ENCODING IMAGISM?
MEASURING LITERARY IMAGEABILITY, VISUALITY AND CONCRETENESS VIA MULTIMODAL WORD EMBEDDINGS

Yuri Bizzoni, Pascale Feldkamp, & Kristoffer L. Nielbo

AARHUS
UNIVERSITY
CENTER FOR HUMANITIES COMPUTING

July 4, 2025 | CCLS2025

CENTER FOR
HUMANITIES
COMPUTING

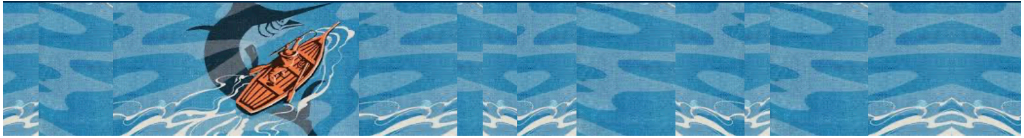
OUTLINE

1. Introduction
2. Approaches
3. Experiments
4. Conclusion

INTRODUCTION: HOW DOES A TEXT “BRING US THERE”?

“The boy took the old army blanket off the bed and spread it over the back of the chair and over the old man’s shoulders.”

– *The Old Man and the Sea*, Ernest Hemingway



INTRODUCTION: HOW DOES A TEXT "BRING US THERE"?

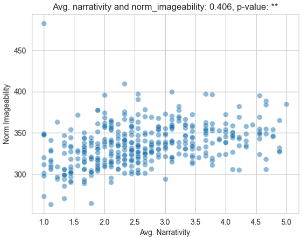
Computational literary studies already employ related measures such as: **imageability**, **concreteness**, **visuality**, abstractness, etc...

Imageability	ease with which a word evokes sensory experiences [Paivio et al., 1968]	
Visuality	extent to which a word is connected with the visual sense [Lynott et al., 2020]	subset
Concreteness	extent to which a word refers to a perceptible entity [Brysbaert et al., 2014]	overlap

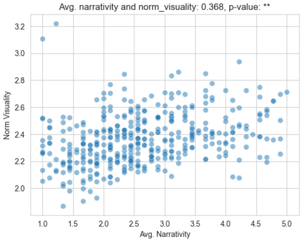
I.a. via MRC psycholinguistics Dictionary [Coltheart, 1981], Concreteness Dictionary [Brysbaert et al., 2014], Lancaster Sensori-motor Norms [Lynott et al., 2020]

INTRODUCTION: FEATURES

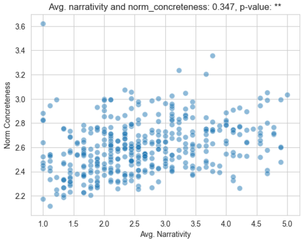
Imageability, **concreteness**, **visuality**: integral to narrative?
Narrativity scores of passages ($n = 401$) from [Piper et al., 2021]:



(a) Imageability



(b) Visuality



(c) Concreteness

** Stronger of correlation with "worldmaking" component (i.e., scoring of "This passage creates a world that I can see and feel")

INTRODUCTION: PROBLEMS

Gauging imageability in literature

Problem 1: Vague definition, conceptual overlaps, & intercorrelation

Problem 2: Limited coverage – MRC Imageability ratings < 5,000 lemmas

Problem 3: Going from words to sentences/passages. Accuracy of aggregated imageability is weak [Verma et al., 2023], probably due to sentence compositionality

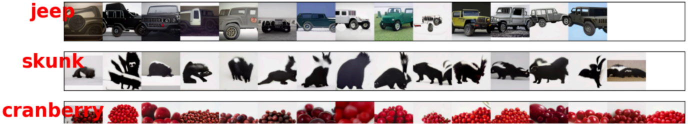
OUTLINE

1. Introduction
2. **Approaches**
3. Experiments
4. Conclusion

OTHER APPROACHES

Text-to-image generation: imageability of a (prompt) text to filter and direct models
- can *improve quality* of generations & can *reduce bias* from low-imageability concepts (like "Parisian" or "Bahamian") [Yang et al., 2020]

> Efforts to measure imageability taking **compositionality** into account [Wu and Smith, 2023; Hessel et al., 2018; Verma et al., 2023];
i.a., consistency of generated images as a proxy for sentence-imageability [Wu and Smith, 2023]
– uneven performance for fiction



OUR SETUP

An initial computational exploration aimed at:

(1) testing the relationship between **imageability** and related constructs such as **concreteness** and **visuality**
(2) evaluating the potential of multimodal embedding-based metrics to model imageability beyond static, word-level ratings

3 dictionaries (Imageability, Visuality, Concreteness)

CLIP model [Radford et al., 2021]; multimodal vision language model trained on image-text pairs. Distributional properties of embeddings – their shapes: **norm** (value-sizes across dimensions) & **entropy** (evenness)

*we also use variance and sparsity in the paper but let's keep it brief

HYPOTHESES

H1: Imageable and concrete words have multimodal embeddings with more localized activation – a few dominant dimensions carry most of the weight. ($\downarrow entropy$)
Example: `dog` tends to yield visually consistent, clustered representations (in model space, paired images typically depict dogs), whereas `beautiful` shows more evenly distributed activation, reflecting its vagueness.

