Challenges and insights from cross-linguistic word-meaning associations: A roadmap for the study of loose colexification

Thomas Brochhagen

Cross-linguistic data formats workshop

MPI-EVA, 2023.12.14

upf. Universitat Pompeu Fabra Barcelona

Departament | Facultat de Traducció i Ciències del Llenguatge

1

The role of (psycho-)metrics in explaining cross-linguistic semantic organization

1 The role of (psycho-)metrics in explaining cross-linguistic semantic organization

2 Challenges raised by the use of (psycho-)metrics in cross-linguistic research

1 The role of (psycho-)metrics in explaining cross-linguistic semantic organization

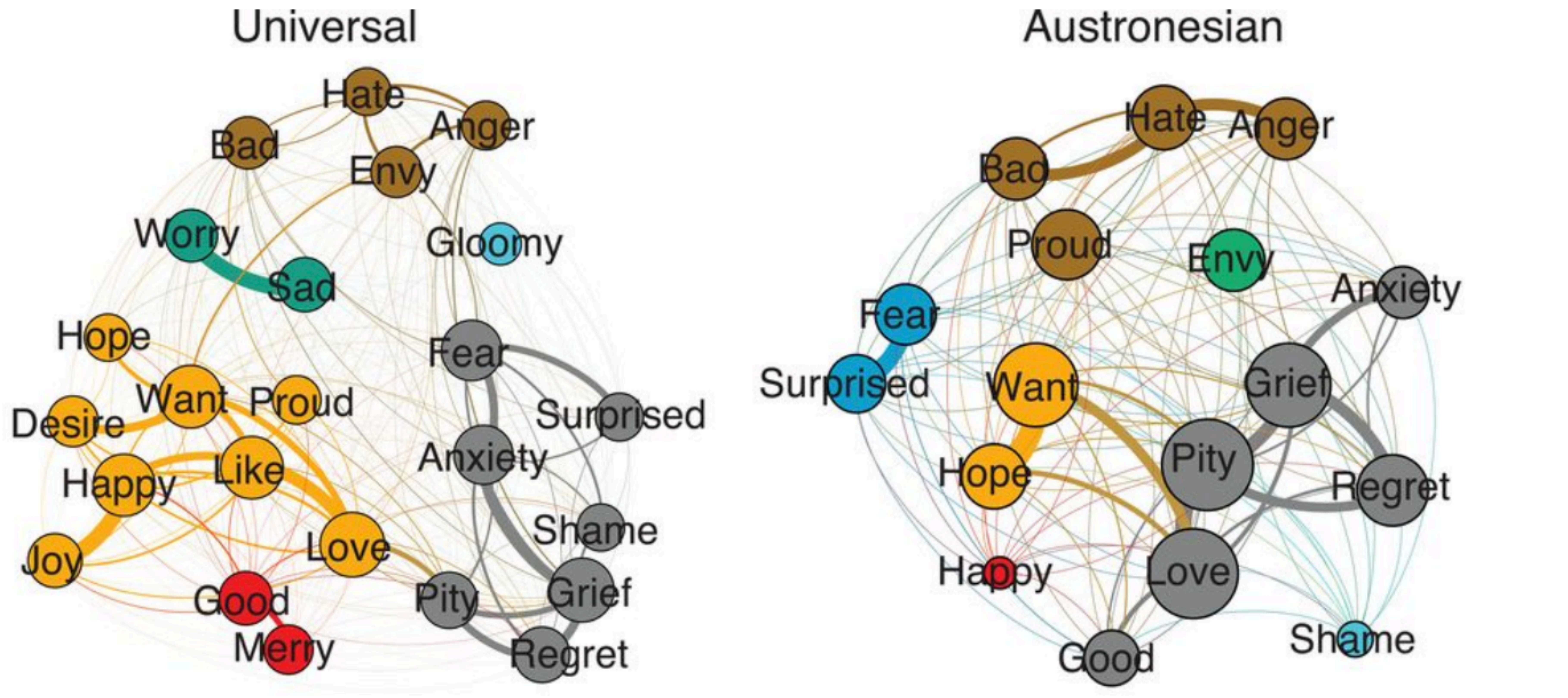
2 Challenges raised by the use of (psycho-)metrics in cross-linguistic research

3 Scalable alternatives: A roadmap for the study of loose colexification

The role of (psycho-)metrics in explaining cross-linguistic semantic organization

