On the distributional hypothesis – why it works and how it doesn't

Janet Pierrehumbert

8 August, ESSLLI 2024

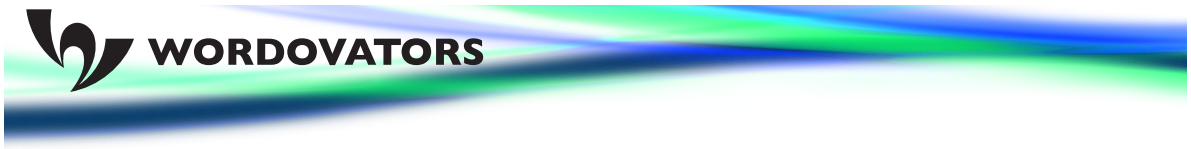
Outline

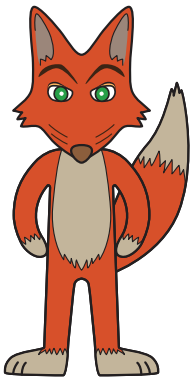
Outline

Words: the locus of associations between sound (or spelling) and meaning.

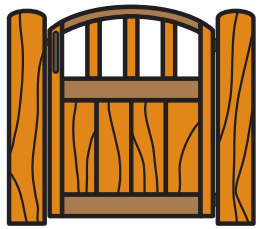
- Phonemic principle: Phonemes ($\sim$ letters) have no meaning in themselves, but can be combined in different ways to express meanings.
- Example: *pats*, *past*, *spat*.
- Words can be similar in form but not in meaning: *bet*, *pet*; *fact*, *faction*.
- They can be similar in meaning but not in form *guru*, *mentor*.

Form-meaning associations must be learned.



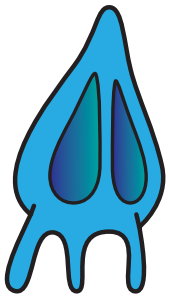


fox
azeria
kettu
zorro



gate
atea
portti
puerta

mushroom
onddo
sieni
seta



?

How is the form obtained? How is the meaning obtained?

- Extensional semantic theory (basis for theories of Skinner and Quine).
  - The child observes articulatory and acoustic events and builds phonological generalizations about them.
  - The child observes objects and actions, and builds semantic generalizations about them.
  - Statistical associations between these cognitive representations of the real world are stored in memory.
- But this is not how current NLP algorithms work.

The distributional hypothesis for word meanings

- Assumption: the meanings of words can be known by the company they keep. The company is other words.
- Foundation for all static word embeddings (Latent Semantic Analysis, word2vec, Glove, etc) and all neural network NLP models.
- Both Word2vec and the modern transformer models (BERT, GPT etc.) are trained using Masked Language Modelling: Try to predict a hidden word from its context.
  - Ex1: You don't have to never eat meat, just treat red [MASK] like caviar or gold-dusted cake.
  - Ex2: You are conflating the issue. Slavery was not moral but it was [MASK] legal.