H2: Alternatively, imageable words may activate more dimensions overall — reflecting richer, more distributed sensory associations. ($\uparrow norm, \uparrow entropy$)
Example: `blanket` evokes touch, warmth, shape, and color, while `justice` is more abstract and spikes in fewer, narrowly focused embedding dimensions.

OUR EXPERIMENT SETUP

Word-level analysis: relationship between human imageability scores and multimodal embedding shapes for dictionary entries of the MRC imageability dictionary

Sentence-level analysis: We compare dictionary scores with multimodal embeddings shapes in literary texts

Literary case study: We examine the discriminatory power of these embedding-based metrics where imageability is expected to differ: **imagist poems** versus **love poems**

OUTLINE

EXPERIMENT I: CORRELATIONS W. HUMAN-CODING

Visuality shows weaker correlations with imageability and concreteness: the dictionary-based concept of imageability ($n = 4,800$ words) is more closely related to concreteness

Entropy and norm exhibit medium correlations with human-coded scores ($\rho > 0.5$) – especially **imageability** & **concreteness** * When using embeddings from BERT, we do not see correlations

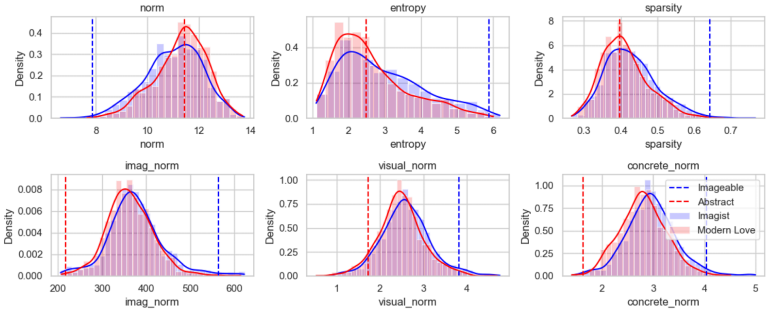
EXPERIMENT II: FROM WORDS TO LITERARY SENTENCES

	data	n	imag	visual	concrete
Norm	*The Old Man and the Sea* (1952)	1,928	-0.61	-0.63	-0.62
	Chicago (1880-2000)	9,000	-0.31	-0.37	-0.36
Entropy	*The Old Man and the Sea*	-	0.60	0.63	0.62
	Chicago	-	0.31	0.37	0.37

EXPERIMENT III: IMAGISM

Imagism : emphasis on clear, visual language and rejection of abstraction [Pound, 1913]

- Highly imageable: `"Homespun, dyed butternuts dark gold color."`
- Non-imageable: `"Of insidious intent"`



OUTLINE

CONCLUSIONS

Dictionaries (visual, imageability, concreteness) diverge
… we could use some work on the differences between imageability and concreteness, especially as it relates to narrativity, immersion / transportation

Embedding-based metrics show promise for sentence-or paragraph imageability assessment, considering proposed sensitivity to context and compositionally

CONCLUSIONS

- Highly concrete and imageable words exhibit a more evenly distributed, multimodal representation – with more dimensions contributing weakly (↑ entropy) and less overall activation (↓ norm). Multiple weak signals rather than a few strong ones
- Abstract words show sharper, more localized activation patterns (↓ entropy, ↑ norm)

Still an open question whether embeddings might actually outperform lexicon-based methods in capturing sentence-level human imageability judgments!

THANK YOU

```python
if any(questions):
    try:
        answer()
    except:
        print(":)")
else:
    thankyou()
```

pascale.feldkamp@cas.au.dk
chc.au.dk

SLIDES:
https://pfeldkamp.github.io/pascalefeldkamp/pdfs/CCLS25imageability.pdf

QR 2 SLIDES

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 33A(4):497–505. Place: United Kingdom Publisher: Taylor & Francis.

Hessel, J., Mimno, D., and Lee, L. (2018). Quantifying the Visual Concreteness of Words and Topics in Multimodal Datasets. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2194–2205, New Orleans, Louisiana. Association for Computational Linguistics.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., and Carney, J. (2020). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52(3):1271–1291.

Paivio, A., Yuille, J. C., and Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 76(1):1–25. Place: US Publisher: American Psychological Association.

Piper, A., Bagga, S., Monteiro, L., Yang, A., Labrosse, M., and Liu, Y. L. (2021). Detecting narrativity across long time scales. In *Proceedings of CHR 2021: Computational Humanities Research Conference*, pages 319–333. CEUR.

Pound, E. (1913). A few don’ts by an Imagiste. *Poetry*, 1(6):200–206.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR. ISSN: 2640-3498.

Verma, G., Rossi, R., Tensmeyer, C., Gu, J., and Nenkova, A. (2023). Learning the Visualness of Text Using Large Vision-Language Models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2394–2408, Singapore. Association for Computational Linguistics.

Wu, S. and Smith, D. (2023). Composition and Deformance: Measuring Imageability with a Text-to-Image Model. In Akoury, N., Clark, E., Iyyer, M., Chaturvedi, S., Brahman, F., and Chandu, K., editors, *Proceedings of the 5th Workshop on Narrative Understanding*, pages 106–117, Toronto, Canada. Association for Computational Linguistics.

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., and Russakovsky, O. (2020). Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558. arXiv:1912.07726 [cs].