

Using word token embeddings

Katrin Erk

Visualization

Visualizing tokens in space

- Down-project a BERT space to 2 dimensions, then we can visualize many tokens of the same lemma
- The ContextAtlas does this with tokens from English Wikipedia
 - You can choose the lemma, and the layer to visualize
 - Example: “fire” <https://storage.googleapis.com/bert-wsd-vis/demo/index.html?#word=fire>
 - In the notebook: How to make your own visualizations, using your own corpus

Clustering

Usage types for diachronic lexical semantics

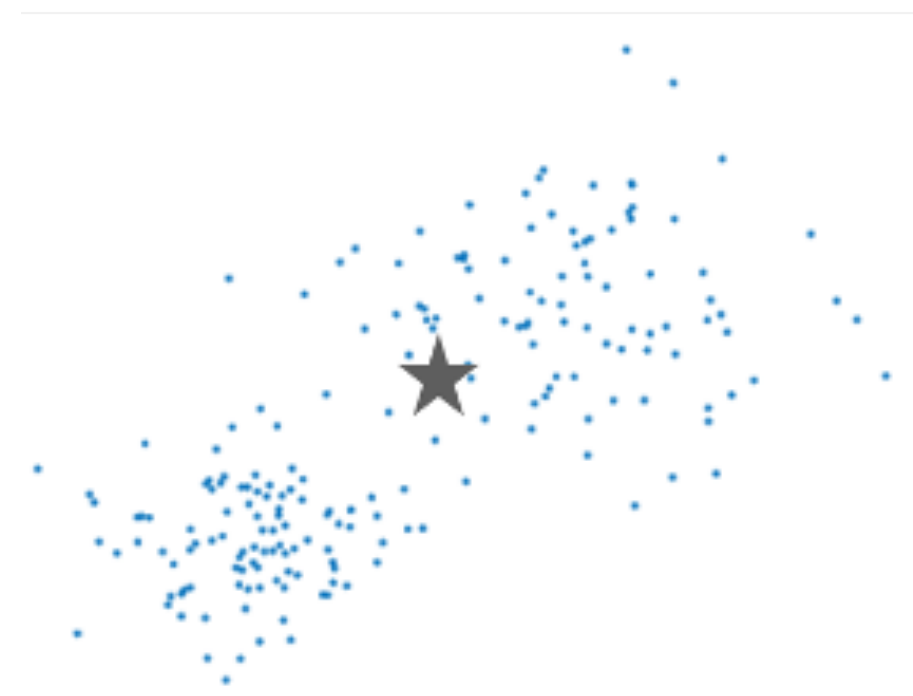
- Giulianelli, del Tredici, Fernandez: “Analysing lexical semantic change with contextualized word representations”
- Pre-trained contextualized language model, applied to text data from different time periods
- They cluster embeddings into usage types
Change in probability distribution over usage types as indicating lexical change
- Usage types analysis:
 - senses of polysemous, homonymous words are often separated
 - some clusters separate literal vs figurative uses
 - argument structure and syntax influence clustering
 - some errors: different senses lumped together, or a sense spread over multiple clusters

Clustering word tokens

Word token embeddings are vectors, like word type embeddings, so you can cluster them in the same way

Clustered tokens of a lemma turn out to be a useful intermediate level of representation in between keeping all tokens and collapsing to a single vector

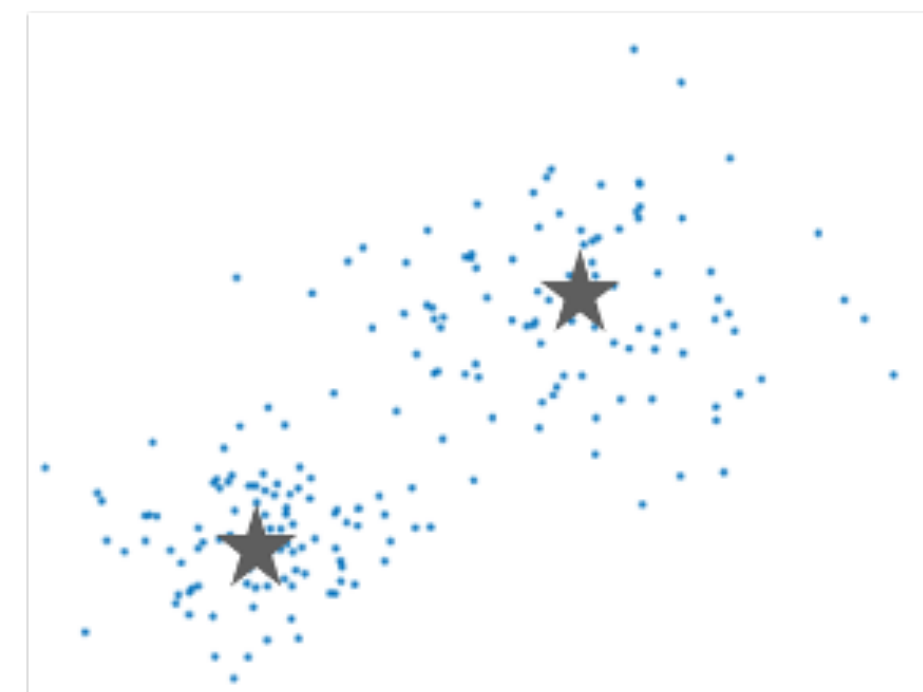
Single-Prototype



“bank”

