Compositionality in Semantic Spaces

Martha Lewis

ILLC
University of Amsterdam

2nd Symposium on Compositional Structures

Glasgow, UK

Outline

Outline

How can we understand new combinations of concepts?

> *When a male octopus spots a female, his normally grayish body suddenly becomes striped. He swims above the female and begins caressing her with seven of his arms.*
>
> *Cherries jubilee on a white suit? Wine on an altar cloth? Apply club soda immediately. It works beautifully to remove the stains from fabrics.*

Steven Pinker. The Language Instinct: How the Mind Creates Language (Penguin Science) (pp. 1-2).

How can we understand new combinations of concepts?

> *When a male octopus spots a female, his normally grayish body suddenly becomes striped. He swims above the female and begins caressing her with seven of his arms.*
>
> *Cherries jubilee on a white suit? Wine on an altar cloth? Apply club soda immediately. It works beautifully to remove the stains from fabrics.*

Steven Pinker. The Language Instinct: How the Mind Creates Language (Penguin Science) (pp. 1-2).

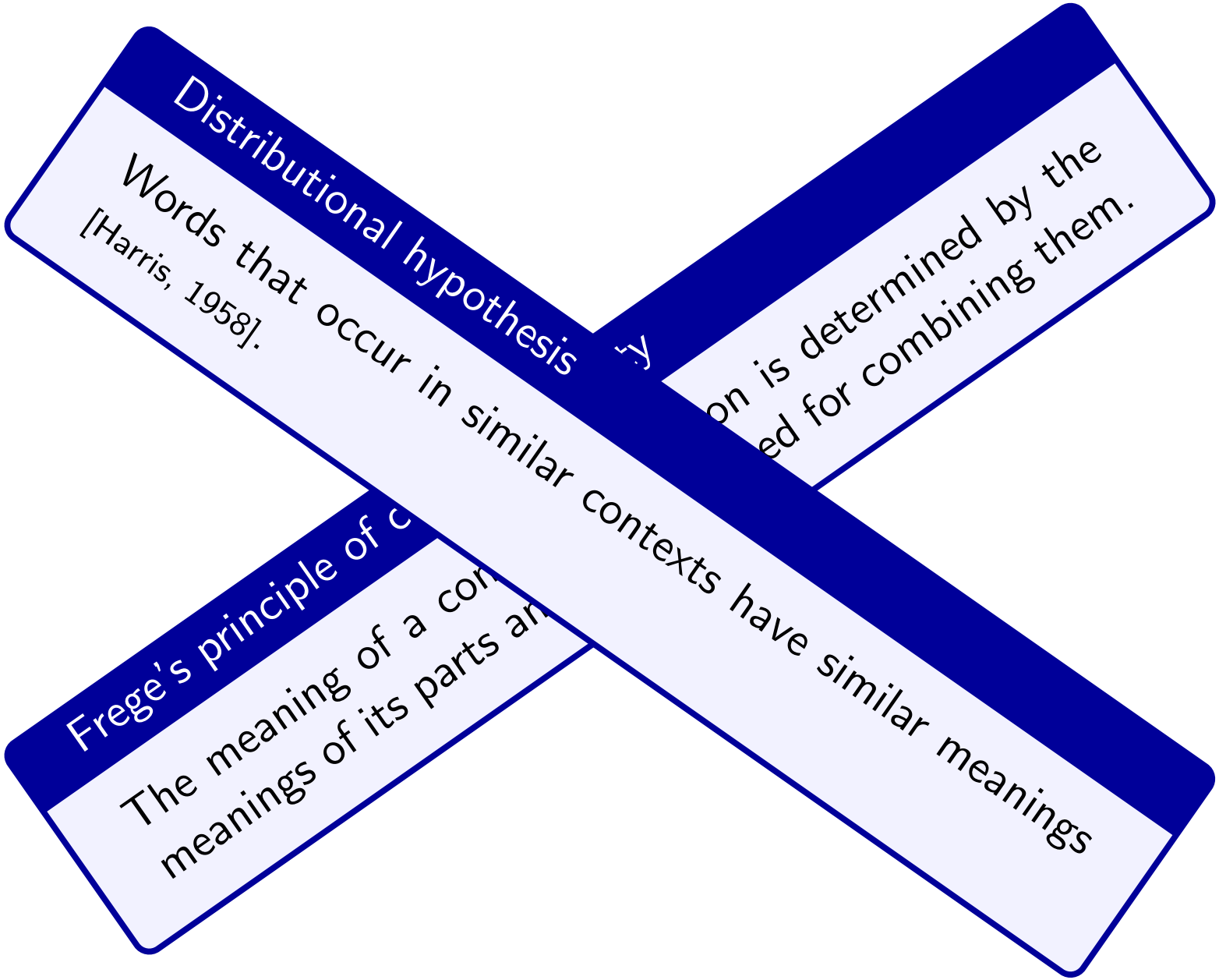
... And how can we get computers to do the same?

Compositional Distributional Semantics

Frege's principle of compositionality

The meaning of a complex expression is determined by the meanings of its parts and the rules used for combining them.

Compositional Distributional Semantics



The meaning of words

Distributional hypothesis

Words that occur in similar contexts have similar meanings [Harris, 1958].

U.S. Senate, because they are	?	, like to eat as high on the
It made him	?	.
sympathy for the problems of	?	beings caught up in the
peace and the sanctity of	?	life are not only religious
without the accompaniment of	?	sacrifice.
a monstrous crime against the	?	race.
this mystic bond between the	?	and natural world that the
suggests a current nostalgia for	?	values in art.
Harbor" in 1915), the	?	element was the compelling
an earthy and very	?	modern dance work,
To be	?	, he believes, is to seek one's
Ordinarily, the	?	liver synthesizes only enough
nothing in the whole range of	?	experience more widely
It is said that fear in	?	beings produces an odor that
megatons: the damage to	?	germ plasm would be such

The meaning of words

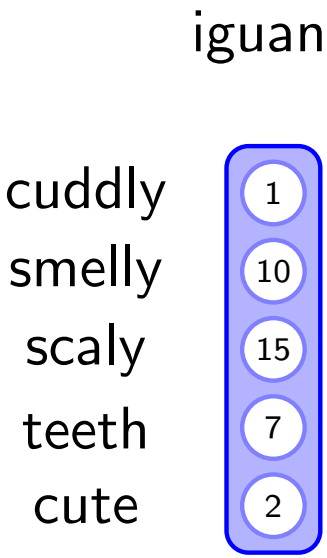
Distributional hypothesis

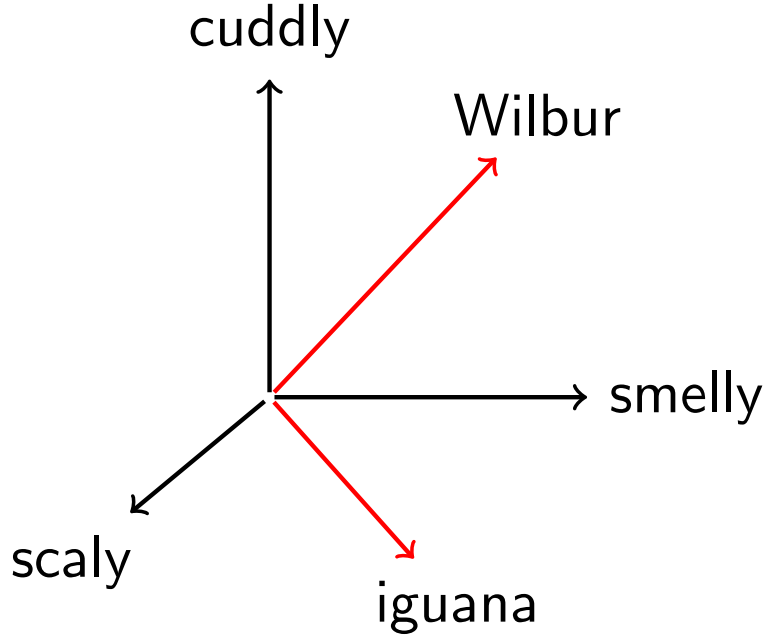
Words that occur in similar contexts have similar meanings [Harris, 1958].