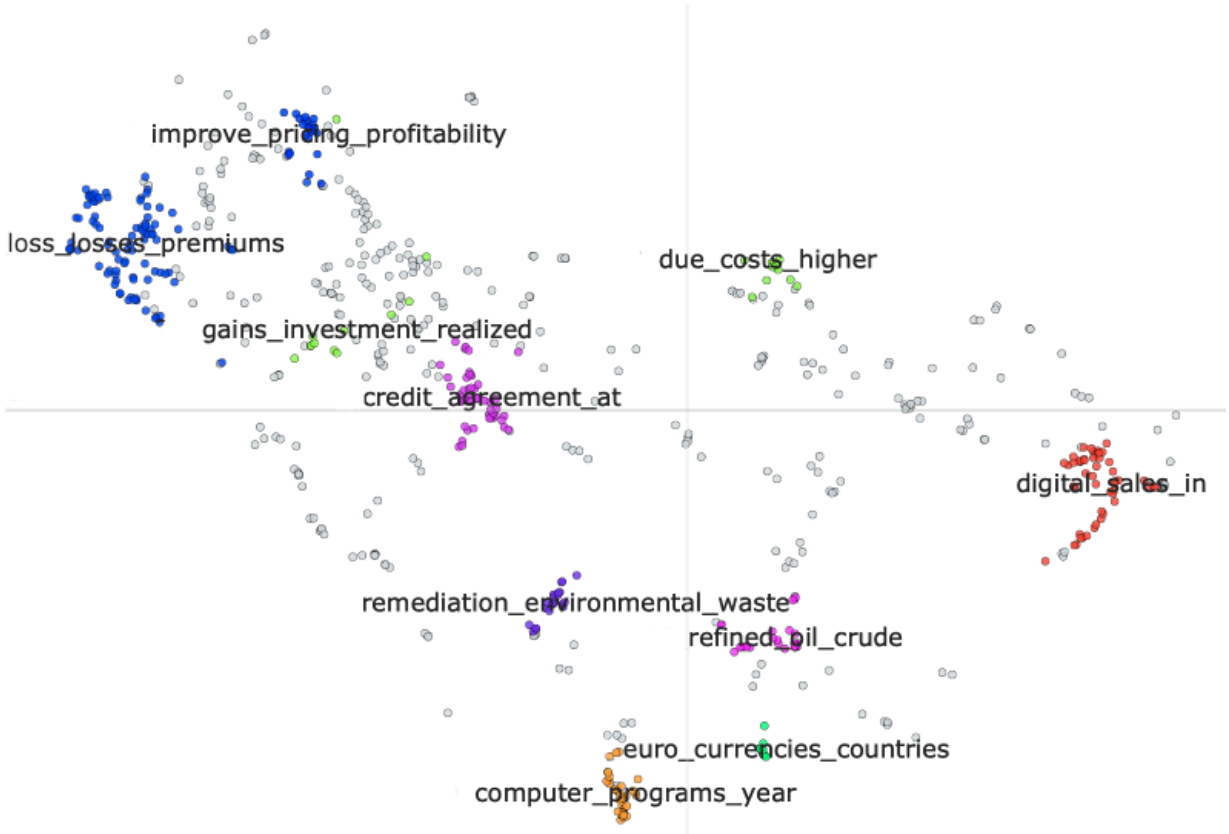
More precisely

- Training: Optimize the weights in a NN model so that the word that DID occur is assigned a higher probability than a random selection of words that did NOT occur.
- Result: A continuous meaning hyperspace of hundreds of dimensions. The vectors for words that are similar in their cooccurrence patterns are closer together (according to cosine similarity); vectors for dissimilar words are more orthogonal.
- Approach can be applied to build representations of phrases or topics.

Note: No extensional semantics. No external reality.

An example outcome

2D projection of semantic clusters for sentences in Securities Exchange Commission filings from a project on forecasting credit rating changes (Drinkall et al, 2024)



Why does this work?

A Big Question of Generative Linguistics: How can babies learn language systems without overt feedback?

- Humans – including babies – are continually attempting to predict what will come next.
- If their prediction is wrong, they are surprised and update their mental models.
- LLMs models are trained in this cognitively realistic manner, using the Masked Language Modelling task.
- The model attempts to predict words that were hidden. Discrepancies between predictions are actual outcomes are used to update the model. Usually using the information-theoretic measure of cross-entropy, a close relative of surprisal.

Sources of prediction

- Morphosyntax
  - Local constraints on word sequencing.
  - Example: placement of function words relative to content words.
- Topic of discussion.
  - Different topics of discussion activate and deactivate vocabulary fields.
  - Quantified in computational linguistics via measures of "burstiness" (which I will explain).