average over all tokens

Multi-Prototype



“bank”

cluster tokens into multiple groups

Exemplar



“bank”

storing all the individual tokens

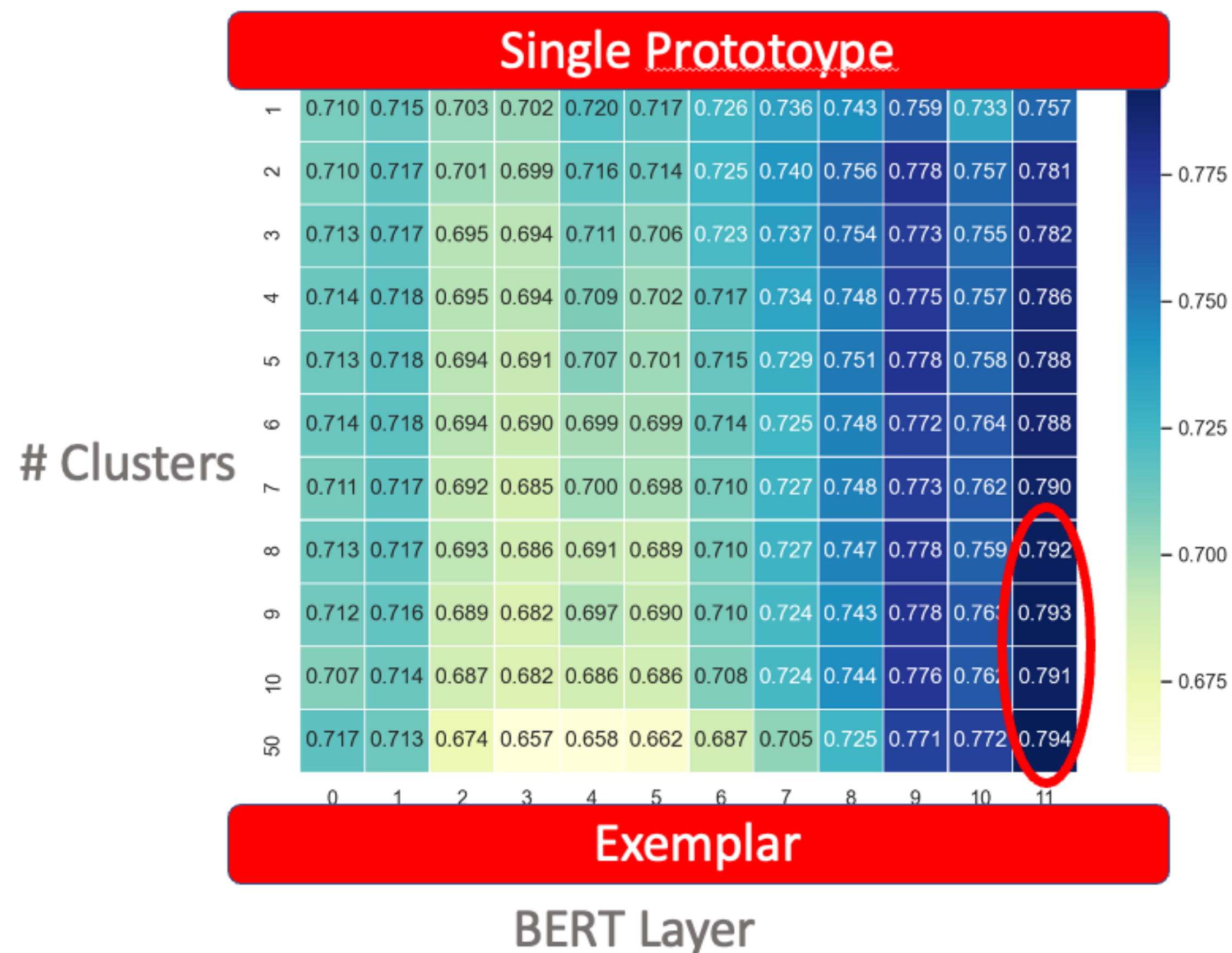
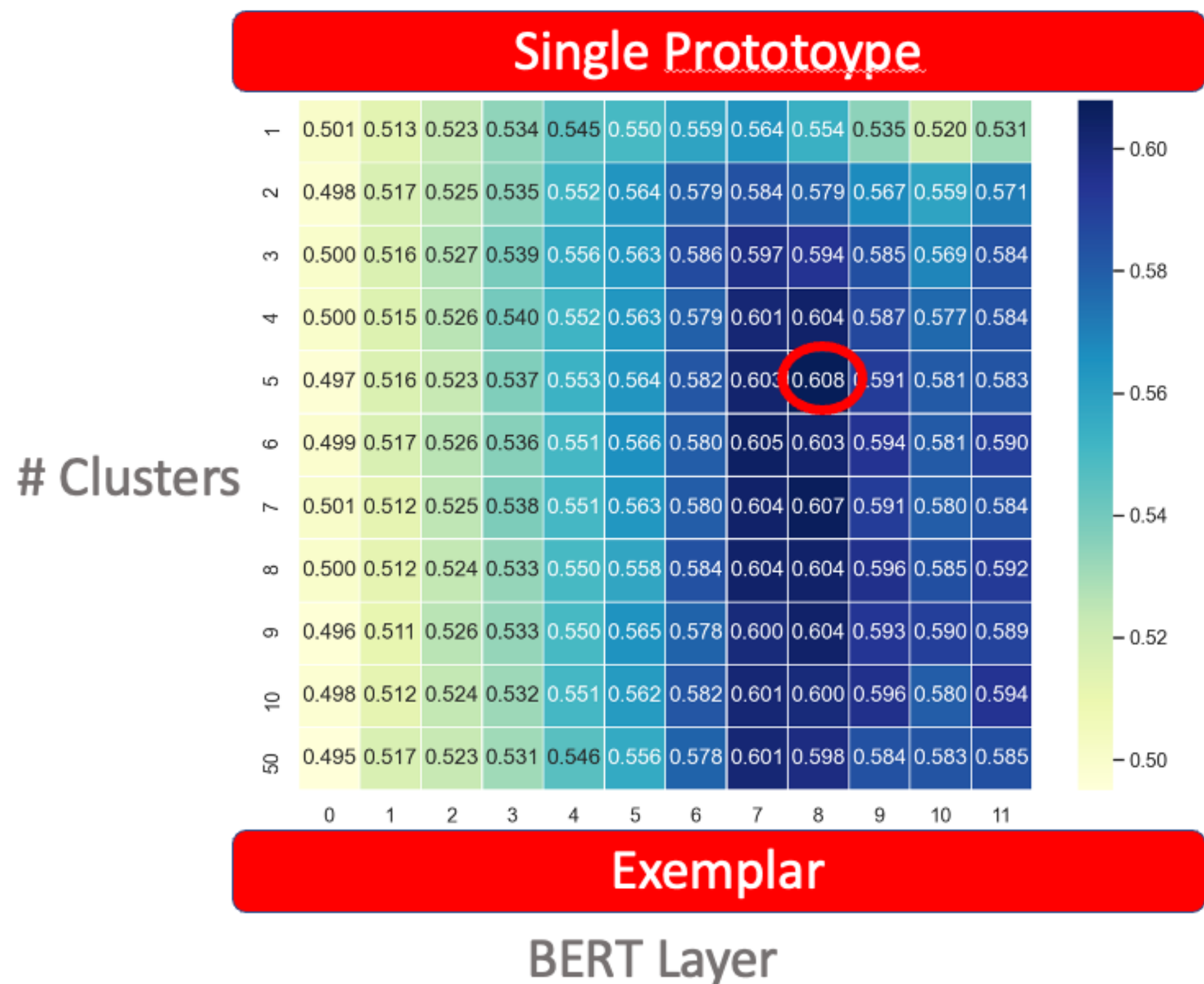
Recovering the taxonomic/topical similarity distinction