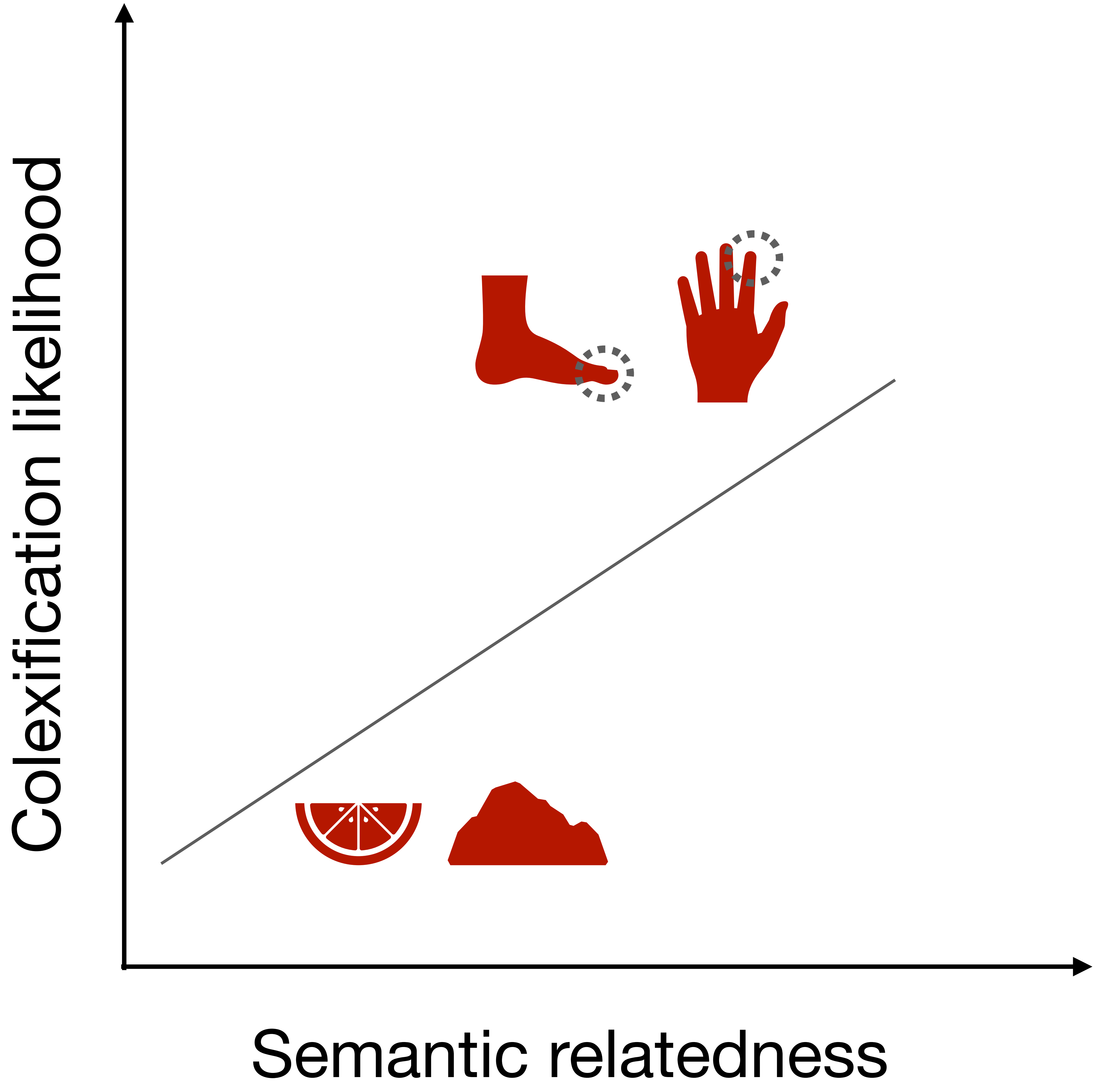
Rzymski, Christoph and Tresoldi, Tiago et al. 2019. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies.

Leverage distribution of word-meaning associations to infer regularities directly



Jackson, J., J. Watts, T. Henry, J.-M. List, P. Mucha, R. Forkel, S. Greenhill, R. Gray, and K. Lindquist (2019): Emotion semantics show both cultural variation and universal structure. Science 366.6472. 1517-1522

Study word-meaning associations through the relationship meanings stand in



Xu, Y., Duong, K., Malt, B. C., Jiang, S., & Srinivasan, M. (2020). Conceptual relations predict colexification across languages. Cognition, 201, 104280.

Brochhagen, T., & Boleda, G. (2022). When do languages use the same word for different meanings? The Goldilocks principle in colexification. Cognition, 226, 105179

Brochhagen, T., Boleda, G., Gualdoni, E., & Xu, Y. (2023). From language development to language evolution: A unified view of human lexical creativity. Science, 381(6656)

Three challenges for the study word-meaning associations through the relationship meanings stand in

First challenge: ad-hoc English-based enrichments

Resources like CLICS* do not ship with (psycho)metrics about forms, meanings, or their relationship

We have to rely on enrichments that range from the relatively straightforward (vision, associativity, affectiveness) to the more ad-hoc (WordNet)

These tend to be English-based

*Tjuka, Annika, Robert Forkel, and Johann-Mattis List. 2022. Linking Norms, Ratings, and Relations of Words and Concepts Across Multiple Language Varieties. Behavior Research Methods 54. 864–884

Second challenge: decontextualized (psycho-)metrics

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. Cognition, 122(3), 280–291

Brochhagen, T. (2020). Signalling under Uncertainty: Interpretative Alignment without a Common Prior. The British Journal for the Philosophy of Science, 71(2), 471–496

Brochhagen, T. (2021). Brief at the Risk of Being Misunderstood: Consolidating Population- and Individual-Level Tendencies. Computational Brain & Behavior, 4(3), 305–317

