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

Thomas Brochhagen
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

Leverage distribution of word-meaning associations to infer regularities directly

Study word-meaning associations through the relationship meanings stand in


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

Second challenge: decontextualized (psycho-)metrics


...
Third challenge: phylogenetic and geographic bias


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

Scalable alternatives: A roadmap for the study of loose colexification

strict colexification

loose colexification

straight: RECTILINEAR, HONEST

straightforward: SIMPLE/EASY TO UNDERSTAND

Scalable alternatives: A roadmap for the study of loose colexification


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 $l$

Certainty* of loosely colexifying is given by

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

if $i \neq j$

* may also be considered "degree of partial colexification" but less conceptually clear what this means
Byte-Pair Encoding based notion of loose colexification

Find all mergers of forms of language \( l \)

Get frequency of mergers

Certainty of loosely colexifying is given by

\[
\text{freq}(m_{li}, j) = \frac{1}{\text{freq}(m_{li}, j) - 1}
\]

if \( i \neq j \)

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

\[
\text{freq}(m_{i,j}^l)
\]
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

\[
\text{freq}(m_{i,j})
\]

Certainty* of loosely colexifying is given by

\[
\frac{1}{\text{freq}(m_{i,j})} \quad \text{if} \quad i \neq j
\]

* may also be considered “degree of partial colexification” but less conceptually clear what this means
Non-identical top
(Most overlap without identity)

earth^h earth^hquake
soil soiled
dust custom
clif: climb
island slave
shore short
water waterfal:
foam foal
point pointed
high^tide high^h

Non-zero bottom
(Minimal overlap)

adultery of:ering
adultery sorcerer
perjury of:ering
Measures of relationship between meanings

Scalable

Large amounts of data available

Multilingual


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
Add again the figure about the relationship of forms and meanings according to ERC.
Add again the figure about the relationship of forms and meanings according to ERC.

Semantic similarity + communicative need

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