- Reminder:
 - taxonomic similarity (“similarity”): words that are close together in an ontology, like “cat” and “dog”
 - topical similarity (“relatedness”): words that co-appear in stories, but don’t refer to similar kinds of entities, like “dog” and “kennel”
- Count-based distributional model: Context window size determines type of similarity
- Word token embeddings: **Different BERT layers approximate the two kinds of similarity**
- Chronis and Erk 2020, “When is a bishop not like a rook? When it's like a rabbi! Multi-prototype BERT embeddings for estimating semantic relationships.”
- BERT-base, approximating human similarity ratings with two datasets:
 - SimLex (Hill et al 2015) for taxonomic similarity (“similarity”)
 - MEN (Bruni et al 2013) for topical similarity (“relatedness”)

Recovering the taxonomic/topical similarity distinction (BERT-Base)

Taxonomic sim. best performance: layer 7/8

Topical sim. best performance: last layer



No clear recommendation on #clusters

Inspection of multi-prototype embeddings: the noun “fire”

Again, clusters don't just reflect word sense, but also stories that people often tell with a word

Erk & Chronis 2023, “Word embeddings are word story embeddings (and that's fine).” British National Corpus data, 5 clusters for “fire”

- **Cluster 0: emotion, transformative fire**

- Changez said nothing, but shuffled backwards , away from the fire of Anwar's blazing contempt
- Never again, ... would the ceremonies be performed; gone were the offerings, the blood-shedding, the fire and incense

- **Cluster 1: destructive fire**

- There was a fire at Mr's store and they called it arson.
- An electrical short circuit started the fire, they think.