…

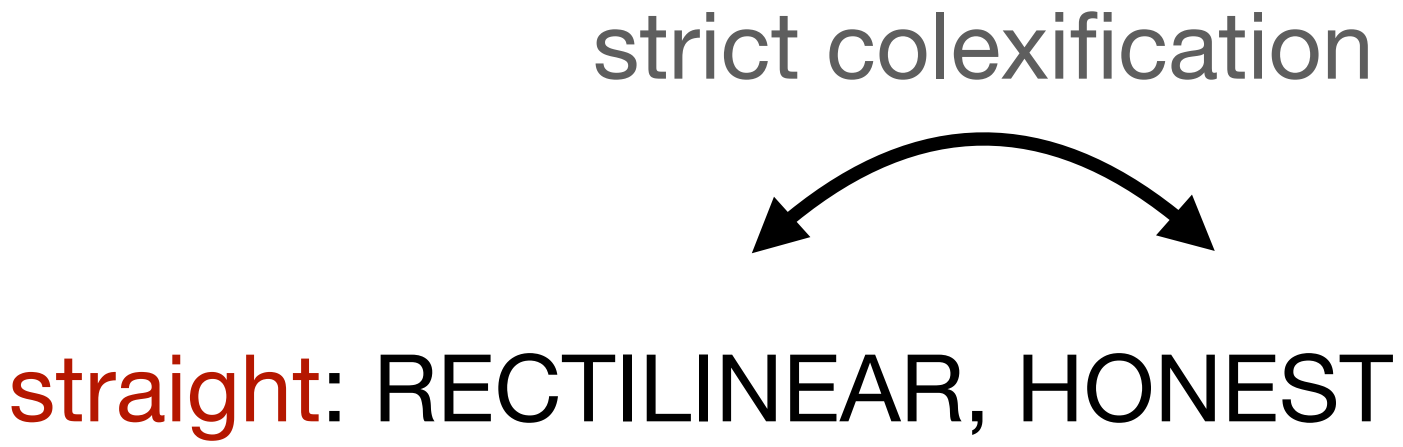
Third challenge: phylogenetic and geographic bias

Guzmán Naranjo, M., & Becker, L. (2021). Statistical bias control in typology. Linguistic Typology, 26(3), 605–670

Hartmann F. & Jäger G. (under review). Gaussian process models for geographic controls in phylogenetic trees. Open Research Europe

Scalable alternatives: A roadmap for the study of loose colexification

Scalable alternatives: A roadmap for the study of loose colexification

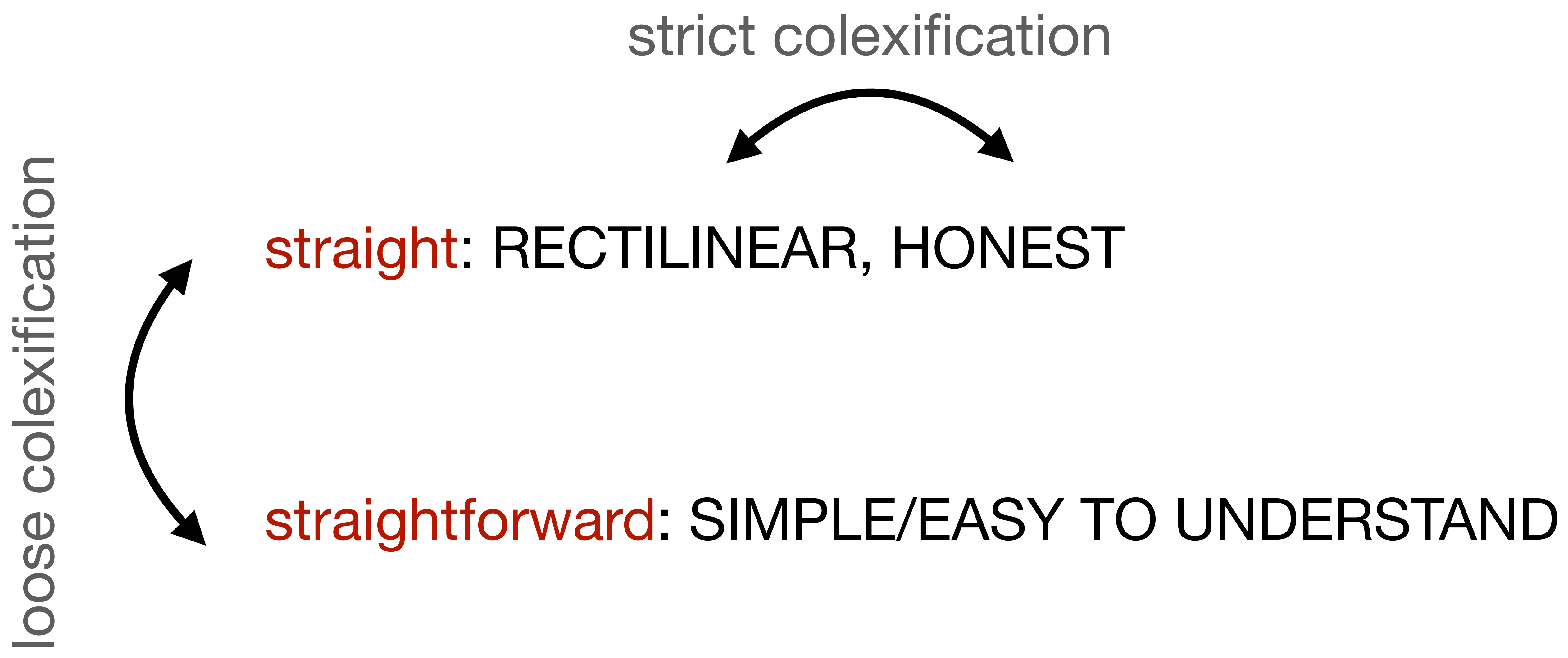


strict colexification

straight: RECTILINEAR, HONEST

List, J.-M. (2023): Inference of partial colexifications from multilingual wordlists. *Frontiers in Psychology* 14.1156540. 1-10.

