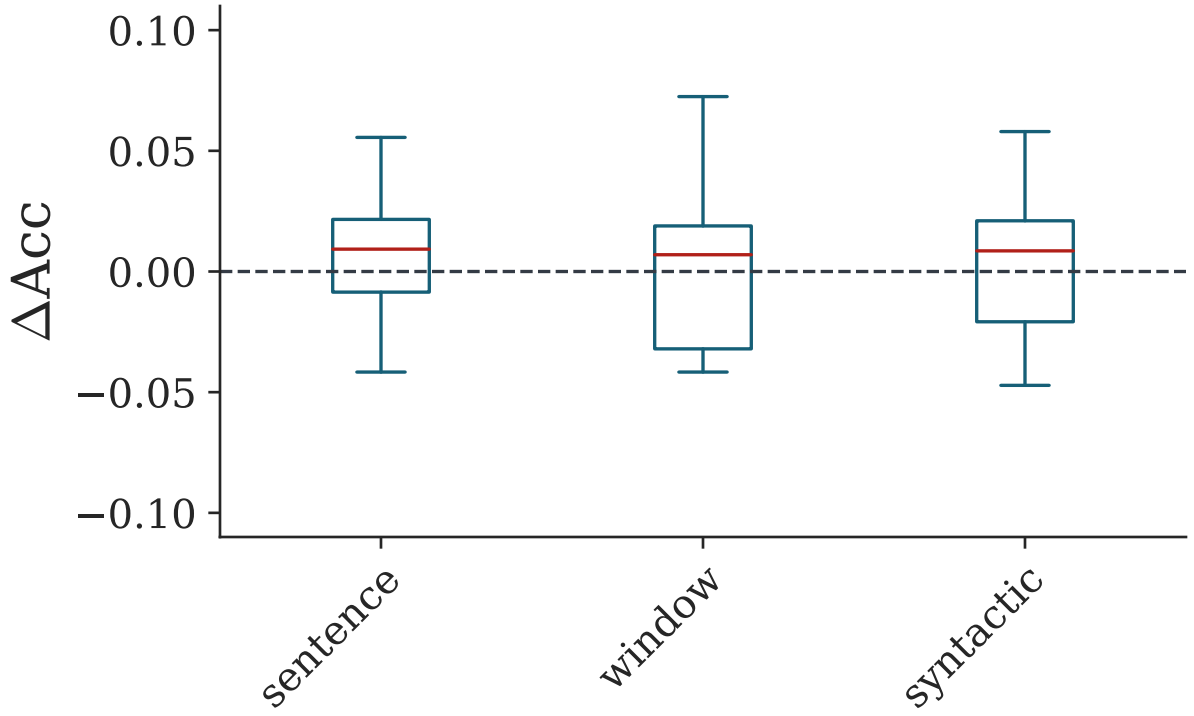

Figure 1: Distribution of multi-class accuracy differences between models trained with and without co-occurrences separated based on co-occurrence removal method. For McRae dataset and English corpora.