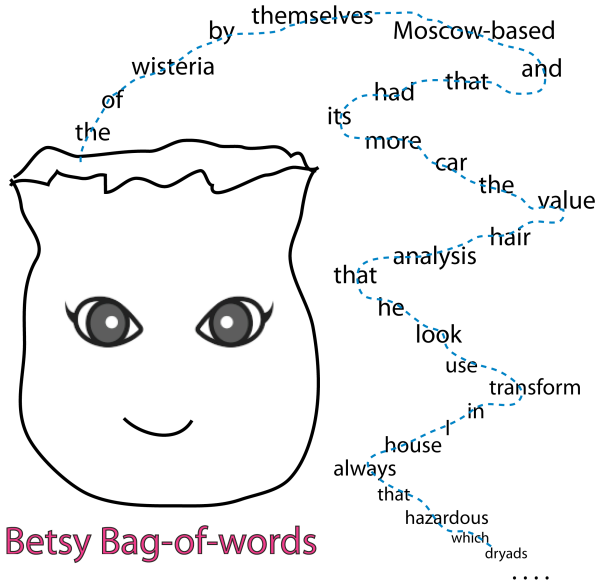
Outline

Example of a burst

Darwin occurs many times within the following text, because the text is about Darwin.

Darwin spent the summer of 1825 as an apprentice doctor <...> Darwin found lectures dull and distressing, so he neglected his studies <...> In Darwin second year at the university he joined the Plinian Society, a student natural-history group <...> One day, Grant praised Lamarck's evolutionary ideas. Darwin was astonished by Grant's audacity <... > Darwin was rather bored by Robert Jameson's natural-history course <..> He learned the classification of plants, and assisted with work on the collections of the University Museum, one of the largest museums in Europe at the time ...

Quantifying burstiness.

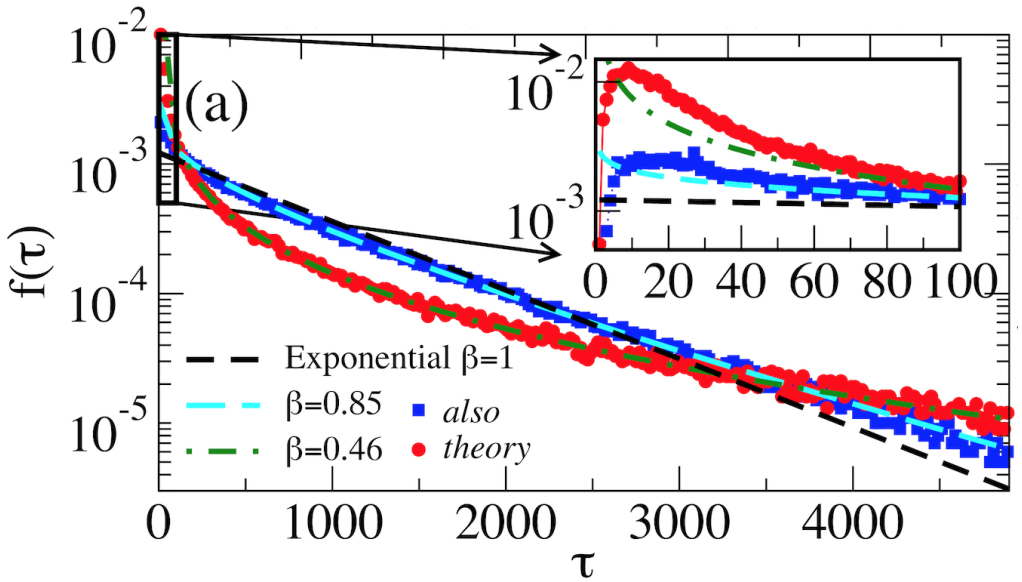


Baseline: Naive Bag of Words model for constructing word sequences. This is a Poisson process.

- The distribution of counts of words in documents is predicted to be a Poisson counting distribution. Deviations are mainstay of document retrieval and topic classification.
- Altmann, Pierrehumbert & Motter (2009): Model of recurrence time distributions for all 2000 words occurring 10,000 times or more in some Usenet discussion groups.