Scalable alternatives: A roadmap for the study of loose colexification



strict colexification

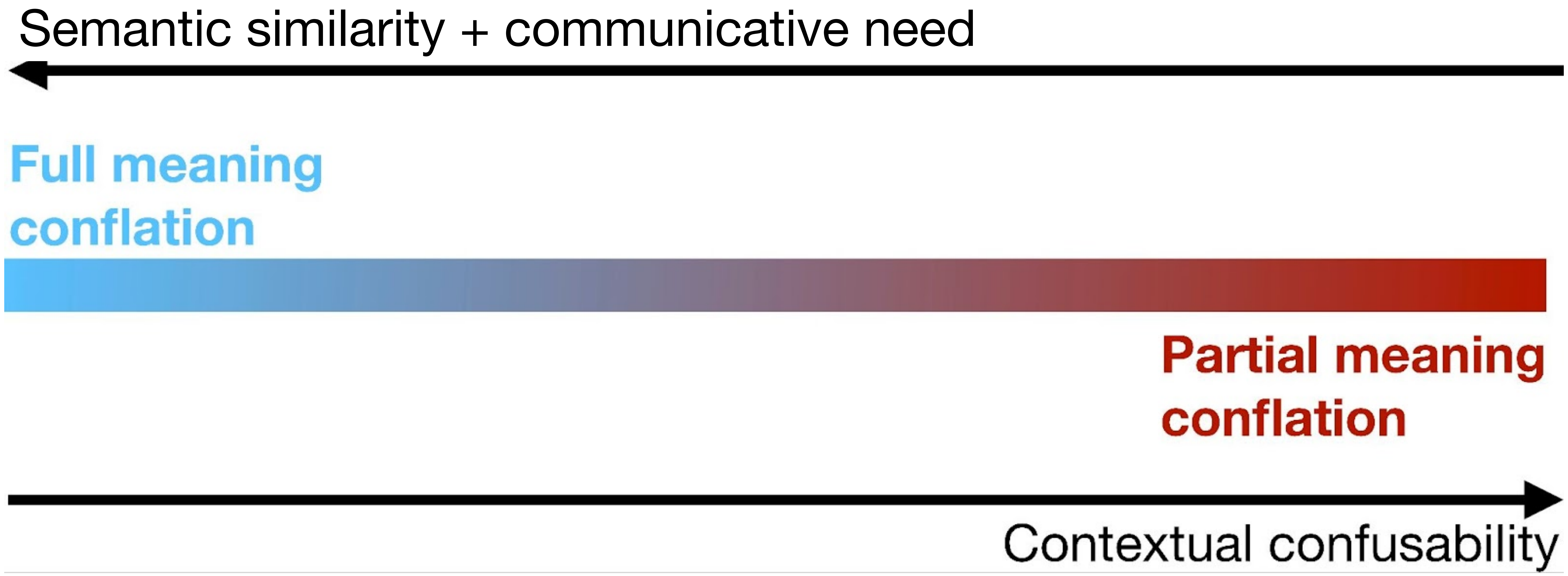
straight: RECTILINEAR, HONEST

straightforward: SIMPLE/EASY TO UNDERSTAND

loose colexification

List, J.-M. (2023): Inference of partial colexifications from multilingual wordlists. *Frontiers in Psychology* 14.1156540. 1-10.

Scalable alternatives: A roadmap for the study of loose colexification


Semantic similarity + communicative need
Full meaning conflation
Partial meaning conflation
Contextual confusability

Brochhagen, T., & Boleda, G. (2022). When do languages use the same word for different meanings? The Goldilocks principle in colexification. Cognition, 226, 105179

Norcliffe, E. & Majid, A. (2023). Partial and full colexification in the perception domain. Proceedings of ICLC16.

Scalable alternatives: A roadmap for the study of loose colexification

Abandon psychometrics, at least for now

Scalable alternatives: A roadmap for the study of loose colexification

Abandon psychometrics, at least for now

Unsupervised graded measure of loose colexification

Scalable alternatives: A roadmap for the study of loose colexification

Abandon psychometrics, at least for now

Unsupervised graded measure of loose colexification

Scalable measures of relationship between meanings

Byte-Pair Encoding based notion of loose colexification

```
aaabdaaabac
```

Byte-Pair Encoding based notion of loose colexification

```
aaabdaaabac

ZabdZabac
```

Byte-Pair Encoding based notion of loose colexification

```
aaabdaaabac

ZabdZabac

ZYdZYac
```

Byte-Pair Encoding based notion of loose colexification

```
aaabdaaabac

ZabdZabac

ZYdZYac

XdXac
```