Norms (English) — Multi-class

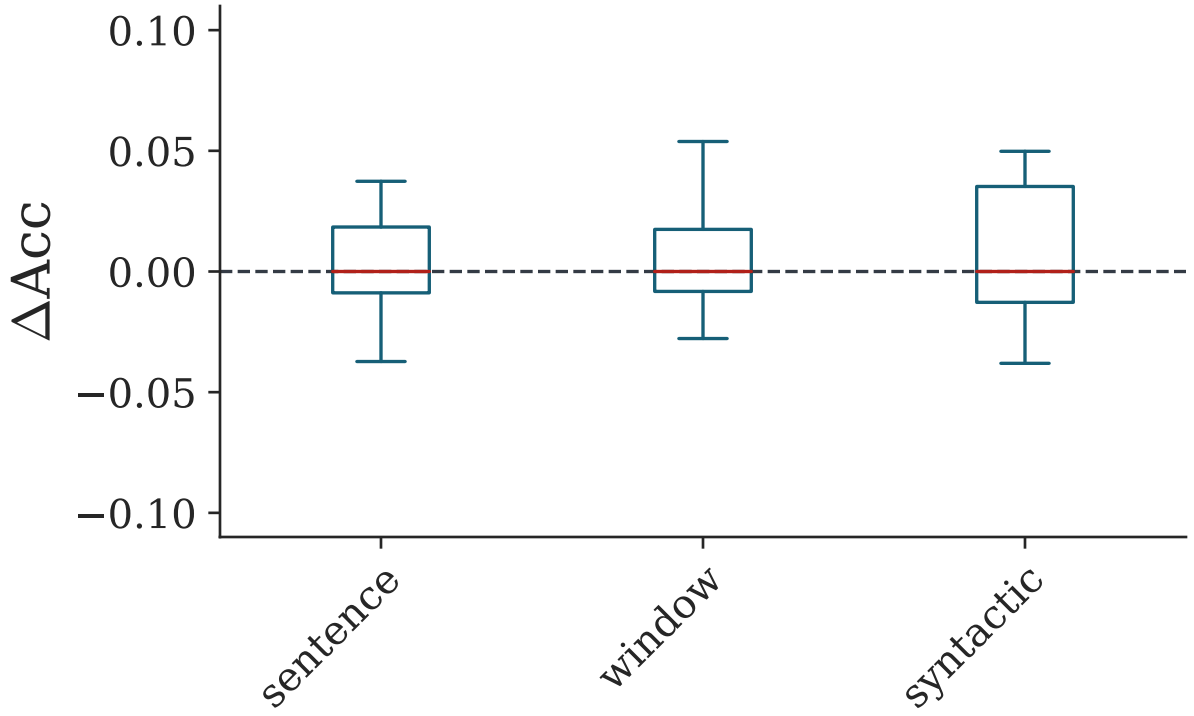

Figure 1: Distribution of multi-class accuracy differences between models trained with and without co-occurrences separated based on co-occurrence removal method. For Norms dataset and English corpora.

McRae (Spanish) — Multi-class

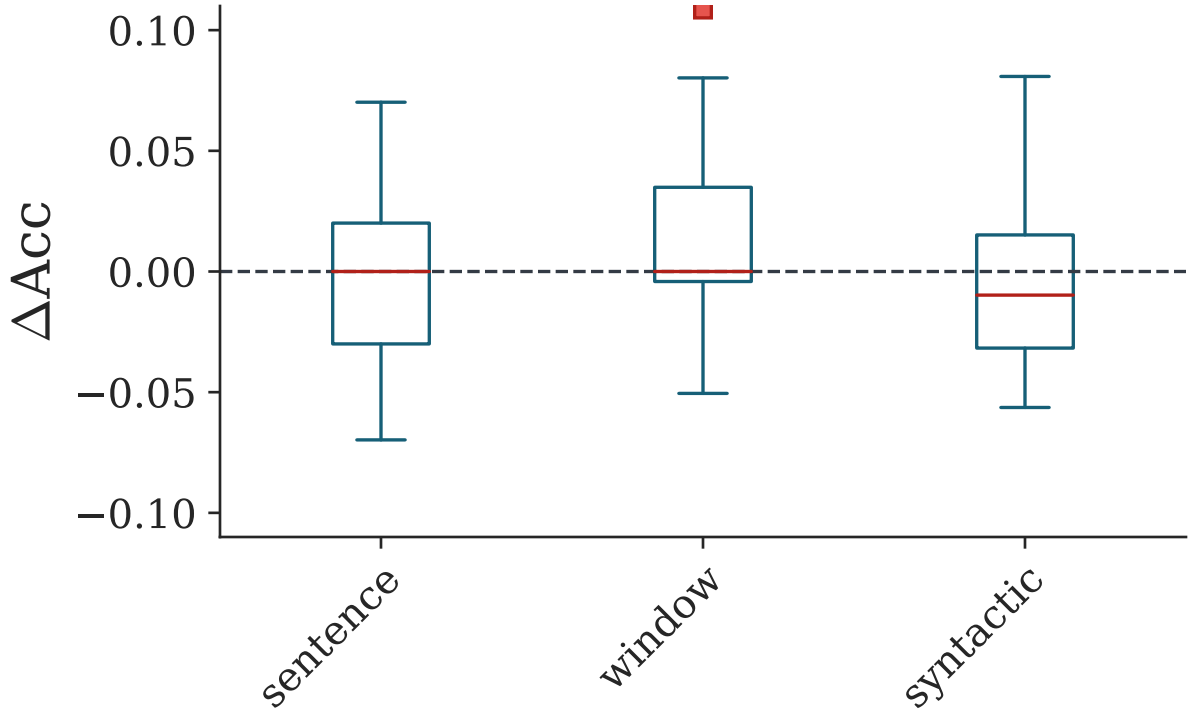

Figure 1: Distribution of multi-class accuracy differences between models trained with and without co-occurrences separated based on co-occurrence removal method. For McRae dataset and Spanish corpora.

Binary Results

- Typically quite high (80-90% accuracy)
- Require negative examples (clearly not trivial)
- Similar to multi-class (no real distinct patterns regarding removal of co-occurrences).

End

Thanks. Any questions, remarks, or criticisms?