- **Cluster 2: artillery**

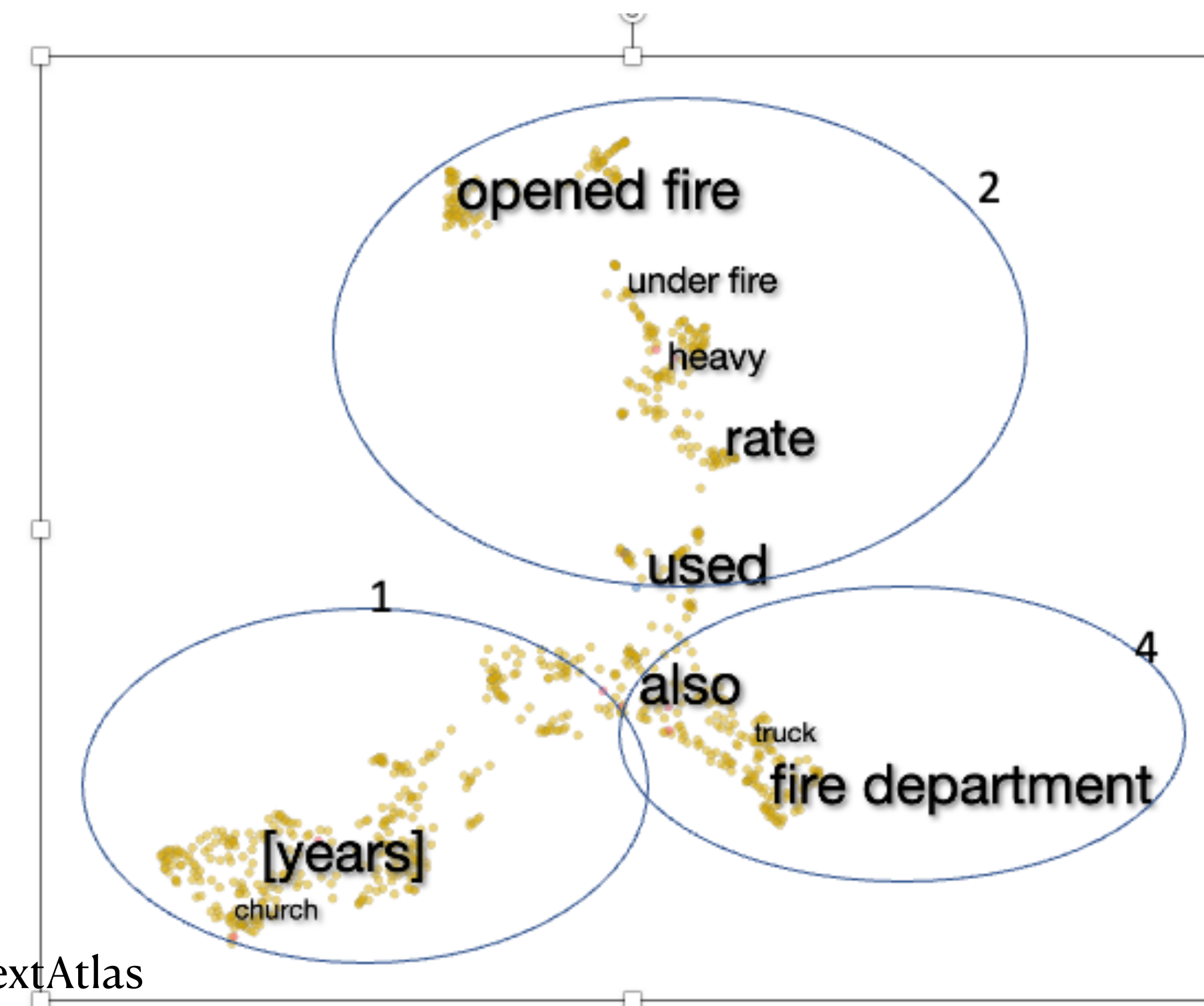
- small-arms fire

- **Cluster 3: hearth**

- or reading in the shadow of a fire;
- The bar is warm and cosy, with an open fire and oak beams.

- **Cluster 4: compound nouns/fire control**

- half a layer of fire cement
- fire alarms were installed



Multi-prototype embeddings are useful

- Soper and Koenig 2023, “Modeling the Role of Polysemy in Verb Categorization”
 - Data: Humans categorize verbs by dragging them into groups (Majewska et al 2021)
 - Modeling: Multi-prototype embeddings are much better than word type vectors, and better than single-prototype BERT
 - Possible reason: Vectors are less “confused” by mixture of contexts from senses
- But importantly, the paper also gets at the nature of word meaning