U.S. Senate, because they are	**human**	, like to eat as high on the
It made him	**human**	.
sympathy for the problems of	**human**	beings caught up in the
peace and the sanctity of	**human**	life are not only religious
without the accompaniment of	**human**	sacrifice.
a monstrous crime against the	**human**	race.
this mystic bond between the	**human**	and natural world that the
suggests a current nostalgia for	**human**	values in art.
Harbor" in 1915), the	**human**	element was the compelling
an earthy and very	**human**	modern dance work,
To be	**human**	, he believes, is to seek one's
Ordinarily, the	**human**	liver synthesizes only enough
nothing in the whole range of	**human**	experience more widely
It is said that fear in	**human**	beings produces an odor that
megatons: the damage to	**human**	germ plasm would be such

Distributional Semantics

- Words are represented as vectors
- Entries of the vector are derived from how often the target word co-occurs with the context word





Similarity is given by cosine distance:

$$sim(v, w) = \cos(\theta_{v,w}) = \frac{\langle v, w \rangle}{||v||||w||}$$

The role of compositionality

Compositional distributional models

We can produce a sentence vector by composing the vectors of the words in that sentence.

$$\overrightarrow{s} = f(\overrightarrow{w_1}, \overrightarrow{w_2}, \ldots, \overrightarrow{w_n})$$

Three generic classes of CDMs:

- *Vector mixture* models [Mitchell and Lapata (2010)]
- *Tensor-based* models [Coecke, Sadrzadeh, Clark (2010); Baroni and Zamparelli (2010)]
- *Neural* models [Socher et al. (2012); Kalchbrenner et al. (2014)]

Applications (1/2)

Why are CDMs important?

The problem of producing robust representations for the meaning of phrases and sentences is at the heart of every task related to natural language.

- **Paraphrase detection**

  Problem: Given two sentences, decide if they say the same thing in different words

  Solution: Measure the cosine similarity between the sentence vectors

- **Sentiment analysis**

  Problem: Extract the general sentiment from a sentence or a document

  Solution: Train a classifier using sentence vectors as input

Applications (2/2)

- **Textual entailment**

  Problem: Decide if one sentence logically infers a different one

  Solution: Examine the feature inclusion properties of the sentence vectors

- **Machine translation**

  Problem: Automatically translate one sentence into a different language

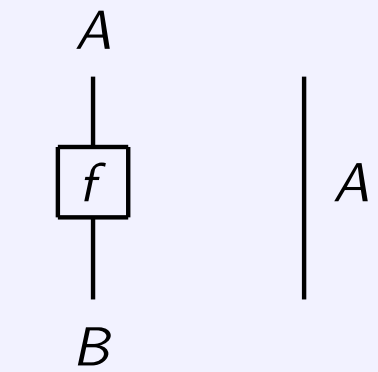
  Solution: Encode the source sentence into a vector, then use this vector to decode a surface form into the target language

- *And so on. Many other potential applications exist...*

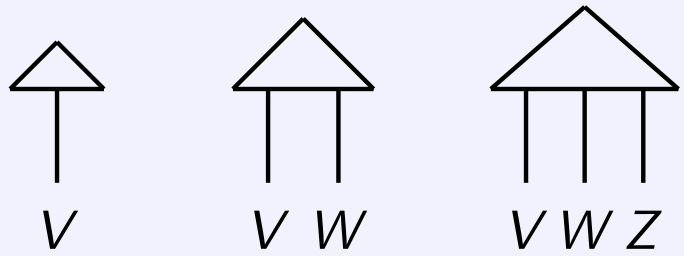
A general programme

1. a Choose a compositional structure, such as a pregroup or combinatory categorial grammar.
   b Interpret this structure as a category, the **grammar category**.
2. a Choose or craft appropriate meaning or concept spaces, such as vector spaces, density matrices, or conceptual spaces.
   b Organize these spaces into a category, the **semantics category**, with the same abstract structure as the grammar category.
3. Interpret the compositional structure of the grammar category in the semantics category via a functor preserving the necessary structure.
4. Bingo! This functor maps type reductions in the grammar category onto algorithms for composing meanings in the semantics category.

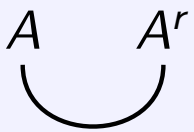
Diagrammatic calculus: Summary



morphisms

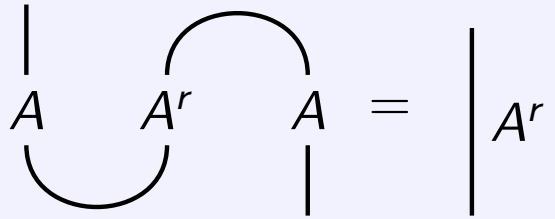


tensors



ϵ-map

η-map



$(\epsilon^r_A \otimes 1_A) \circ (1_A \otimes \eta^r_A) = 1_A$

Quantizing the grammar

Coecke, Sadrzadeh and Clark (2010):

Pregroup grammars are structurally homomorphic with the category of finite-dimensional vector spaces and linear maps (both share compact closure)

- In abstract terms, there exists a structure-preserving passage from grammar to meaning:

$$\mathcal{F} : \text{Grammar} \rightarrow \text{Meaning}$$

- The meaning of a sentence $w_1 w_2 \dots w_n$ with grammatical derivation α is defined as:

$$\overrightarrow{w_1 w_2 \dots w_n} := \mathcal{F}(\alpha)(\overrightarrow{w_1} \otimes \overrightarrow{w_2} \otimes \dots \otimes \overrightarrow{w_n})$$

Pregroup grammars

A pregroup grammar $P(\Sigma, \mathcal{B})$ is a relation that assigns grammatical types from a Compact CC freely generated over a set of atomic types $\mathcal{B}$ to words of a vocabulary Σ.

- Atomic types $x \in \mathcal{B}$ have morphisms

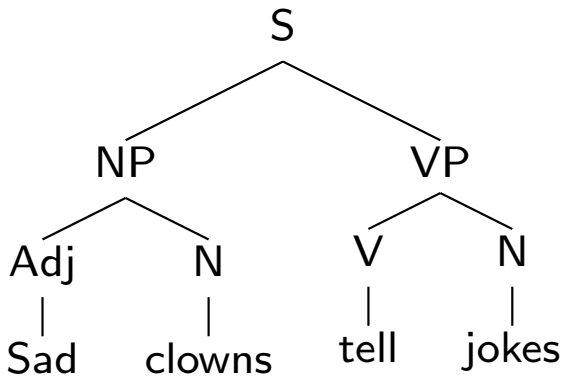
$$\epsilon^r_x : x \cdot x^r \to 1, \qquad \epsilon^l_x : x^l \cdot x \to 1$$
$$\eta^r_x : 1 \to x^r \cdot x, \qquad \eta^l_x : 1 \to x \cdot x^l$$

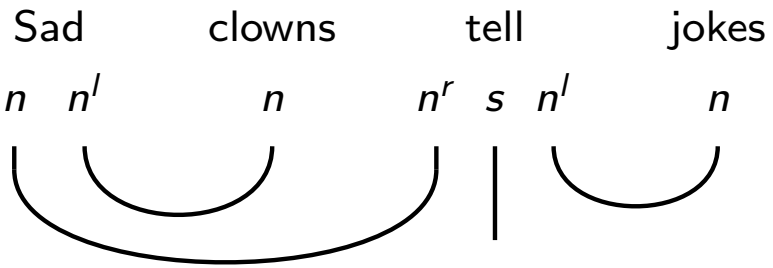
- Elements of the pregroup are basic (atomic) grammatical types, e.g. $\mathcal{B} = \{n, s\}$.
- Atomic grammatical types can be combined to form types of higher order (e.g. $n \cdot n^l$ or $n^r \cdot s \cdot n^l$)
- A sentence $w_1 w_2 \dots w_n$ (with word w_i to be of type t_i) is grammatical whenever:

$$t_1 \cdot t_2 \cdot \dots \cdot t_n \to s$$

Pregroup derivation: example

$$p \cdot p^r \to 1 \to p^r \cdot p \qquad p^l \cdot p \to 1 \to p \cdot p^l$$





$$\begin{aligned} n \cdot n^l \cdot n \cdot n^r \cdot s \cdot n^l \cdot n &\to n \cdot 1 \cdot n^r \cdot s \cdot 1 \\ &= n \cdot n^r \cdot s \\ &\to 1 \cdot s \\ &= s \end{aligned}$$

A functor from syntax to semantics

We define a strongly monoidal functor $\mathcal{F}$ such that:

$$\mathcal{F} : P(\Sigma, \mathcal{B}) \to \textbf{FVect}$$

$$\begin{aligned}
\mathcal{F}(p) &= P \quad \forall p \in \mathcal{B} \\
\mathcal{F}(1) &= \mathbb{R} \\
\mathcal{F}(p \cdot q) &= \mathcal{F}(p) \otimes \mathcal{F}(q) \\
\mathcal{F}(p^r) = \mathcal{F}(p^l) &= \mathcal{F}(p) \\
\mathcal{F}(p \leq q) &= \mathcal{F}(p) \to \mathcal{F}(q) \\
\mathcal{F}(\epsilon^r) = \mathcal{F}(\epsilon^l) &= \text{inner product in } \textbf{FVect} \\
\mathcal{F}(\eta^r) = \mathcal{F}(\eta^l) &= \text{identity maps in } \textbf{FVect}
\end{aligned}$$

