Efficient Compression in Locomotion Verbs Across Languages

Thomas A. Langlois[1,2], Roger Levy[1], Nidhi Seethapathi[1], Noga Zaslavsky[2]

[1]Massachusetts Institute of Technology (MIT) Brain and Cognitive Sciences Department (BCS) [2]New York University (NYU) Department of Psychology

Background

- Languages shaped by a drive for efficient communication[1]
  - Evidence from color word adjectives[1], object nouns[2]
- New perceptual domain: time-varying visual stimuli
  - Builds on evidence from static visual stimuli[1,2]
- Action words (verbs) not studied in terms of efficiency
  - Prior work on locomotion verbs, but not in terms of IB[3]

Question: Do locomotion verbs show pressure for efficiency?

Stimuli & naming data

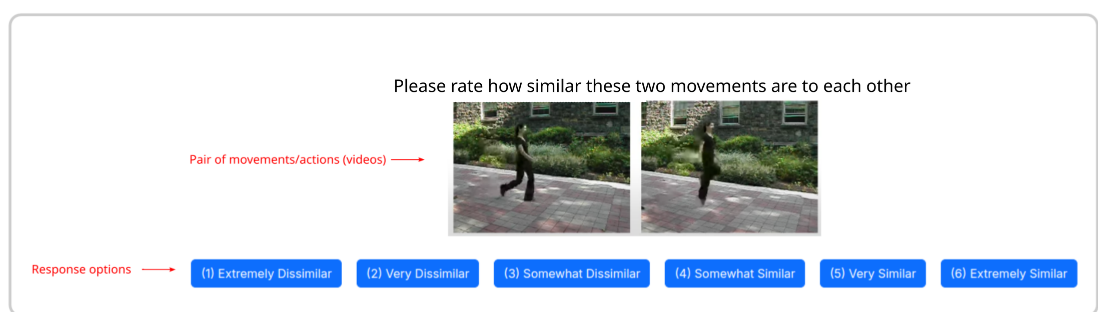
- Motion stimuli evaluated using existing locomotion dataset[3]

- Verb meanings evaluated using existing verb annotations[3]
- English, Spanish, Dutch, Japanese verbs (Malt et al. 2008)