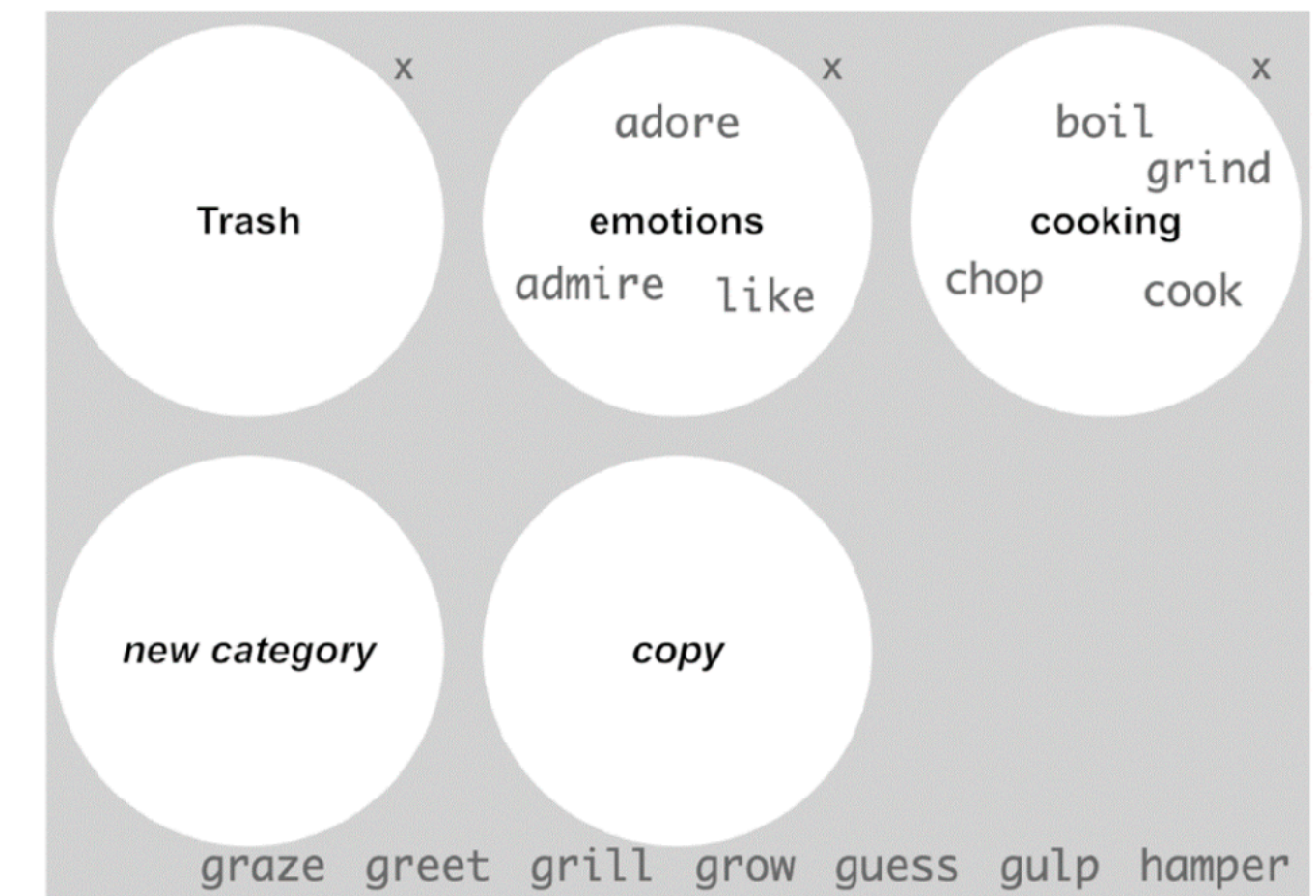


Figure 1: Screen interface for the SpA-Verb sorting task (Majewska et al. 2021)

Embeddings and the nature of polysemy

- What is the nature of polysemy? Some possibilities:
 - A. Single, underspecified meaning underlying all the polysemous senses
Specific sense selected in context by selection of specific features from the underspecified representation, or other interaction of context with the single meaning, e.g. in Pustejovsky's Generative Lexicon
 - B. Fixed list of (related) senses, stored separately
 - C. There is no lexicon, senses are created by the context
How I understand Elman 2009: Humans have knowledge about typical events, and that together with the uttered word creates meaning in context
 - D. A mixture of pragmatic modulation and stored ambiguity: Humans store sufficiently frequently observed usages (Recanati 2017)
- Soper and Koenig: static vectors implement A, contextualized embeddings as implementing C (but I think contextualized embeddings could also be doing D)

Things to know about BERT embeddings

Towards a best practice for using token embeddings:

Things to figure out

- Should you average over layers, or use one layer?
 - Our experience: use single layer to control level of similarity, but no consensus on this
- Should you average over tokens, or keep them separate?
 - Our experience: Don't average over all tokens, or you lose polysemy information. Use multi-prototype clusters. They have been found to be very good in several studies now, but again, no consensus.

Towards a best practice for using token embeddings:

Things to figure out

Ethayarajh 2019, “How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings”

- **Anisotropy** problem:

Embeddings don't fill the entire semantic space, they tend to live in a narrow cone.

Result: They all are highly similar to each other

Timkey and van Schijndel 2021, “All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality”;

- **Dimensions differ in granularity, some have much higher variability than others**
- **When computing cosine similarity, these “rogue dimensions” dominate all, hence the anisotropy**
- They show some simple fixes, including z-scoring:
 - For each dimension d of the space: compute mean and standard deviation of values across all vectors
 - Vector with value x on dimension d : normalize to $x_{\text{new}} = (x - \text{mean}) / \text{stdev}$

Towards a best practice for using token embeddings: Things to figure out

Mickus et al, 2020, “What do you mean, BERT? Assessing BERT as a Distributional Semantics Model”:

- **BERT embeddings contain “I am from sentence₁/sentence₂” information!**
 - BERT is trained not only on masked word prediction, but also on next sentence prediction: given sentences s_1 , s_2 , does s_2 directly follow s_1 in the corpus?
 - Consequence: Embedding also contains information “I am from sentence₁” or “I am from sentence₂”
 - This distorts similarity predictions
 - Workaround: for lexical semantics purposes, only feed in one sentence at a time, if possible
- **Does BERT over-emphasize lemmas?**
 - Do tokens of a single lemma cohere in space? Clustering analysis
 - One in four tokens would be better assigned to another cluster than their own lemma.
 - They say this is low cohesion. But isn't this cohesion too high? We would want tokens of a polysemous word to be close to their synonyms, not to their lemma centroid
 - I haven't followed up on this to see if BERT underestimates cross-lemma synonymy, but this should be checked