[Kartsaklis, Sadrzadeh, Pulman and Coecke, 2016]

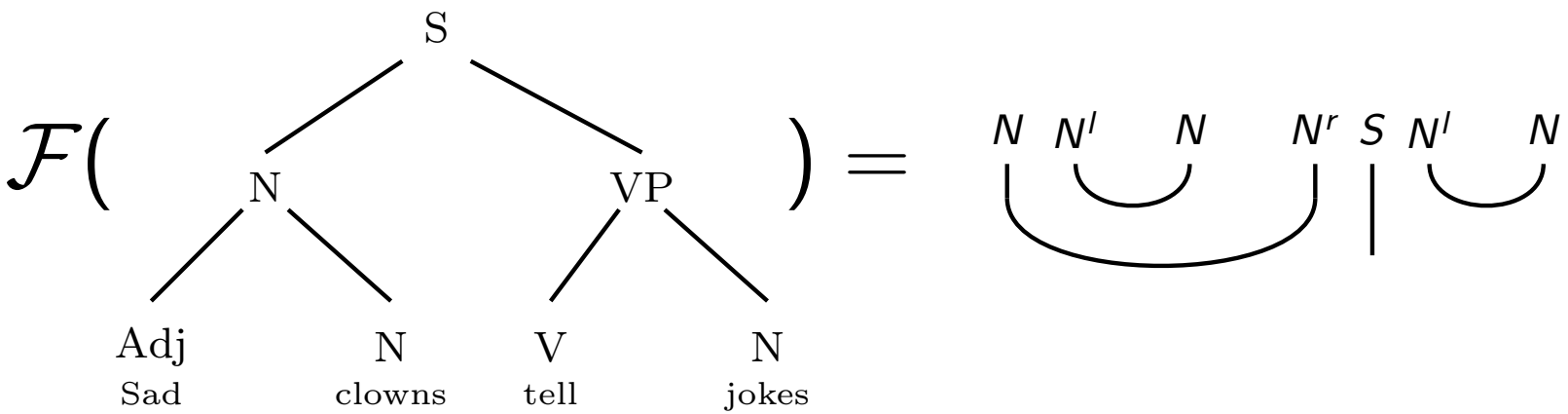
A multi-linear model

The grammatical type of a word defines the vector space in which the word lives:

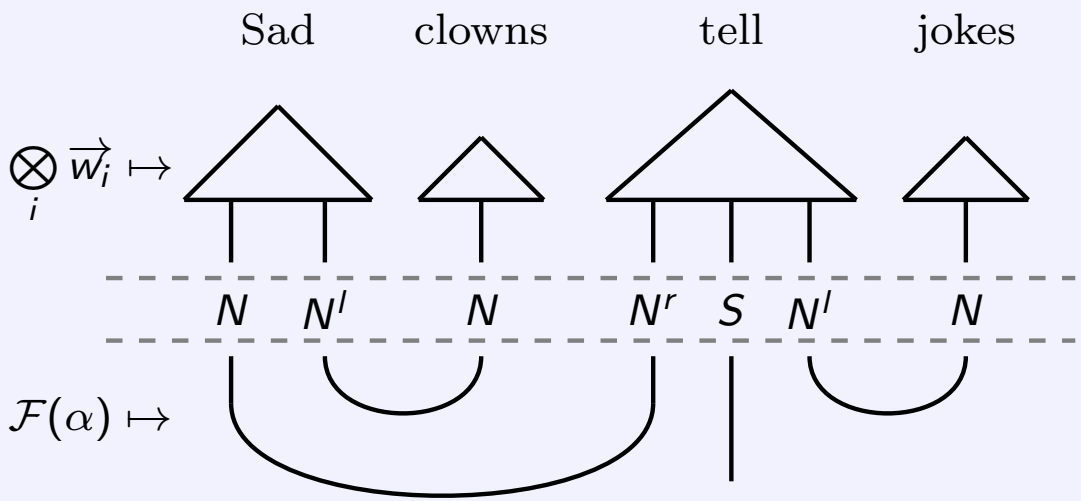
- Nouns are vectors in N;
- adjectives are linear maps $N \to N$, i.e elements in $N \otimes N$;
- intransitive verbs are linear maps $N \to S$, i.e. elements in $N \otimes S$;
- transitive verbs are bi-linear maps $N \otimes N \to S$, i.e. elements of $N \otimes S \otimes N$;

- The composition operation is tensor contraction, i.e. elimination of matching dimensions by application of inner product.

Graphical calculus: example



$\mathcal{F}(\alpha)(\overrightarrow{trembling} \otimes \overrightarrow{shadows} \otimes \overrightarrow{play} \otimes \overrightarrow{hide\text{-}and\text{-}seek})$



Summary

- Formal semantic approaches are good at composition...
- ... but the things they compose are featureless atoms
- Distributional semantics give a much richer meaning to their atoms...
- ... but have no obvious compositional mechanisms.
- The categorical compositional model marries the two by interpreting both the grammar and the sematics category as being compact closed.
- The structure and interactions of the grammar category are mapped over to the semantic category.

Outline

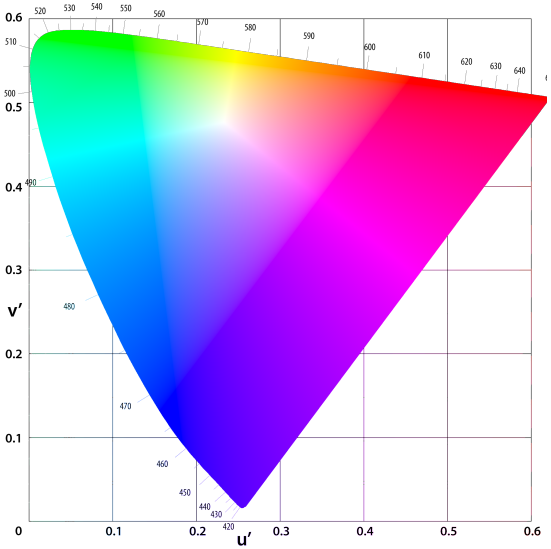
A general programme

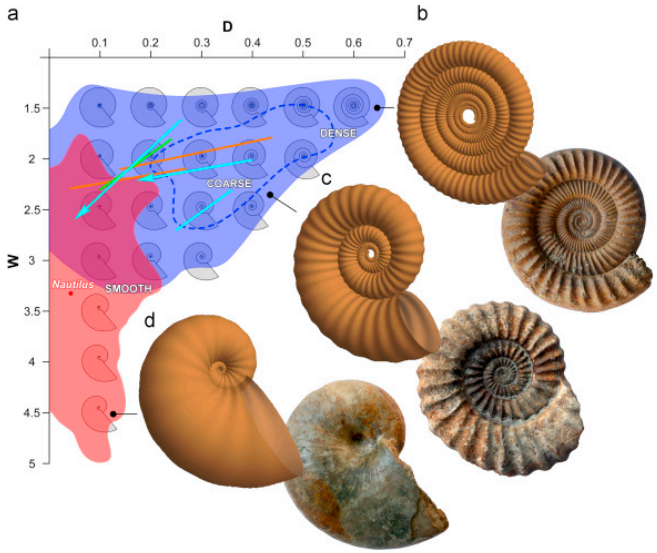
1. a Choose a compositional structure, such as a pregroup or combinatory categorial grammar.
   b Interpret this structure as a category, the **grammar category**.
2. a Choose or craft appropriate meaning or concept spaces, such as vector spaces, density matrices, or conceptual spaces.
   b Organize these spaces into a category, the **semantics category**, with the same abstract structure as the grammar category.
3. Interpret the compositional structure of the grammar category in the semantics category via a functor preserving the necessary structure.
4. Bingo! This functor maps type reductions in the grammar category onto algorithms for composing meanings in the semantics category.