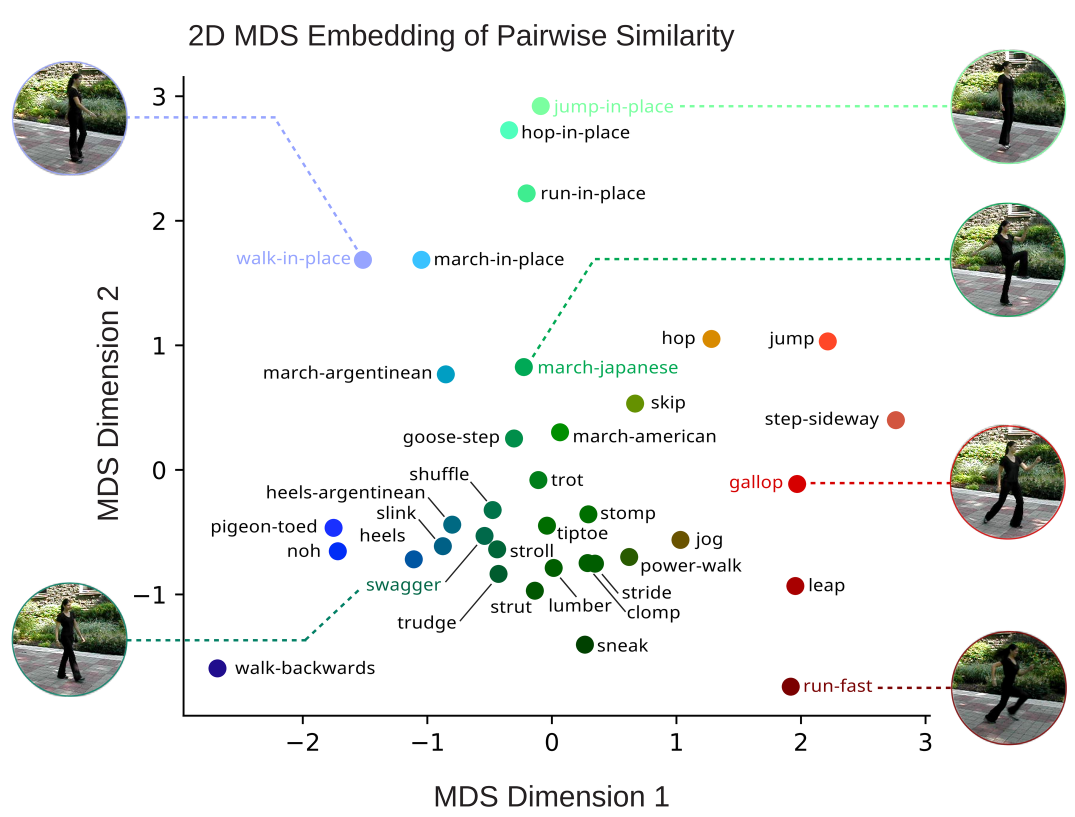
Similarity data

- Estimating a psychological locomotion similarity space

Similarity Task

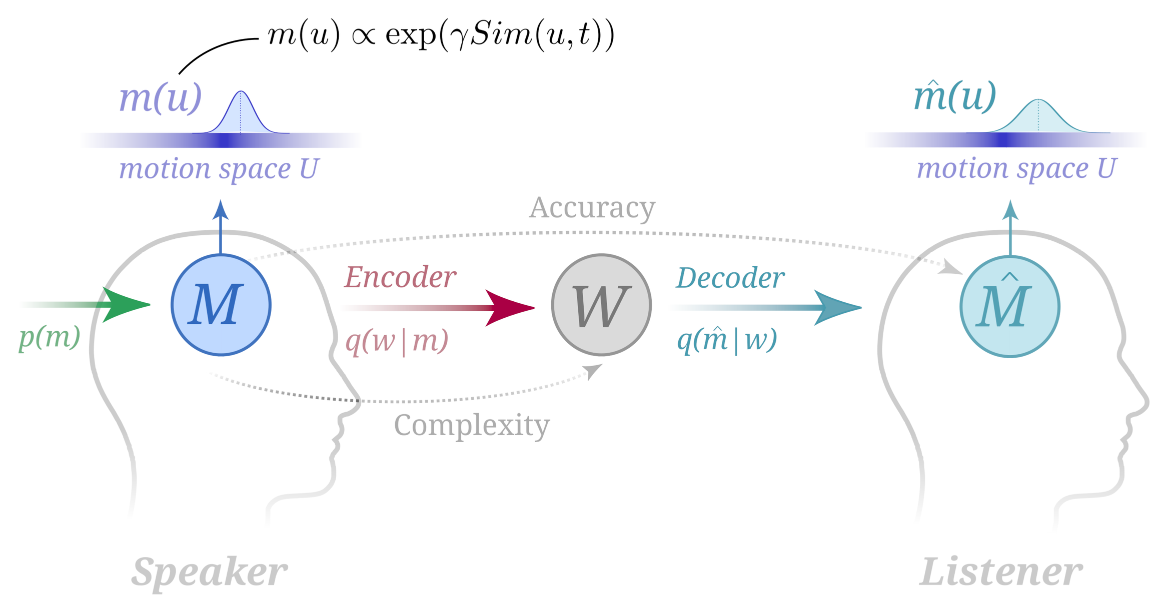


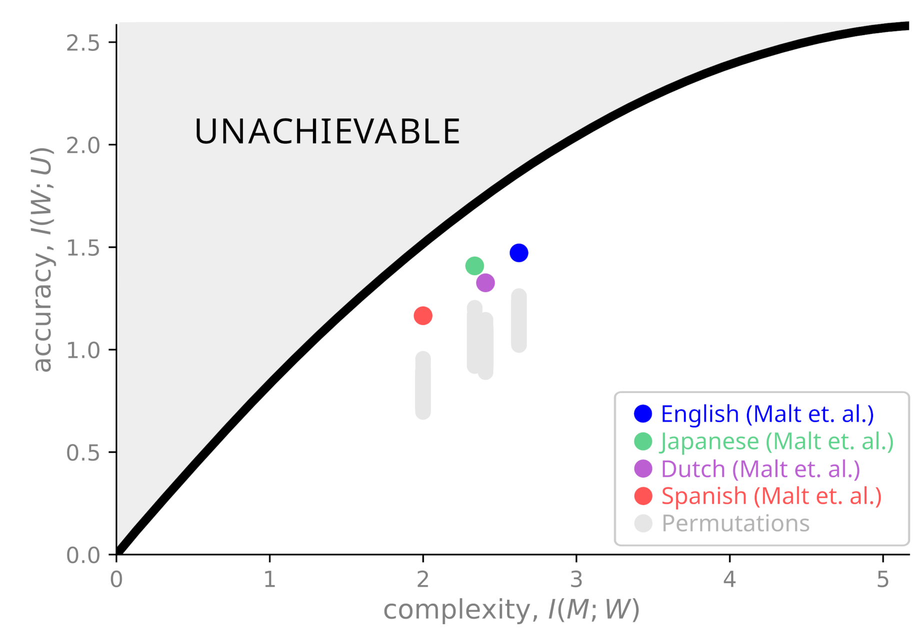
Locomotion Space U (MDS)