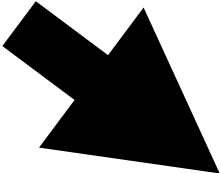
Multilingual BERT

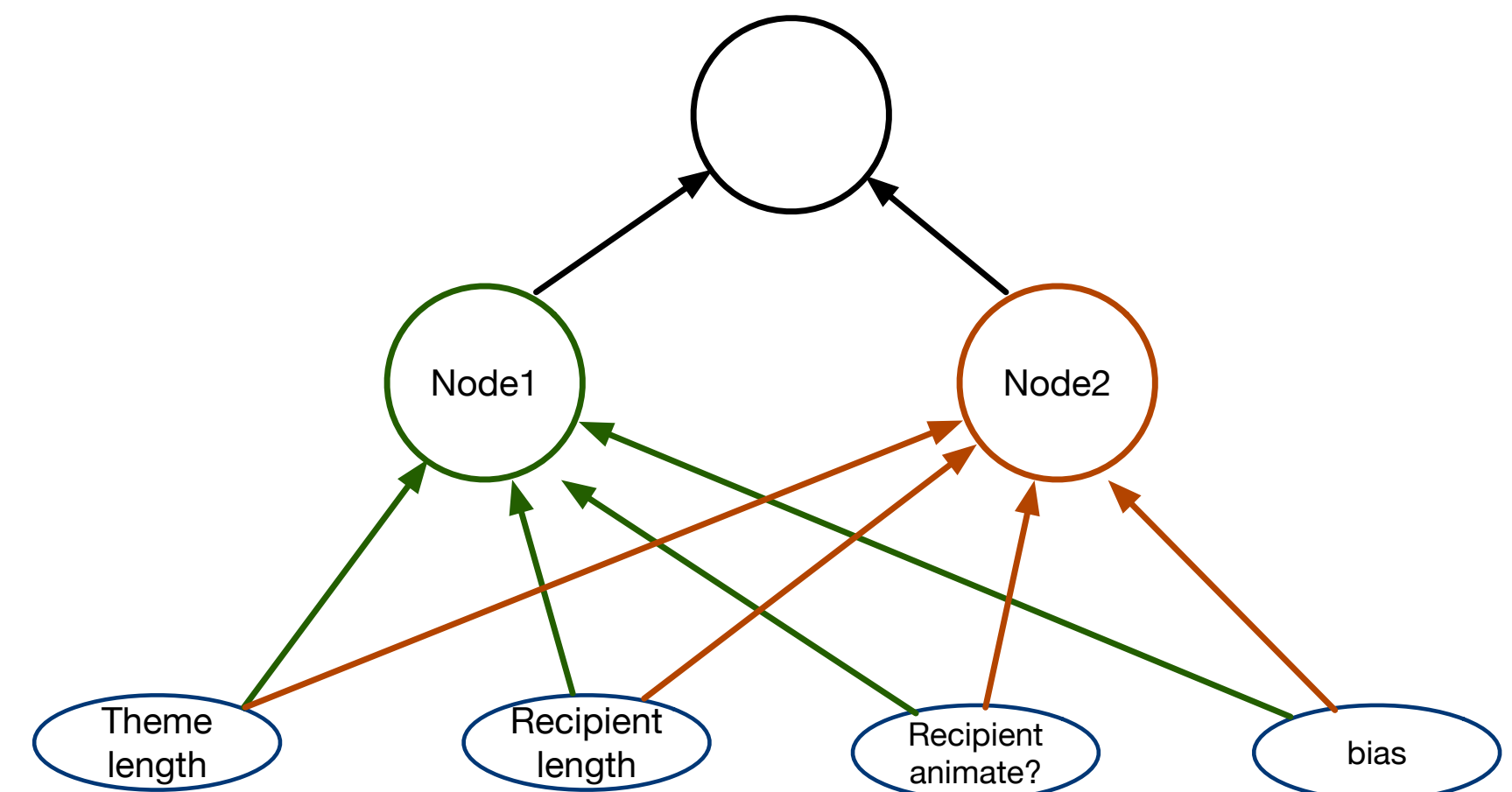
Multilingual BERT, MBERT: pre-trained on monolingual corpora in 104 languages, shared vocabulary of word pieces (that is all that is shared)

- Pires et al 2019, “How multilingual is Multilingual BERT?”
 - Interesting properties: Use MBert embeddings as a basis for training NLP tools for one language, then transfer to another language without training data: works quite well
 - Works best with languages with some lexical overlap, but some success even with no lexical overlap at all
- But in lexical semantics tasks, MBERT is usually worse than a BERT trained only on the target language
 - Example: Vulic et al 2020, “Probing Pretrained Language Models for Lexical Semantics”