Conceptual spaces

Conceptual spaces [Gärdenfors, 2014] can provide a more cognitively realistic semantics.

$$noun \in COLOUR \otimes SHAPE \otimes \cdots$$





[Moulton et al., 2015]

Convex algebras

- Notation. For a set X we write $\sum_i p_i \, |x_i\rangle$ for a finite formal convex sum of elements of X, where $p_i \in \mathbb{R}^{\geq 0}$ and $\sum_i p_i = 1$.
- A **convex algebra** is a set A with a **mixing operation** α satisfying:

$$\alpha(|a\rangle) = a$$

$$\alpha(\sum_{i,j} p_i q_{i,j} \, |a_{i,j}\rangle) = \alpha(\sum_i p_i \, |\alpha(\sum_j q_{i,j} \, |a_{i,j}\rangle)\rangle)$$

- Examples: Real or complex vector spaces, simplices, semilattices, trees
- A **convex relation** between convex algebras (A, α) and (B, β) is a relation that commutes with forming convex combinations, i.e.

$$(\forall i. R(a_i, b_i)) \Rightarrow R(\sum_i p_i a_i, \sum_i p_i b_i)$$

A category for conceptual spaces

- We define the category **ConvexRel** as having convex algebras as objects and convex relations as morphisms
- **ConvexRel** is compact closed with $\times$ as monoidal product.
- We build a functor from pregroup grammar in the same way: choose a space N for nouns, a space S for sentences.
- Drawback: how do we start to build word representations?

Change of category - **CPM(FVect)**

- We can use density matrices, and more generally, positive operators rather than vectors to represent words.
- Positive operators A, B have the Löwner ordering $A \sqsubseteq B \iff B - A$ is positive.
- We use this ordering to represent entailment, and introduce a graded version - useful for linguistic phenomena.
- We will show that graded entailment lifts compositionally to sentence level.
- Similar approaches in [Sadrzadeh et al., 2018]

Words as positive operators

- A positive operator P is self-adjoint and $\forall v \in V. \langle v| P |v\rangle \geq 0$
- Density matrices are convex combinations of projectors: $\rho = \sum_i p_i |v_i\rangle \langle v_i|$, s.t. $\sum_i p_i = 1$
- We can view concepts as collections of items, with p_i indicating relative frequency.
- For example:

$$\begin{aligned}[[pet]] =& p_d |dog\rangle \langle dog| + p_c |cat\rangle \langle cat| + \\ & p_t |tarantula\rangle \langle tarantula| + ... \end{aligned}$$

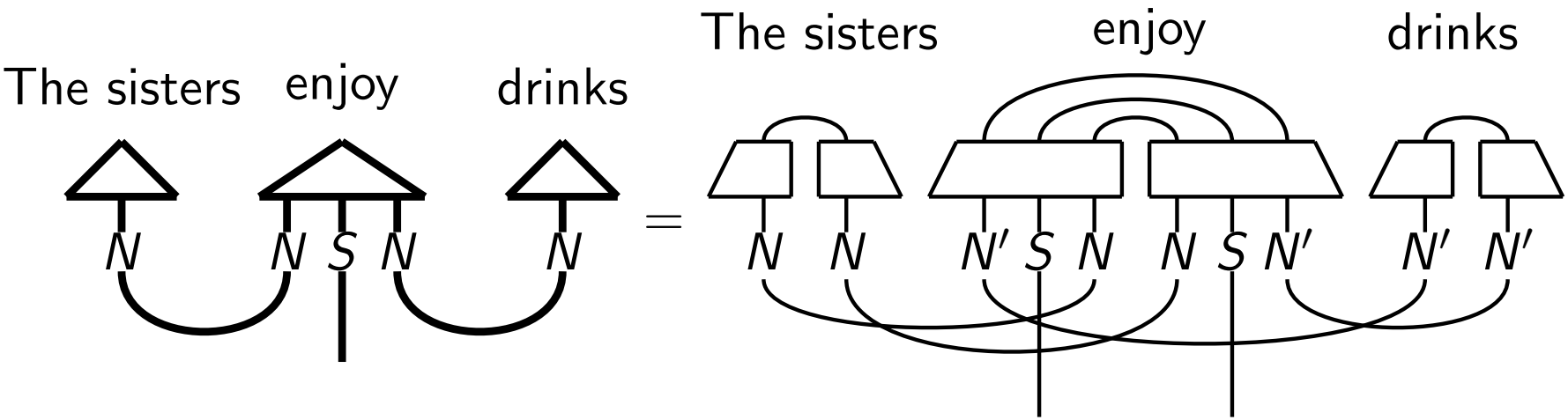
$$\text{where} \quad \forall i. p_i \geq 0 \quad \text{and} \quad \sum_i p_i = 1$$

- There are various choices for normalisation of the density matrices...

Sentence meaning in **CPM(FVect)**

- We assign semantics via a strong monoidal functor $S : \mathbf{Preg} \to \mathbf{CPM}(\mathbf{FVect})$
- Let $w_1 w_2 ... w_n$ be a string of words with corresponding grammatical types t_i in $\mathbf{Preg}_{\{n,s\}}$ such that $t_1, ... t_n \xrightarrow{r} s$
- Let $[\![w_i]\!]$ be the meaning of word w_i in **CPM(FVect)**. Then the meaning of $w_1 w_2 ... w_n$ is given by:

$$[\![w_1 w_2 ... w_n]\!] = S(r)([\![w_1]\!] \otimes ... \otimes [\![w_n]\!])$$



So how do we do graded hyponymy?

- Recall that positive operators A, B have the Löwner ordering $A \sqsubseteq B \iff B - A$ is positive.
- We say that A is a *hyponym* of B if $A \sqsubseteq B$
- We say that A is a *k-hyponym* of B for a given value of k in the range $(0, 1]$ and write $A \preccurlyeq_k B$ if:

$$B - kA \text{ is positive}$$

- We are interested in the maximum such k.

Theorem

For positive self-adjoint matrices A, B such that $supp(A) \subseteq supp(B)$, the maximum k such that $B - kA \geq 0$ is given by $1/\lambda$ where λ is the maximum eigenvalue of $B^{+}A$.

k-hyponymy interacts well with compositionality

- We would like our notion of entailment to work at the sentence level.
- Since sentences are represented as positive operators, we can compare them directly.
- If sentences have similar structure, we can also give a lower bound on the entailment strength between sentences based on the entailment strengths between the words in the sentences.

Example

Suppose $[\![dog]\!] \preccurlyeq_k [\![pet]\!]$ and $[\![park]\!] \preccurlyeq_l [\![field]\!]$. Then

$$[\![\textit{My dog runs in the park}]\!] \preccurlyeq_{???} [\![\textit{My pet runs in the field}]\!]$$

How should we build density matrices for words?