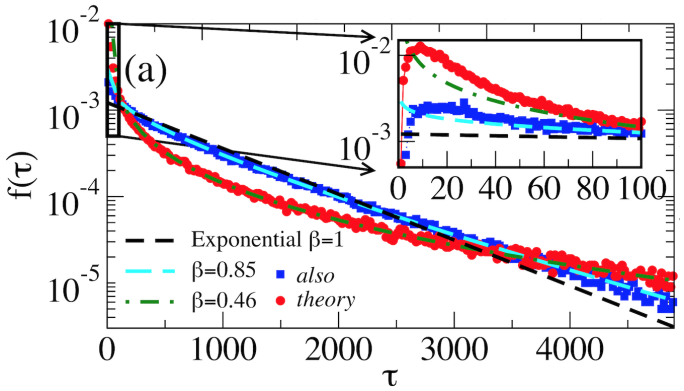
Recurrence time distributions for two words with the same frequency in one group: theory and also

μ is the average time between tokens of the same word, and τ is time. Poisson Process predicts that the probability distribution of τ is exponential: $f(\tau) = \mu e^{-\mu\tau}$.... What is it really?



Linear x axis, log y axis $\rightarrow$ exponential function appears as a straight line. Left corner blown up in inset.

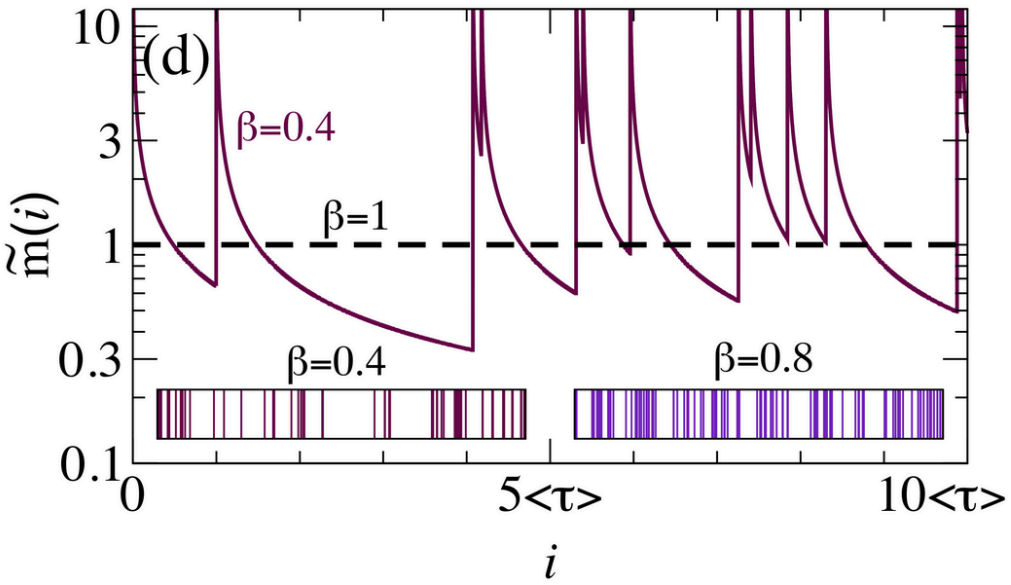
Some observations



- The distributions are not exponential. (They are stretched-exponential, with one free parameter β.)
- The words differ in β.
- theory, a topical noun, has a lower β than also, an adverb.
- NOTE: Inset shows that the frequency boost peaks at a lag of $\sim$10. At very short lags, the word frequency is below the model prediction (because of the syntax).

General pattern as seen in time:

Likelihood of using some some topical word, in relation to a constant overall likelihood based on its average frequency.



- As soon as the word is used, the likelihood of (re)use shoots up. NB order of magnitude effect, for this example.
- Then it decays slowly (by a power law).

Relation of burstiness to semantic categories

.

The contexts for Darwin (a named entity) are very distinctive because a window around Darwin includes words that are less used if the topic is very different. What about other kinds of words?

Burstiness	**Description**	**Examples of words**
Most	Entities	Africa, Bible, Darwin
Very	Predicates and Relations	blue, die, religion
Somewhat	Modifiers and Operators	believe, everyone, ten, tall
Least	Higher Level Operators	hence, not, certainly, often

(Altmann, Pierrehumbert, and Motter, 2009)

Outline

Scalar adjectives

"The Dakota is a tall building." TRUE or FALSE?