Byte-Pair Encoding based notion of loose colexification

Find all mergers of forms of language l

Byte-Pair Encoding based notion of loose colexification

Find all mergers of forms of language l

Get frequency of mergers

Byte-Pair Encoding based notion of loose colexification

Find all mergers of forms of language l

Get frequency of mergers

For each pair of forms, i and j, find least frequent merger common to both

freq$(m_{i,j}^{l})$

Byte-Pair Encoding based notion of loose colexification

Find all mergers of forms of language l

Get frequency of mergers

For each pair of forms, i and j, find least frequent merger common to both

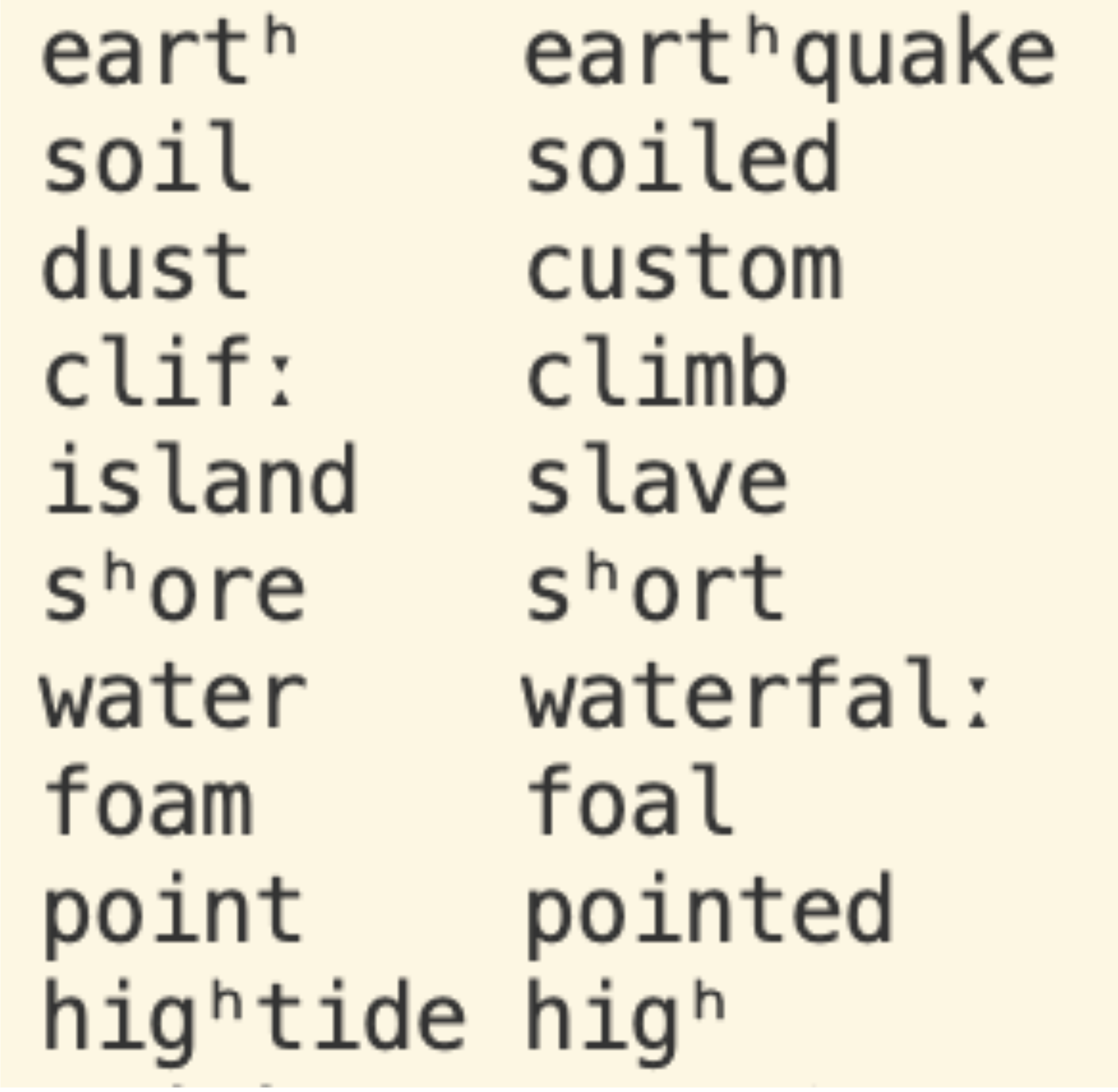
$$\text{freq}(m^l_{i,j})$$

Certainty* of loosely colexifying is given by