Probing

Probing

- Conneau et al 2018, “What you can cram into a single $\$ \& ! \# ^*$ vector: Probing sentence embeddings for linguistic properties” (titled in honor of my colleague Ray Mooney), Tenney et al 2019, “BERT Rediscovered the Classical NLP Pipeline”
 - Run a model over a sentence to compute embeddings
 - Train a classifier (for example this ) to guess only from embeddings:
 - part of speech, verb tense, does this word denote an organization, do these two words co-refer?
 - If this works, conclude the information is encoded in the embedding



Probing for semantic features

Petersen and Potts 2023, “Lexical Semantics with Large Language Models: A Case Study of English break”

- Relevant features from the linguistic literature:
Transitive (John broke the vase), Unaccusative (the vase broke),
Metaphorical (we broke the law), violation (we broke the law), ...
- Probing: which of these features are encoded in token embeddings of “break”?
 - Low layer embeddings don’t strongly encode those features, but high layers do
- Clustering: Does it match manual sense annotation?
 - Mostly yes. Interestingly, outliers tend to be sense blends

Interesting discussions in the Petersen and Potts paper

- Views of the lexicon in theoretical linguistics and distributional models
 - **high dimensionality**: yes for linguistics, and both type and token embeddings
 - **contextual modulation**: “A word sense will be influenced by its immediate morphosyntactic context as well as the broader context of use.” yes for linguistics and token embeddings, no for type embeddings
 - **discreteness**: features in lex. sem. entries are discrete, and highly structured. yes for linguistics, no for distributional models
- Potential and limitation of this analytical approach
- “Overall, then, a theory of lexical semantics that draws heavily on LLMs as investigative tools, and even as ways to state theoretical ideas, is likely to become more usage-based than traditional theories would assume. This could lead them to focus less on pure representation and more on what is actually communicated between people when they communicate.”

A general-purpose probe: Mapping word tokens to features

Chronis et al 2023, “A Method for Studying Semantic Construal in Grammatical Constructions with Interpretable Contextual Embedding Spaces.”

- Probing classifier: map word tokens to interpretable feature space: semantic property collections from psychology & neuroscience
 - McRae et al, Buchanan et al: “What are properties of a strawberry? List up to 5”
-> is sweet, is a fruit, has pits, is red, ... #features: 2500 (McRae), 4000 (Buchanan)
 - Binder et al: “When you think of a strawberry, do you think of it as something you can easily see?” 65 features corresponding to particular brain regions

A general-purpose probe: Mapping word tokens to features

Predicted features for “fire”, 5 multi-prototype centroid embeddings:

Buchanan		Binder		McRae	
1. figurative	animal, color, light, fire, burn	1. figurative	Color, Needs, Harm, Cognition, Temperature	1. figurative	has_legs, is_hard, different_sizes, has_4_legs, is_large
2. destructive	destroy, build, cause, break, person	2. destructive	Unpleasant, Fearful, Sad, Consequential, Harm	2. destructive	different_colors, a_mammal, made_of_paper, made_of_cement, inbeh_-_explodes
3. artillery	act, weapon, kill, loud, human	3. artillery	UpperLimb, Communication, Social, Audition, Head	3. artillery	a_weapon, used_for_killing, made_of_metal, is_loud, used_for_war
4. cooking	hot, food, wood, burn, heat	4. cooking	Pleasant, Needs, Happy, Near, Temperature	4. cooking	found_in_kitchens, used_for_cooking, requires_gas, an_appliance, is_hot
5. N-N compounds	person, place, work, office, law	5. N-N compounds	Biomotion, Face, Speech, Body, Unpleasant	5. N-N compounds	has_doors, used_for_transportation, a_bird, has_feathers, beh_-_eats

Table 3: The most distinctive features for each prototype of *fire* multi-prototype embeddings, in each of the three interpretable semantic spaces.