Communication model & Efficiency analysis

- Verb meanings across multiple languages show a pressure for efficiency





Theoretical Model (IB)

Speaker conveys a mental locomotion representation *m(u)* via a naming policy *q(w|m)*

Listener infers *$\hat{m}$(u)* via decoder *q($\hat{m}$|w)* using Bayes' rule

Theoretical Encoder

Complexity
$$I_q(M;W) = \sum_m \sum_w q(w|m)p(m)\log\left[\frac{q(w|m)}{q(w)}\right]$$

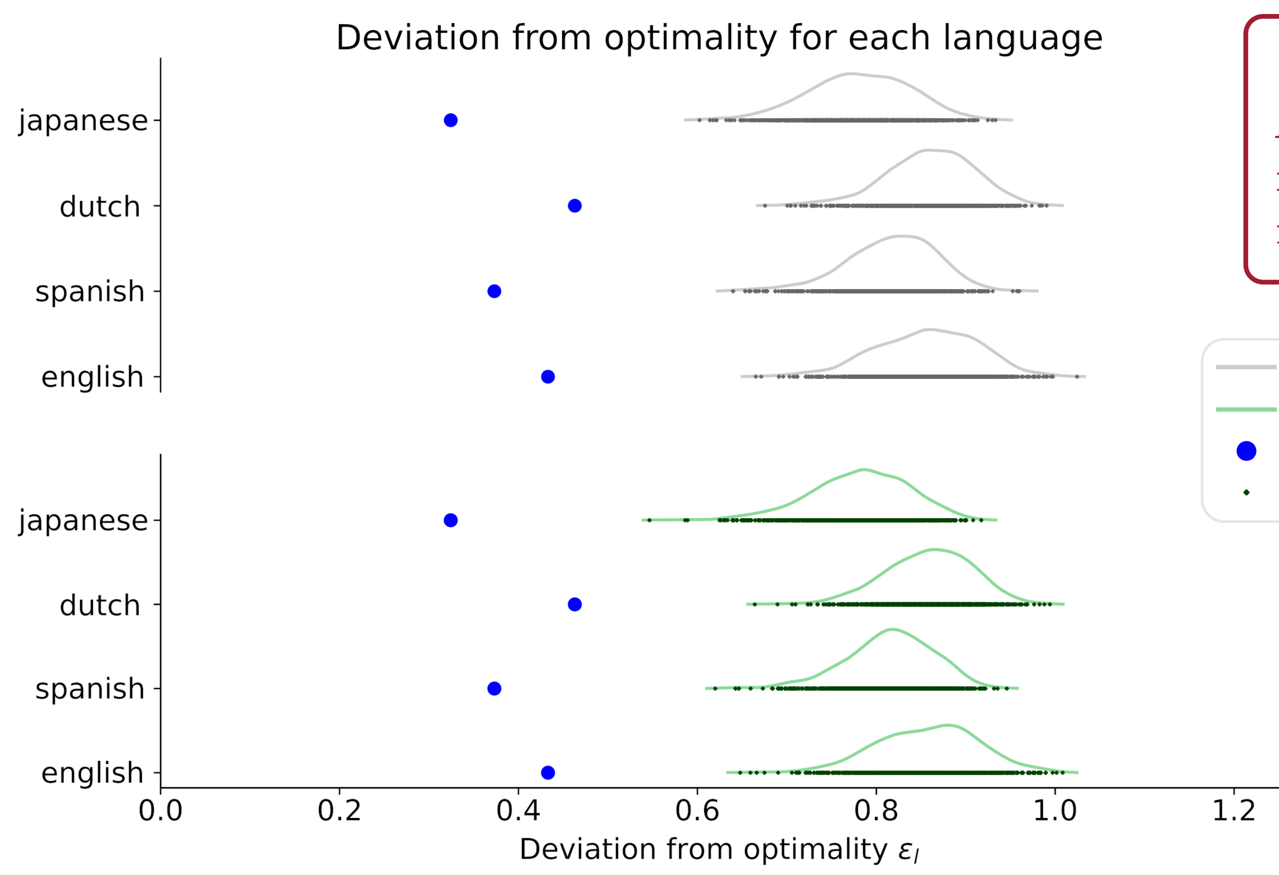
$$\mathcal{F}_\beta[q(w|m)] = I_q(M;W) - \beta I_q(W;U)$$