"tall" does not designate any particular degree of height. It asserts a relationship between a value on a contextually relevant scale, and what value would be typical for the context.

Problem for the distributional hypothesis?

Yes: Consider good.

- What scale is relevant? Tastiness? Durability? Moral worth...?
- What is the typical value? The bar is higher for some domains of comparison than others.
- Result: contexts for scalar adjectives are less distinctive than for named entities.

Maybe no:

- Selectional restrictions with nouns.
- Common sentiments and formulaic expressions.

How well do LLMs understand scalar adjectives?

Results on identifying the correct scale from Lin et al. (LREC 2024). Accuracy for three benchmarking sets:

Models	DM	CD	WK
fast-text	0.811	0.656	0.983
BERT-b	$0.829_{\pm 0.010}$	$0.797_{\pm 0.010}$	$\mathbf{0.997}_{\pm 0.004}$
BERT-l	$\mathbf{0.853}_{\pm 0.007}$	$\mathbf{0.805}_{\pm 0.011}$	$\mathbf{0.997}_{\pm 0.006}$
RoBERTa-b	$0.668_{\pm 0.014}$	$0.705_{\pm 0.007}$	$0.906_{\pm 0.018}$
RoBERTa-l	$0.777_{\pm 0.011}$	$0.757_{\pm 0.008}$	$0.977_{\pm 0.010}$

Table 2: Direct scale membership probing results, subscripted numbers are standard deviation in ten runs. Best results per dataset are in bold.

Observation: Good but not perfect.

Results on identifying the intensity on the scale

Compared several methods of calculating intensity. Success is moderate.

Model	Method	DM	CD	WK
fast-text	-	0.637_{WK}	0.685_{DM}	0.836_{CD}
BERT-b	G&A	0.646_{CD}	0.735_{DM}	0.902_{DM}
	Ours	0.639_{CD}	0.706_{DM}	$\mathbf{0.967}_{DM}$
BERT-l	G&A	$\mathbf{0.695}_{CD}$	0.731_{DM}	0.918_{DM}
	Ours	0.673_{CD}	0.727_{DM}	0.902_{DM}
RoBERTa-b	G&A	0.557_{WK}	0.645_{DM}	0.820_{DM}
	Ours	0.648_{CD}	0.748_{DM}	0.934_{DM}
RoBERTa-l	G&A	0.595_{CD}	0.682_{DM}	0.836_{DM}
	Ours	0.664_{CD}	$\mathbf{0.752}_{DM}$	0.934_{DM}

Observation: Worse performance than for identifying the scale.

Why is identifying the intensity harder?

- Building on L. McNally (2017)
  - The setup of the benchmarking sets assume that the same scale is pertinent to all nouns.
  - E.g. cold, cool, warm, hot
  - But the space of potential contrasts differs by context. Compare cold/hot soup versus cold/?hot welcome. So the adjectives are not neatly lined up in the semantic hyperspace.
- But also
  - Insofar as they are lined up, the speaker makes their choice based on external reality. Which the LLMs have no access to.

Next level: Scalar Adverbs. Lorge et al. BlackBox NLP, EMNLP 2023

Category	Adverbs
MODALITY (14.8%)	*{maybe, perhaps, possibly}, arguably, probably, actually, certainly, definitely*
FREQUENCY (5.3%)	*never, occasionally, sometimes, often, generally, usually, frequently, always*
DEGREE (46.8%)	*hardly, slightly, basically, pretty, quite, very, really, completely*

"He is certainly slightly angry" = "I have a high level of belief that the extent to which he is angry is relatively low on a contextually relevant scale of anger."