$$\frac{1}{\text{freq}(m^l_{i,j})} \quad \text{if} \quad i \neq j$$

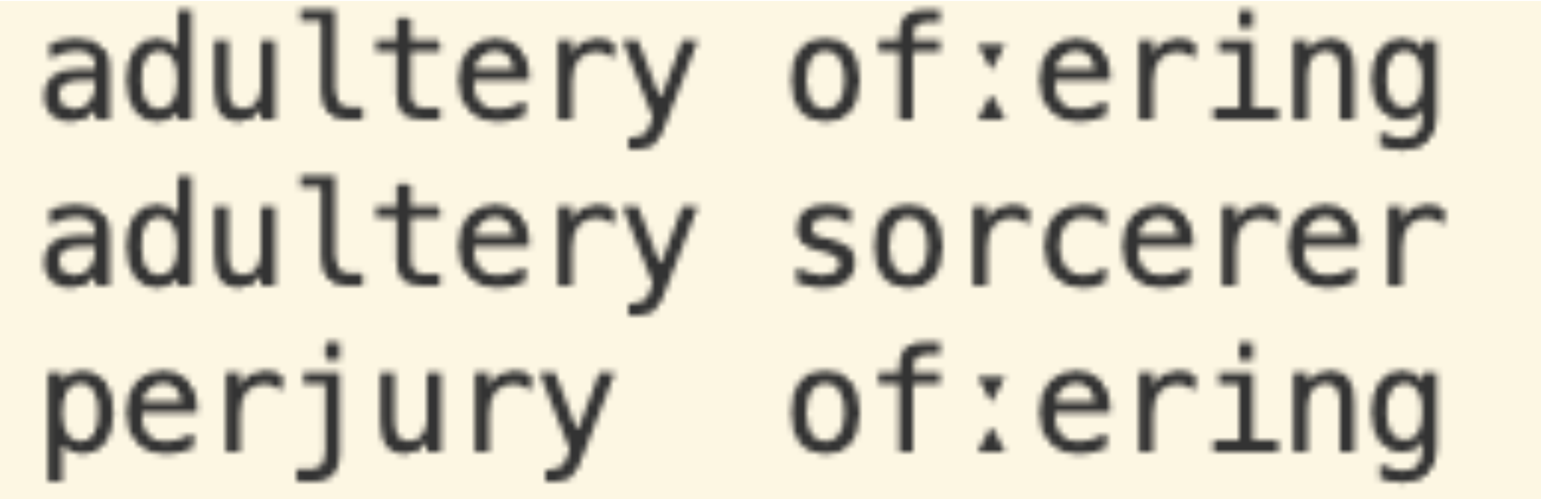
* may also be considered "degree of partial colexification" but less conceptually clear what this means

Non-identical top
(Most overlap without identity)



eartʰ	eartʰquake
soil	soiled
dust	custom
clifː	climb
island	slave
sʰore	sʰort
water	waterfalː
foam	foal
point	pointed
higʰtide	higʰ

Non-zero bottom
(Minimal overlap)



adultery	ofːering
adultery	sorcerer
perjury	ofːering

Measures of relationship between meanings

Scalable

Large amounts of data available

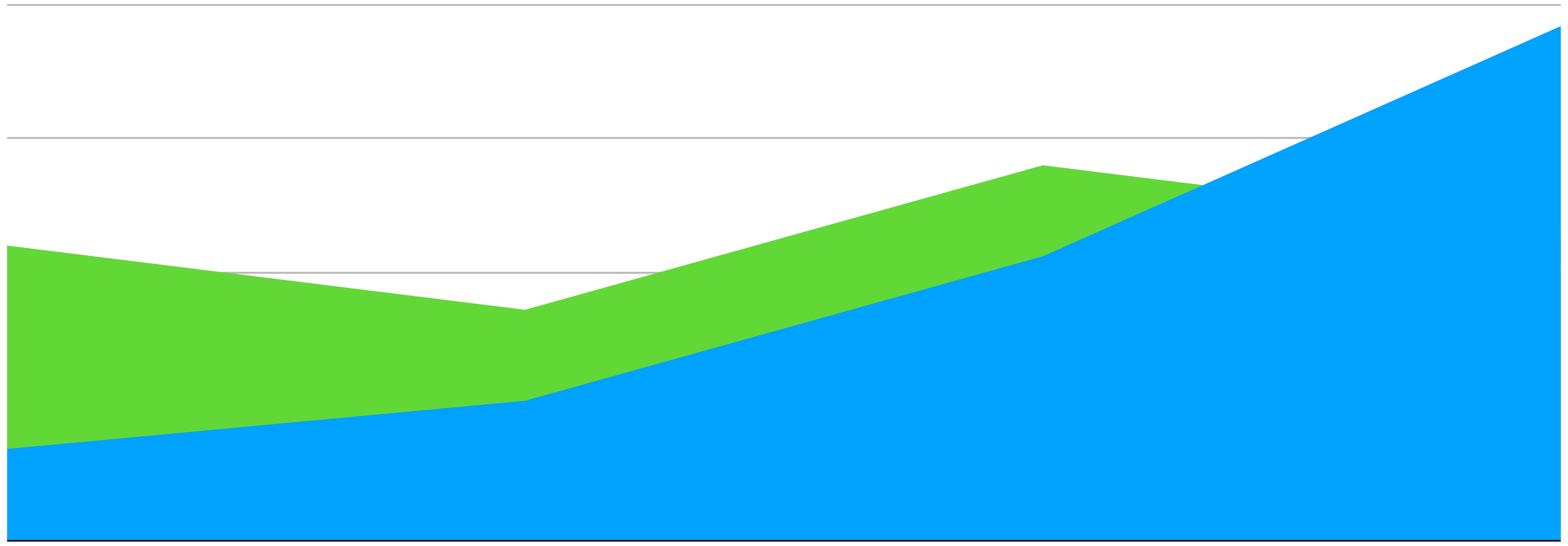
Multilingual

BigScience Workshop et al. (2023). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.

Seifart, F., Paschen, L., & Stave, M. (2022). Language Documentation Reference Corpus (DoReCo) 1.2.