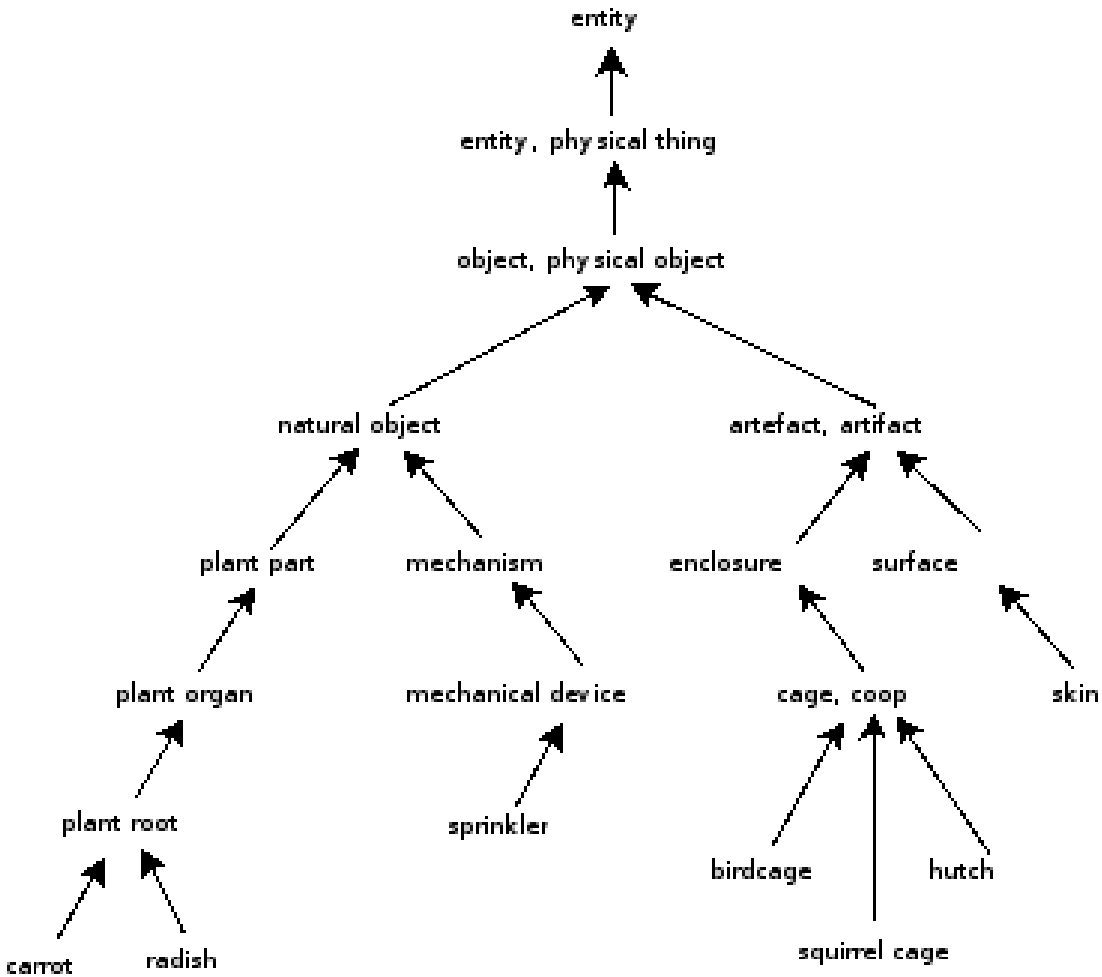


Figure 1. "is a" relation example

WordNet. Princeton University. 2010

Experiments

Hyperlex [Vulić et al., 2017] gold-standard dataset. 2,163 noun pairs human annotated.

Model	Dev	Test
Poincaré embeddings	-	0.512
SDSN (Random)	0.757	0.692
SDSN (Lexical)	0.577	0.544
Density matrices	-	0.551
Density matrices (non)	-	0.631

Experiments

Dataset from [Sadrzadeh et al., 2018] consisting of 23 sentence pairs. Example pairs are:

recommend development $\models$ *suggest improvement*
progress reduce $\models$ *development replace*

With normalization:

Model	ρ	p
Verb-only	0.268	> 0.25
Frobenius mult.	0.508	> 0.05
Frobenius n.c.	0.436	> 0.05
Additive	0.643	> 0.001
Inter-annotator	0.66	-

[Bankova et al., 2019]

Experiments

Dataset from [Sadrzadeh et al., 2018]. Example pairs are:

recommend development $\models$ *suggest improvement*
progress reduce $\models$ *development replace*

Without normalization:

Model	ρ	p
Verb-only	0.370	> 0.1
Frobenius mult.	0.696	$> 5 \cdot 10^{-4}$
Frobenius n.c.	0.306	0.15
Additive	0.737	$> 5 \cdot 10^{-5}$
Inter-annotator	0.66	-

Speculation

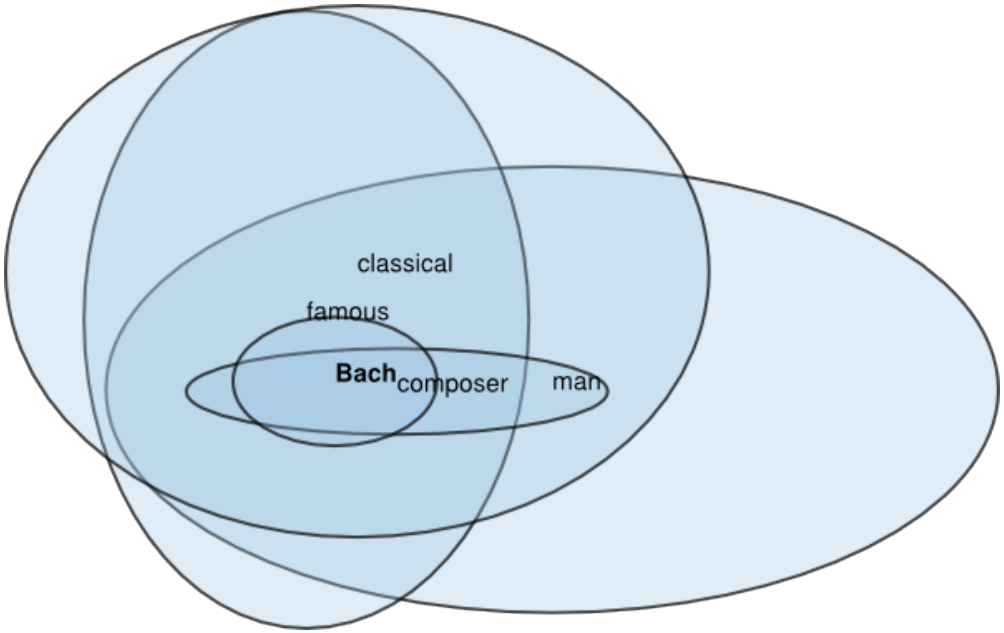


Figure 1: Learned diagonal variances, as used in evaluation (Section 6), for each word, with the first letter of each word indicating the position of its mean. We project onto generalized eigenvectors between the mixture means and variance of query word *Bach*. Nearby words to *Bach* are other composers e.g. *Mozart*, which lead to similar pictures.

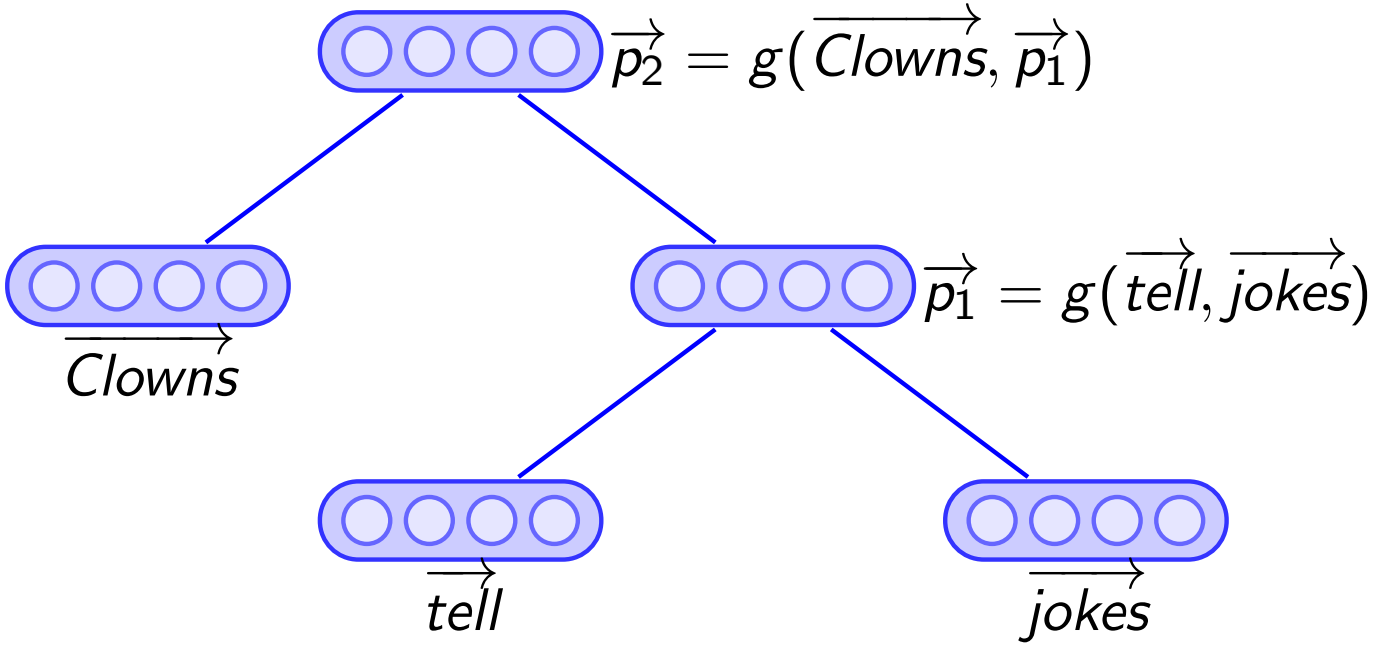
[Vilnis and McCallum, 2014], and similar approaches seen in [Jameel and Schockaert, 2017], [Bražinskas et al., 2017]

Summary

- Vector spaces provide a good means of talking about similarity between words.
- But both conceptual spaces and positive operators give us a richer word representation.
- We don't have a good way of building word representations in conceptual spaces yet, but using information from WordNet to build density matrices seems to give good results with very little effort.