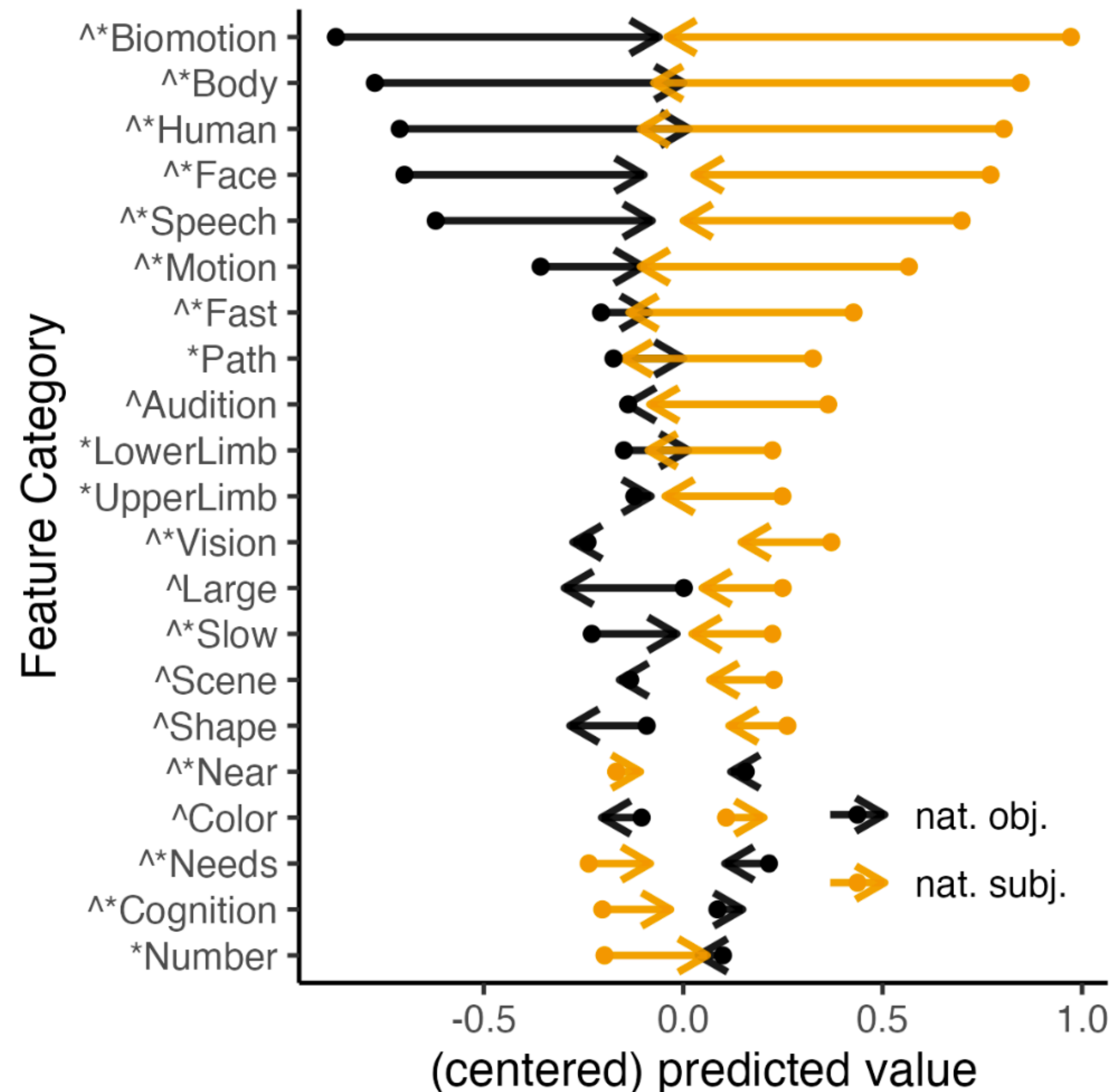
Probing construal in constructions with Binder and Buchanan features

Very similar-looking constructions can have subtle semantic differences

- three beautiful days in New York
- a beautiful three days in New York (AANN construction)
- In the linguistic literature: AANN acceptable only when the noun behaves as a single collective unit and is, in effect, more semantically similar to a unit of measurement than it would be in the unmarked construction.
- Can we see this in the embeddings?
 - Buchanan, 1,000 sentence sample, comparing AANN with default equivalent, top 5 features of head noun token:
 - AANN (relative to default): measure, one, green, unit, grow.
 - default (relative to AANN): animal, leg, child, human, please

Probing construal in constructions with Binder and Buchanan features

- Binder features:
properties of words in
subject and object position
- Normal:
“The chef cut the onion”
- Switched:
“The onion cut the chef”
- How much do properties
depend on the word, how
much on its argument position?



The language model as an artificial participant

Large Language Models (LLMs) as artificial participants

- LLMs do not just provide a static semantic space, they are trained to predict.
So they can also answer questions.
 - Do LLMs learn syntactic knowledge that resembles that of humans? Will they make similar mistakes as humans?
 - Linzen et al 2016, “Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies”
 - Left-to-right text producing model, “The keys on the cabinet _____” :
with what probability will the model produce “is”? “are”?
 - Many papers on syntax in LLMs: Linzen, Warstadt, Bowman, Dillon, Lappin...
 - Can also use masked prediction task: “The keys on the cabinet [MASK] for the garage”
 - Not restricted to syntax: e.g., semantics of constructions: Weissweiler et al, “The Better Your Syntax, the Better Your Semantics? Probing Pretrained Language Models for the English Comparative Correlative”

Large Language Models (LLMs) as artificial participants

- Problem with using LLMs as artificial participants: If they give good answers, does that mean they have understood? Or are they exploiting some surface text patterns?
 - (Does this problem also apply when we use LLMs as a semantic space?)
- Striking in a series of papers by Allyson Ettinger, e.g, Misra et al 2023, “COMPS: Conceptual Minimal Pair Sentences for testing Robust Property Knowledge and its Inheritance in Pre-trained Language Models”
 - The model is good at distinguishing these. So, does it have robust property knowledge?
 - a. A wug is a robin. Therefore, a wug can fly.
 - b. *A wug is a penguin. Therefore, a wug can fly.
 - Apparently not, because it becomes confused by distractors:
 - A wug is a penguin. A dax is a robin. Therefore, a wug can fly
 - Did it just go with the nearest bird-like word?
- Lesson learned: always consider possible shortcuts, and test for them

Treating the model (almost) like a study participant

- example: Weissweiler
- syntax, eg Tal Linzen
- watch out: Allyson Ettinger's results