Scalable measures of relationship between meanings: contextual confusability

Scalable measures of relationship between meanings: contextual confusability



“I stubbed my ___ when I walked into the room”

“I cut my ___ while cooking”

Scalable measures of relationship between meanings: contextual confusability

For each meaning, retrieve a large number of contexts in which it appears

“I stubbed my ___ when I walked into the room”

“I cut my ___ while cooking”

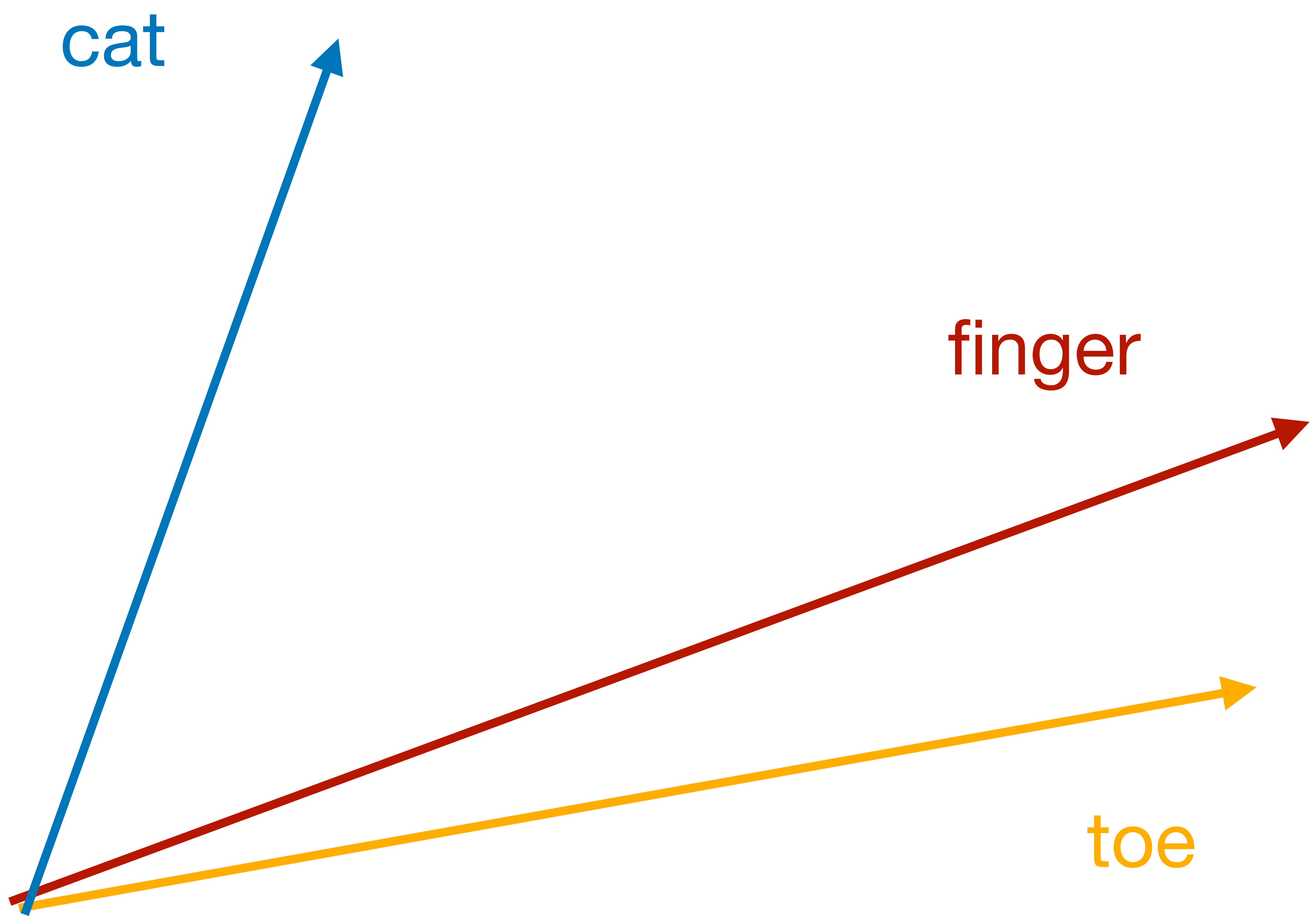
For each context, retrieve language model expectations over the vocabulary

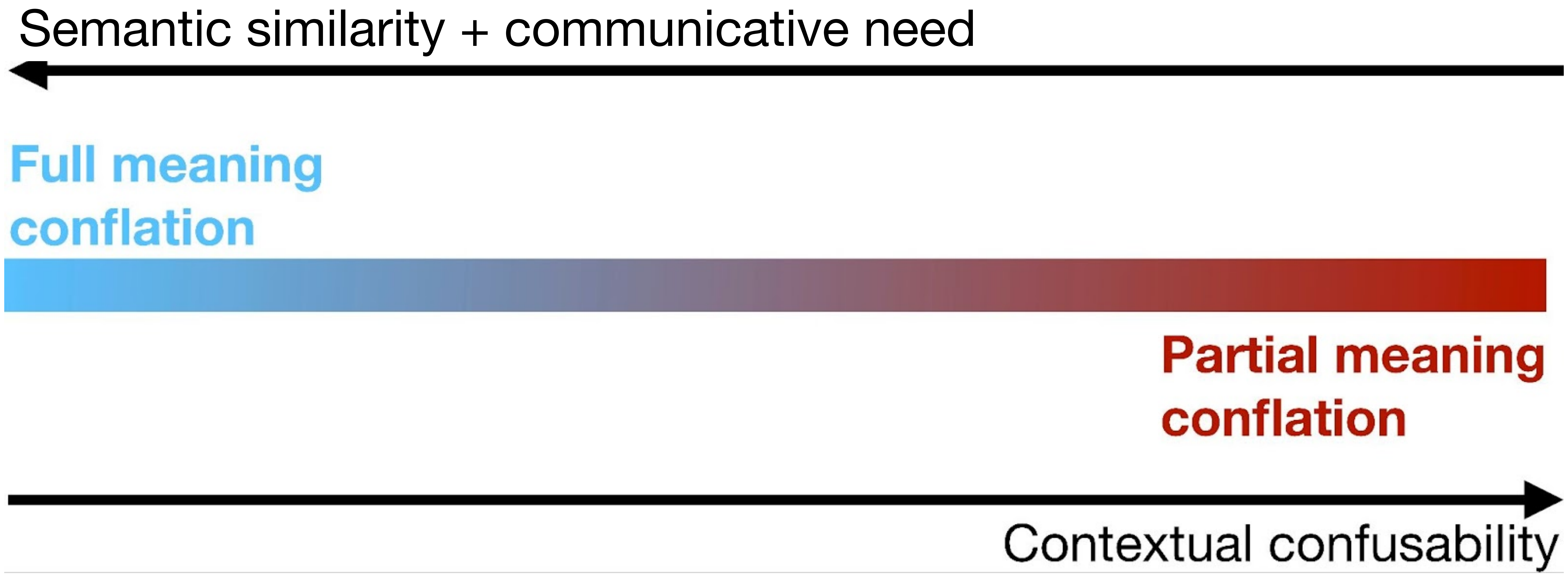
The confusability of two meanings is given by the (inverse of the average of the) sum of the Kullback-Leibler divergence of the expectations across contexts they appear in