Outline

Recursive Neural Networks



$$g_{RNN} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n :: (\overrightarrow{v_1}, \overrightarrow{v_2}) \mapsto f_1\left(M \cdot \begin{bmatrix} \overrightarrow{v_1} \\ \overrightarrow{v_2} \end{bmatrix}\right)$$

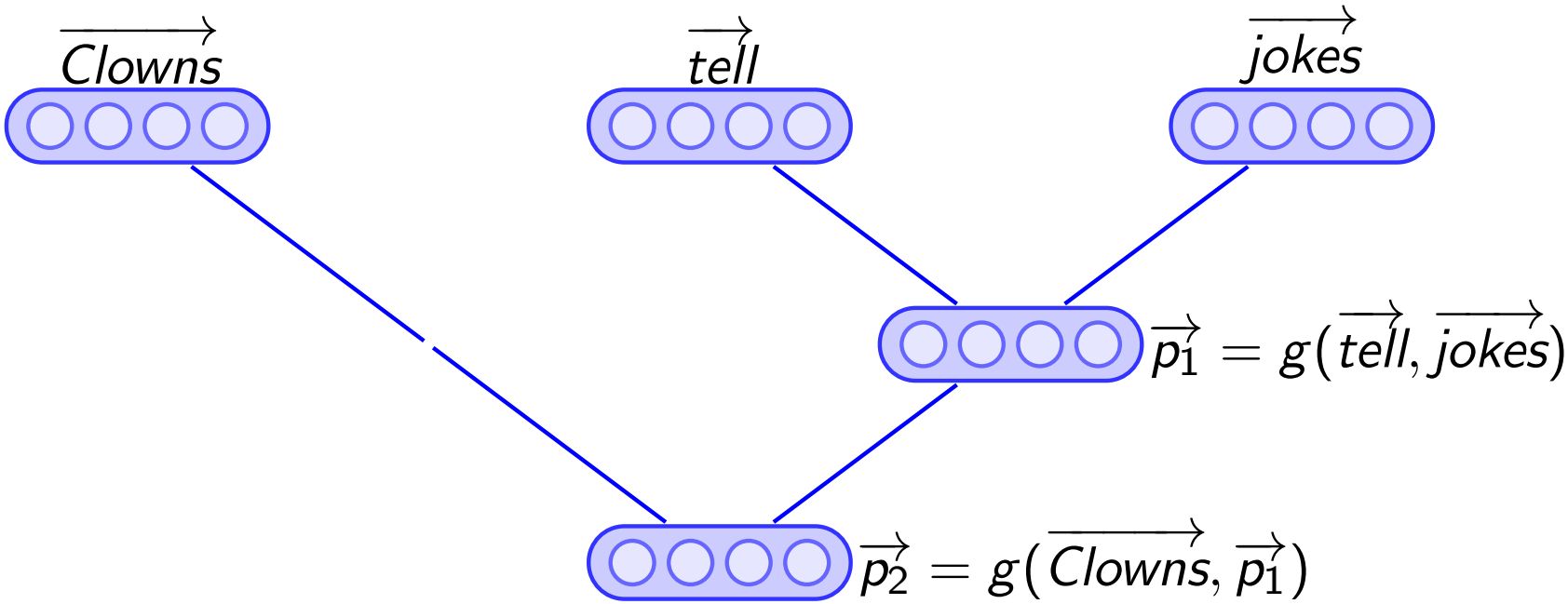
$$g_{RNTN} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n :: (\overrightarrow{v_1}, \overrightarrow{v_2}) \mapsto g_{RNN}(\overrightarrow{v_1}, \overrightarrow{v_2}) + f_2\left(\overrightarrow{v_1}^\top \cdot T \cdot \overrightarrow{v_2}\right)$$

[Socher et al., 2012, Bowman et al., 2014]

A (compact closed) category for neural networks

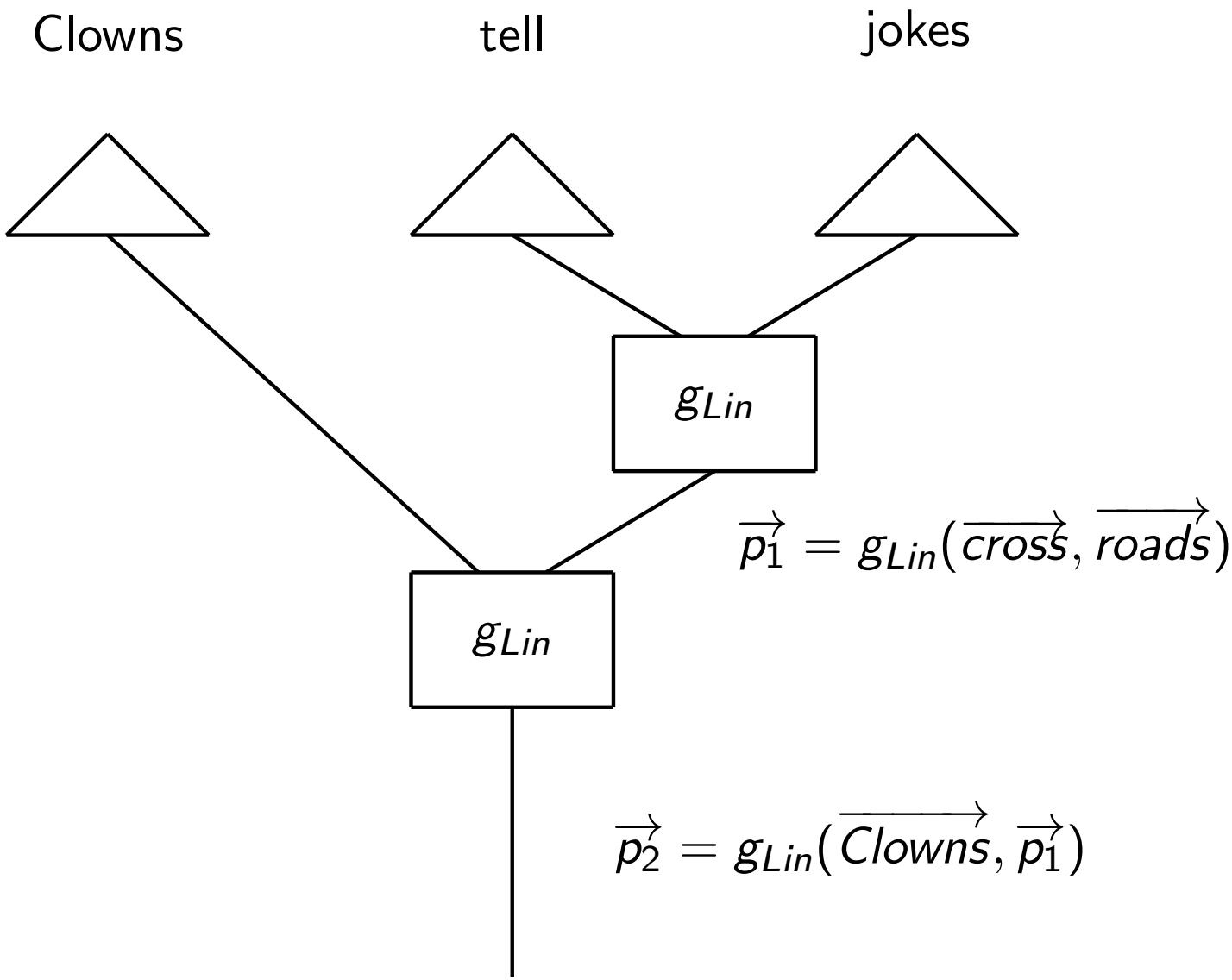
???