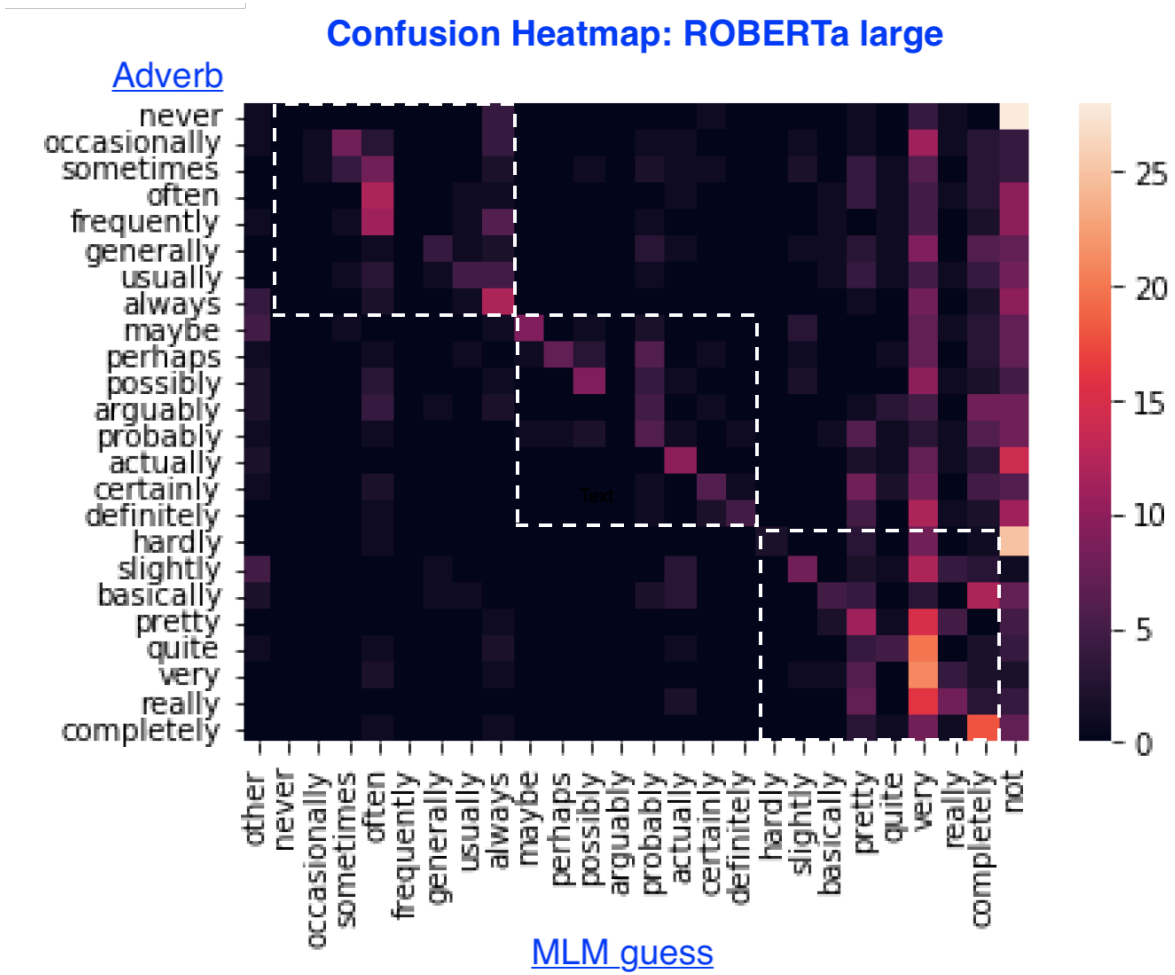
Are scalar adverbs predictable in context?

- MLM task. Left-hand context up to 40 words in Reddit posts.
- Example: "You are conflating the issue. Slavery was not moral but it was [MASK] legal.
- Measure of success: MRR: how close to the top rank (out of 24) was the true answer?

Results (Mean Reciprocal Rank).

	BERTb	BERTl	ROBERTl	GPT2
FREQ	0.11	0.15	0.21	0.04
MODAL	0.09	0.11	0.18	0.05
DEGREE	0.20	0.24	0.28	0.06
Avg.	0.14	0.17	0.22	0.05

- $MRR = 0.10$ means that the correct choice was on the average ranked tenth.
- Contextual cues for the choice of scalar adverb are very weak (weaker than for adjectives, as predicted).
- The distributional hypothesis is not very applicable to high-level operators.