➡ Contextual confusability as average Jeffrey’s divergence between language model expectations in context

Scalable measures of relationship between meanings: Semantic similarity

Semantic similarity as cosine similarity between word embeddings



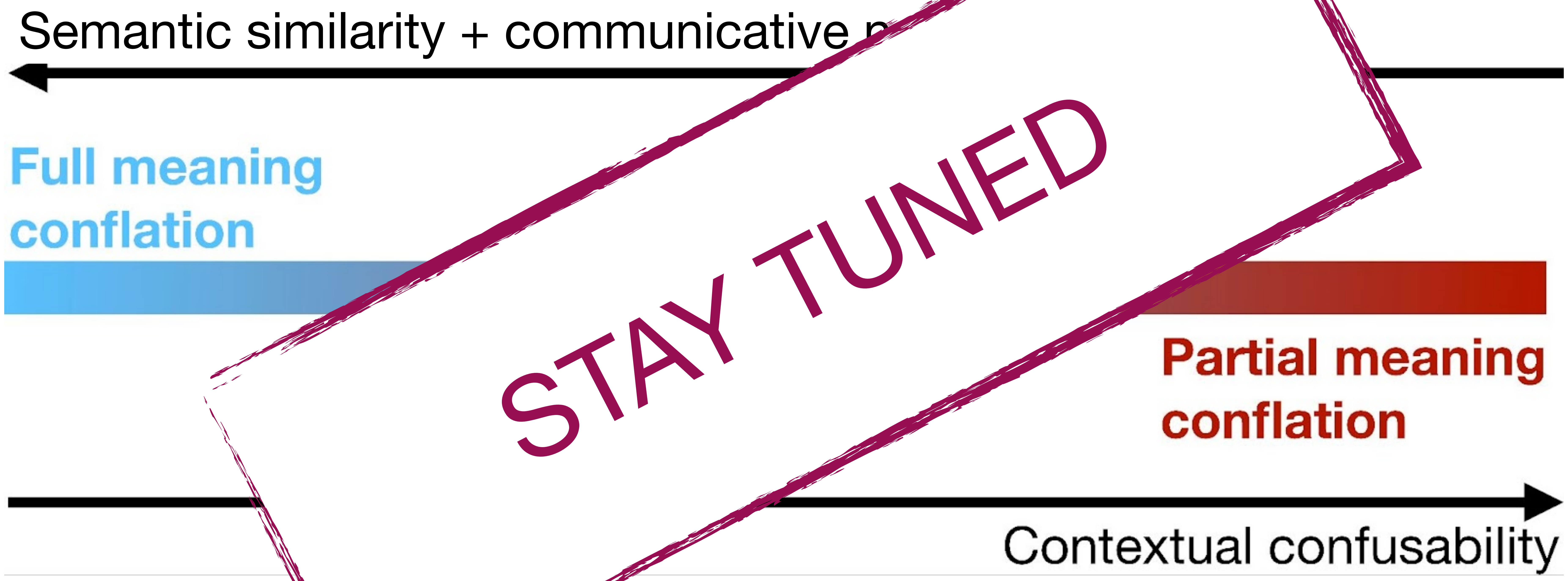


Semantic similarity + communicative need

Full meaning conflation

Partial meaning conflation

Contextual confusability



Semantic similarity + communicative n

Full meaning conflation

Partial meaning conflation

Contextual confusability

STAY TUNED

Coda

- Alternatives to identify and measure loose colexification
- Alternative operationalizations of contextual confusability
- Issues with assessing multilingual LMs quality
- Using GNNs to study partial vs. strict colexification
- Getting around HMC convergence issues when using Gaussian processes for phylo/geo bias

Challenges and insights from cross-linguistic word-meaning associations: A roadmap for the study of loose colexification

Thomas Brochhagen

Cross-linguistic data formats workshop

MPI-EVA, 2023.12.14

upf. Universitat Pompeu Fabra Barcelona

Departament | Facultat de Traducció i Ciències del Llenguatge