Alternatively... linear recursive neural networks

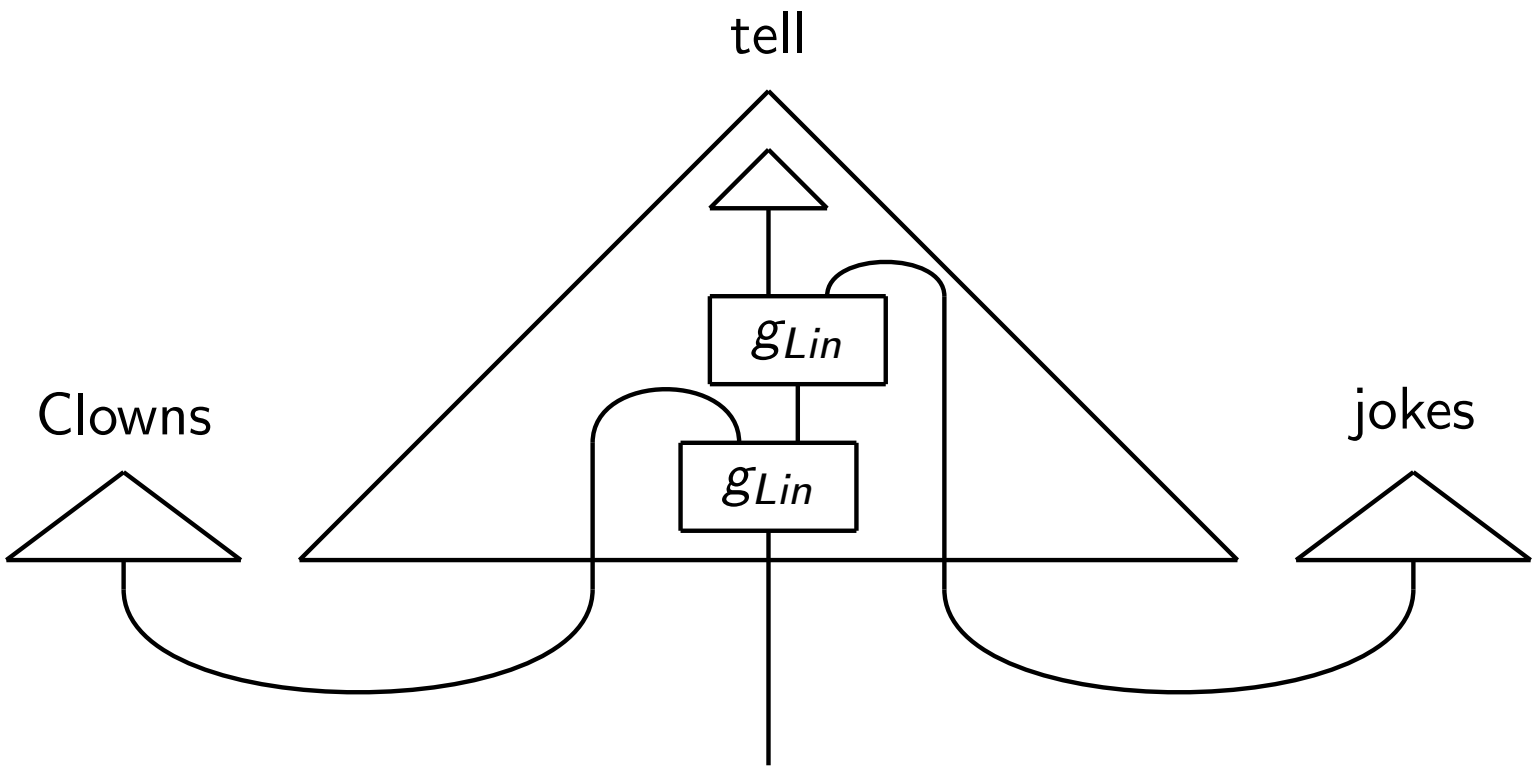


[Lewis, 2019]

Alternatively... linear recursive neural networks

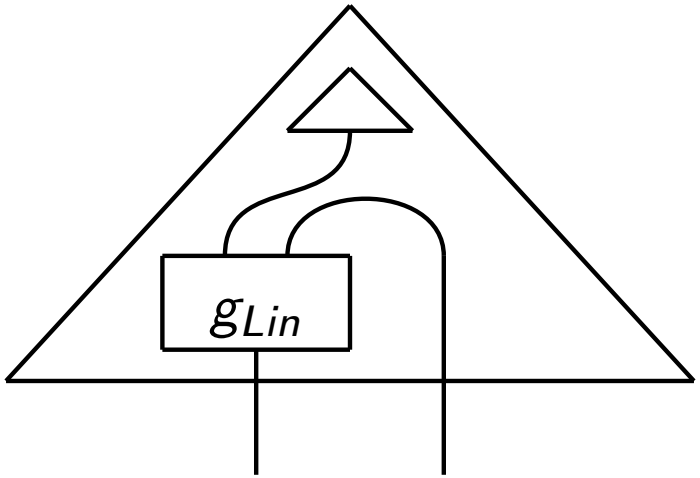


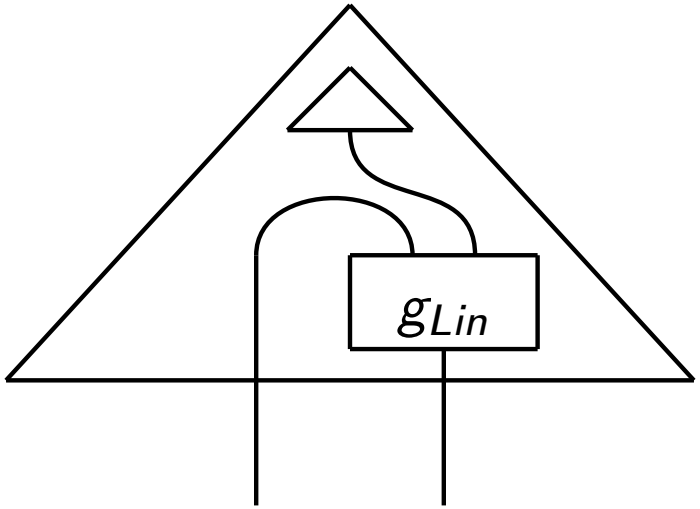
Alternatively... linear recursive neural networks