We want: Bright diagonal with a block structure reflecting confusions within the same semantic class.

We see: Little semantic class structure. Bias towards very and not.

Also tested:

Can LLMs produce logically correct answers for entailments involving scalar adverbs?

- If it is often special, it is at least [MASK] special.
- If it is [MASK] special, it is at least sometimes special.

Short answer: No. Even with generous allowance for multiple correct answers.

Outline

Meta-level reasoning, or semantic ascent.

- You form abstractions over experiences.
- Now you have a mental map or inventory of abstractions.
- You view those as a domain for the next level of abstraction.

Case 1: Referring to text

PERSON: Please translate into Spanish the sentence "What is the capital of France?"

CHATBOT: Paris.

Moderating Hatespeech

Röttger et al. (2021) HateCheck: Functional Tests for Hatespeech Detection Models.

- Diagnostic dataset for hatespeech moderation algorithms.
- 29 linguistically informed cases. Short sentences constructed by templates with exactly one relevant features.
- Reports results as of 2021 for BERT (B-D and B-F; finetuned on two different labelled datasets), Google Jigsaw Perspective; TwoHat SiftNinja.

					B-D	B-F	P	S-N
Counter speech	**F20**: Denouncements of hate that quote it	"Statements like '[IDENTITY] are scum' are deeply hurtful."	non-hate	173	*26.6*	*32.9*	*15.6*	**79.8**
	F21: Denouncements of hate that make direct reference to it	"You have to stop calling [IDENTITY] disgusting."	non-hate	141	*29.1*	*29.8*	*18.4*	**79.4**
Slur usage	**F7**: Hate expressed using slur	"You are just a [SLUR] to me."	hateful	144	60.4	*41.0*	**66.0**	50.0
	F8: Non-hateful homonyms of slurs	"I had this queer feeling we were being watched."	non-hate	30	66.7	**70.0**	63.3	*33.3*
	F9: Reclaimed slurs	"I'm the freshest n*gga in the room."	non-hate	81	***39.5***	*33.3*	*28.4*	*18.5*

- Testing setup with chance level of 50%. Results in red are worse than chance.
- S-N appears to do well on counterspeech because it is strongly biased towards classifying text as non-hateful. Gets 9% right overall on examples that actually are hateful.

A more wide-ranging effort

Thrush et al. (2024) "I am a strange dataset". Asked GPT-type models to complete sentences designed by linguists. Partial list of cases:

Tag	Count	Example		
		Beginning	**False End**	**True End**
Negation & Scope	94	The last word you will read before the period is	not "dog".	actually "dog".
Numerical Operations	62	The number of words in this sentence is	eight.	nine.
Location of Element	55	This sentence	nothing wrong has with the word order.	something wrong has with the word order.
Sub-Word	48	Evary werd en thas sentance iz mispelled	including the words at the end.	except the words at the end.

- The results were terrible. Crowdworkers were pretty good.
- It looks like all cases involve self-referential sentences. However, that's not the only type of meta-level active in language.

Word formation and the type-token distinction

- Best-available models explain how new complex words can be generated by analogy to existing ones.
- connective:connectivity::sensitive:sensitivity::...::cormasive:??
- ?? = cormasivity? cormasiveness?
- The competition rests on how many similar words in the lexicon exhibit one pattern versus the other.
- That's word types, distinct words in the lexicon. Not word tokens, examples of words in running text.
- The child, children pattern fails to generalize no matter how many times you see the words.

Morphological capabilities of GPT-J and GPT4

Work in progress by V. Hofmann et al.

- Competition between -ity and -ness after four suffixes.
  - -ish always takes -ness
  - -able almost always takes -ity
  - -ous weak tendency towards -ness
  - -ive weak tendency towards -ity
- Materials: the PILE (800GB open-access training set) provides 24,385 base word types ($>$ 130M tokens). Also small set of human judgments on nonce words.
- NLP models: GPT-J (open-source). GPT4 (proprietary).
- Cognitive models: Rule-based MGL (Minimum Generalization Learner) vs the analogy engine GCM Generalized Context Model; both trained on word types vs word tokens.