Tradeoff parameter

Accuracy (inversely ~ to distortion)
$$\mathbb{E}_q\left[D[M||\hat{M}]\right] = \mathbb{E}_q\left[\sum_u m(u)\log\frac{m(u)}{\hat{m}(u)}\right]$$

Quantitative analyses

- Permutations of naming data or similarity yield greater deviations from optimality



Quantitative Metric

Deviation from Optimality: measures deviation in IB Objective value between value obtained with real language *p(w|m)* vs. optimal encoder *q(w|m)*

Attested System (human data) Optimal IB system

$$\epsilon_l = \frac{1}{\beta_l}\left(\mathcal{F}_{\beta_l}[p_l(w|m)] - \mathcal{F}^*_{\beta_l}\right)$$

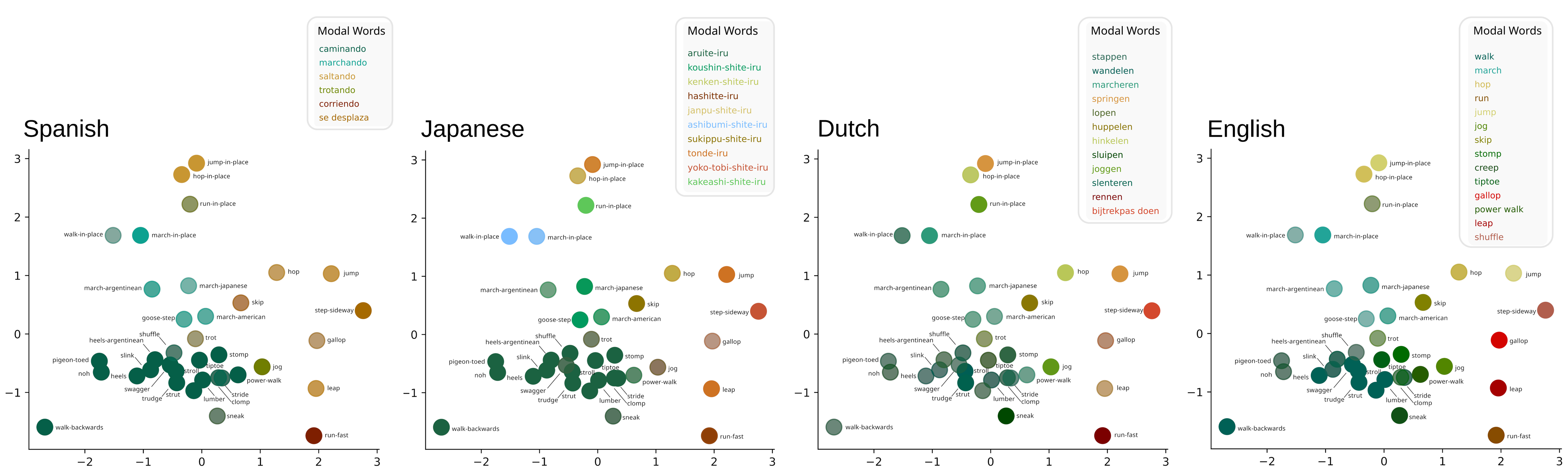
Deviation from optimality

- Naming data permutation analysis indicates that **optimal IB encoders align with attested systems**
- Similarity permutation analysis shows that meaning space reflects **important semantic structure**
- Caveat: gaps remain between optimal IB systems and human languages

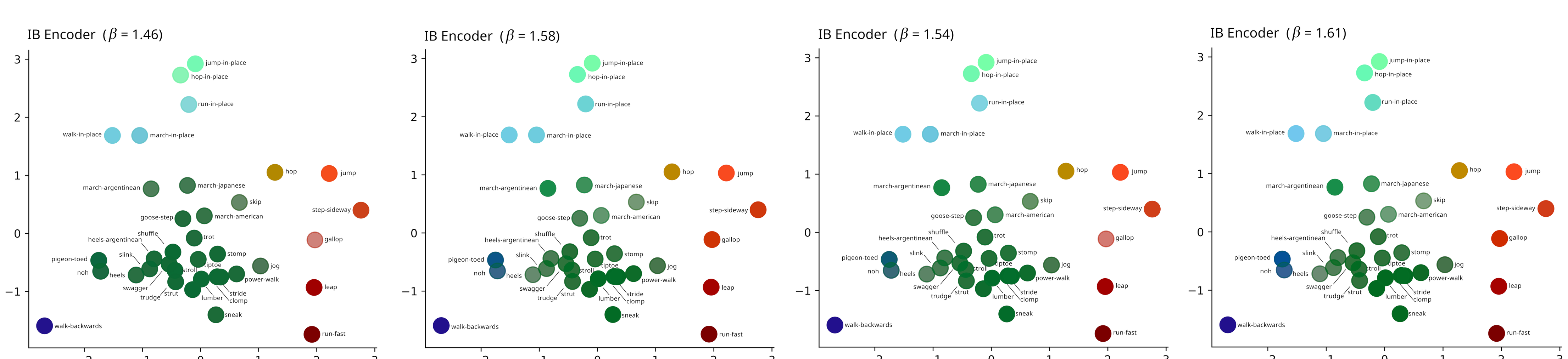
Attested & theoretical systems

- Attested systems *p(w|m)* align with optimal IB encoders *q(w|m)*

Malt et al., (2008)



IB optimal systems



Conclusions

- Languages efficiently compress verb meanings
- Psychological similarity reflects semantic structure
- Optimal IB encoders align with attested systems

- Need to expand stimulus coverage in *U* and languages *l*
  - More ecologically valid and dense motion space
  - More data from a larger variety of languages
- Need to identify best meaning space and *p(m)*
  - Model has uniform *p(m)* but real prior is likely not
  - Meaning space captures structure but gaps remain

References

[1]Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31), 7937–7942

[2]Zaslavsky, N., Regier, T., Tishby, N., & Kemp, C. (2019). Semantic categories of artifacts and animals reflect efficient coding. arXiv preprint arXiv:1905.04562.

[3]Malt, B. C., Gennari, S., Imai, M., Ameel, E., Tsuda, N., & Majid, A. (2008). Talking about walking: Biomechanics and the language of locomotion. Psychological science, 19(3), 232–240.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, 4(1), 109–128.