Results

- GPT-J and GPT-4 function like analogical systems.
  - For cases without variability, MGL and GCM match their behavior equally well; both predict no variability.
  - For cases with variability, the GCM is a better match.
- GPT-J and GPT-4 are trained on word tokens.
  - They match the GCM trained on word tokens.
  - But the GCM trained on word types matches the human judgments better (replicating previous literature).

In short

- The mental lexicon contains word types, which are generalizations over word tokens.
- This is a meta level.
- The GPT models lack this meta level. They don't really HAVE a mental lexicon!

Really? How come?

- GPT models and other LLMs have a fixed size codebook (some 300,000 to 100,000) items. Impossible to list all the English words!
- Less common words are encoded left-to-right as sequences of often meaningless coding units. E.g. BERT has superbizarre as superb+iza+rre (Hofmann et al. 2021).
- The meaning representations are distributed representations optimized over the contexts of the tokens of all the pieces.
- There is no way to "look down" on the model and "see" what words are there.

The problem with meta levels is not just the odd fringe of quotative use of language. It is central to the linguistic system.

Outline

Conclusions

- Strengths of the distributional hypothesis.
  - Without overt feedback, it supports surprisingly powerful language models.
  - Using contextualized prediction and surprisal is grounded in what we know about human cognition.
  - The continuous semantic space makes it possible to quantify semantic similarity in useful ways.
- Weaknesses:
  - Problems with the understanding of less-bursty words, such as operators. These have meanings involving generalizing over relations or relations of relations.
  - Lack of meta-level representations.

Thank you! Are there any questions?

Main References

- E.G. Altmann, J.B. Pierrehumbert, and A.E. Motter (2009) Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words, PLoS One 4(11), e7678). doi:10.1371/journal.pone.0007678
- Church, KW and Gale WA (1995) Poisson mixtures. Nat Lang Eng 1: 163-190.
- Drinkall, F, Pierrehumbert, JB and Zohren, S. (2024) Traditional Methods Outperforam Generative LLMs at Forecasting Credit Rating Changes. arXiv:2407.17624
- Hofmann, V., Pierrehumbert, J.B. and Schütze, H. (2021) Superbizarre is Not Superb: Improving BERT's Interpretations of Complex Words with Derivational Morphology ACL-ICLNLP 2021, 3594–3608. doi =10.18653/v1/2021.acl-long.279.

- Hofmann,V, et al (MS) Derivational Morphology Reveals Analogical Generalization in Large Language Models.
- Lin, F, D Altschuler, and J.B. Pierrehumbert (2024) Probing Large Language Models for Scalar Adjective Lexical Semantics and Scalar Diversity Pragmatics Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING).
- Lorge, I. and J.B. Pierrehumbert (2023) Not Wacky vs. Definitely Wacky: A Study of Scalar Adverbs in Pretrained Language Models. Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP. December, 2023, Singapore. DOI: 10.18653/v1/2023.blackboxnlp-1.23

- McNally, L. (2017) Scalar Alternatives and Scalar Inference Involving Adjectives: A Comment on Van Tiel et al. (2016). In Ostrove et al. (eds.) Asking the Right Questions: Essays in Honor of Sandra Chung. UC Santa Cruz.
- Partee, BH (1992) Syntactic categories and semantic type. In: Rosner, M and Johnson R (eds). Computational Linguistics and Formal Semantics. CUP, 97-126
- Röttger, P., Vidgen, B., Nguyen, D, Waseem, Z, Margetts, J. and Pierrehumbert, J.B. (2021) HateCheck: Functional Tests for Hate Speech Detection Models. ACL-IJNLP 2021, 41-58. doi = 10.18653/v1/2021.acl-long.4.
- Thrush et al. (2024) I am a Strange Dataset: Metalinguistic Tests for Language Models. arXiv.2401.05300v2