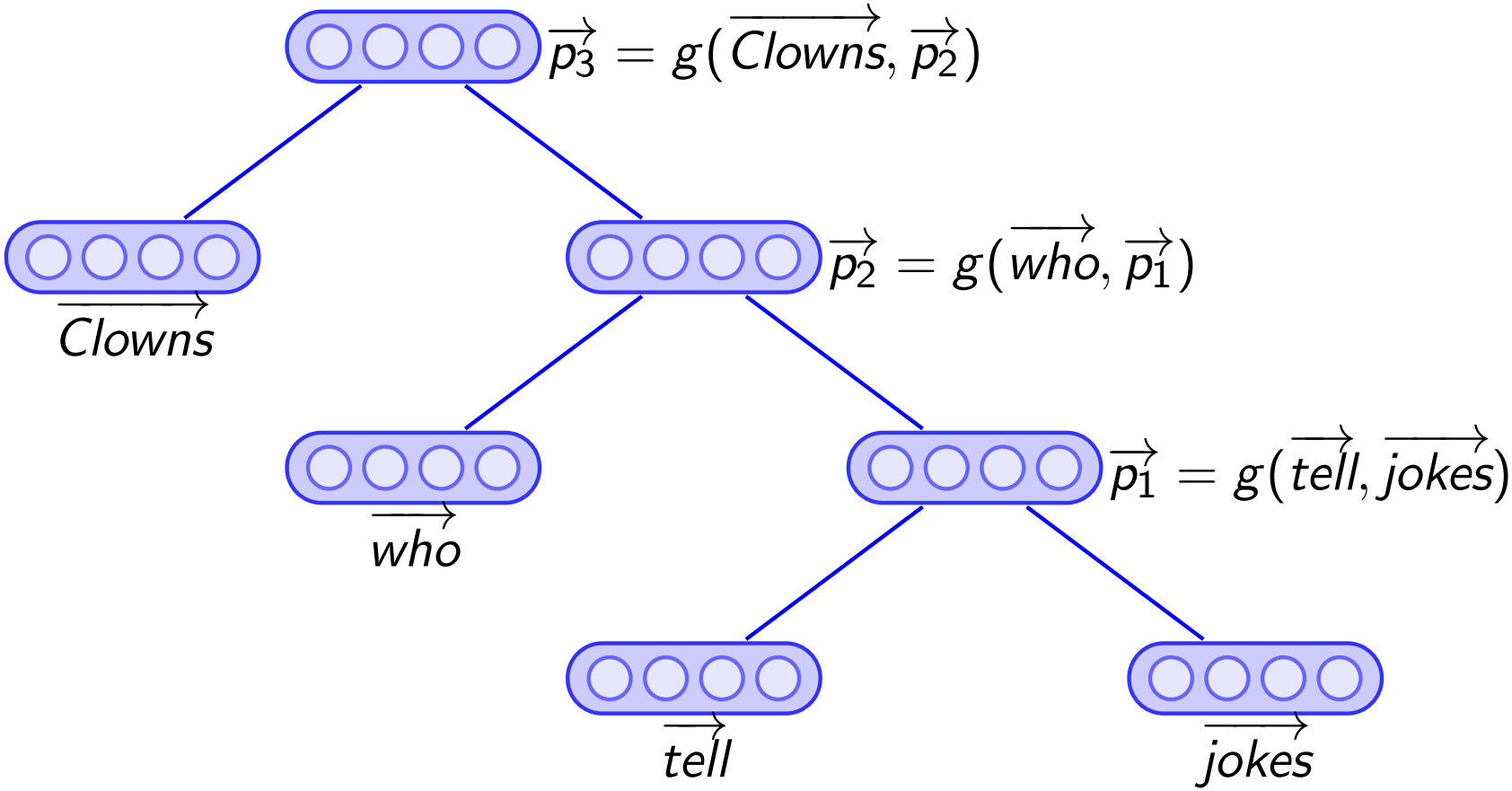
Content/structure split

Adjectives and intransitive verbs



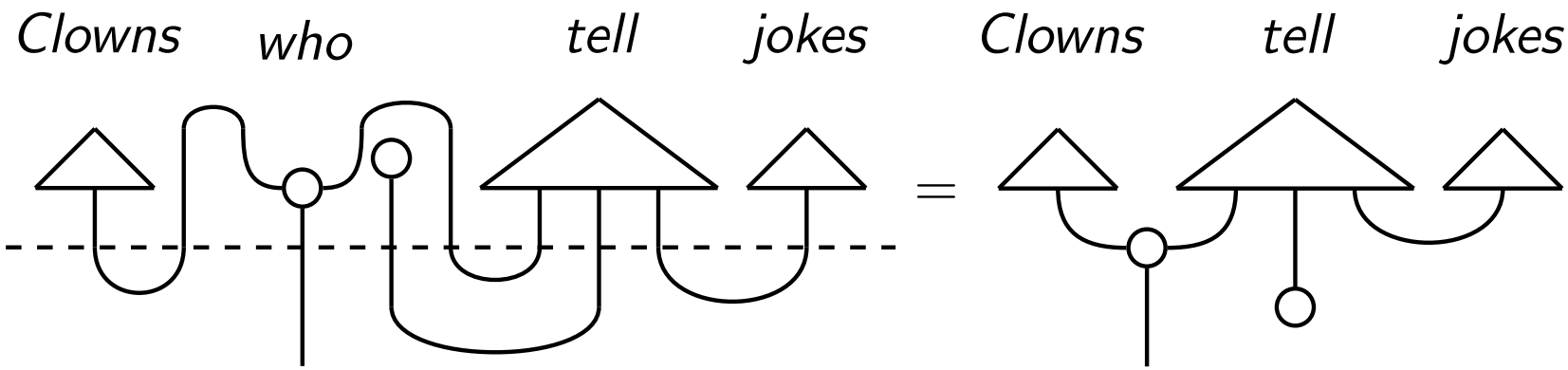


Content/structure split



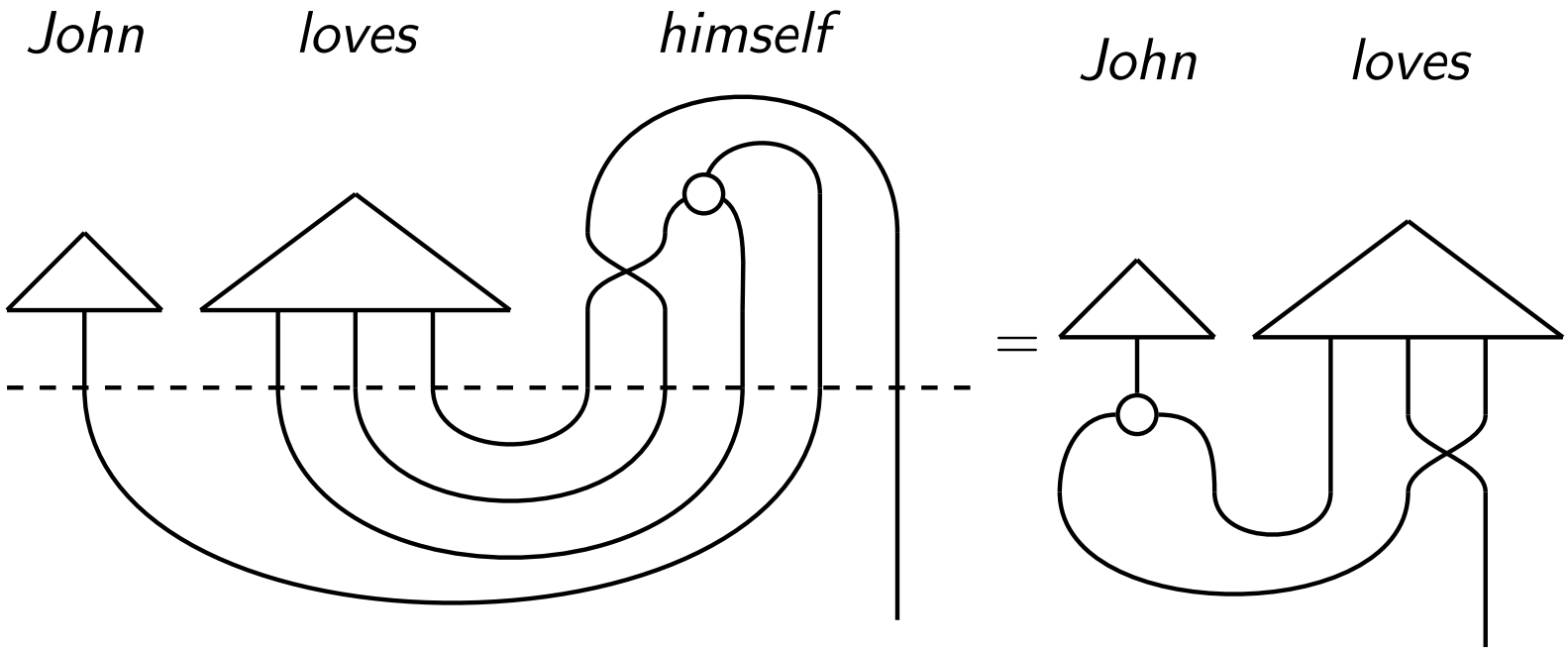
Content/structure split

$who : n^r\, ns^l\, s$



Content/structure split

$himself : ns^r n^{rr} n^r s$



Benefits

- Understanding neural networks as a semantics category for Lambek categorial grammar opens up more possibilities to use tools from formal semantics in computational linguistics.
- We can immediately see possibilities for building alternative networks.
- Decomposing the tensors for functional words into repeated applications of a compositionality function gives options for learning representations.
- Brittleness of word types in DisCo is alleviated

To do

- Incorporate non-linearity
- Extend to other types of network that are currently state-of-the-art
- Testing on data - comparison with standard DisCo representations, examine ability to switch word types, look at specialised tensors

Outline

Summary

- The categorical compositional model of [Coecke et al., 2010] can be instantiated with a choice of grammar category (which we saw yesterday in Sadrzadeh's and Wijnhold's talks).
- We also have a choice of meaning category, allowing us to move towards a richer semantics. Initial results are positive.
- We have started to link to neural network methods for building word vectors.

Outlook

- Develop ways of building word regions from corpora
- Links between positive matrices and multivariate Gaussian word embeddings?
- Building on more cognitively plausible concept representations.
- Starting to model meaning change and game-theoretic models of language [Bradley et al., 2018, Hedges and Lewis, 2018].
- Modelling non-literal uses of language.

Thanks!

NWO Veni grant ‘Metaphorical Meanings for Artificial Agents’

References I

Bankova, D., Coecke, B., Lewis, M., and Marsden, D. (2019).
Graded entailment for compositional distributional semantics.
arXiv preprint arXiv:1601.04908.
to appear in Journal of Language Modelling.

Bowman, S. R., Potts, C., and Manning, C. D. (2014).
Recursive neural networks can learn logical semantics.
arXiv preprint arXiv:1406.1827.

Bradley, T.-D., Lewis, M., Master, J., and Theilman, B. (2018).
Translating and evolving: Towards a model of language change in discocat.
arXiv preprint arXiv:1811.11041.

Bražinskas, A., Havrylov, S., and Titov, I. (2017).
Embedding words as distributions with a bayesian skip-gram model.
arXiv preprint arXiv:1711.11027.

Coecke, B., Sadrzadeh, M., and Clark, S. (2010).
Mathematical Foundations for a Compositional Distributional Model of Meaning. Lambek Festschrift.
Linguistic Analysis, 36:345–384.

Gärdenfors, P. (2014).
The geometry of meaning: Semantics based on conceptual spaces.
MIT Press.

Hedges, J. and Lewis, M. (2018).
Towards functorial language-games.
arXiv preprint arXiv:1807.07828.

References II

Jameel, S. and Schockaert, S. (2017).

Modeling context words as regions: An ordinal regression approach to word embedding.
In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 123–133.

Lewis, M. (2019).

Compositionality for recursive neural networks.
IfCoLog Journal of Applied Logics.
to appear.

Moulton, D., Goriely, A., and Chirat, R. (2015).

The morpho-mechanical basis of ammonite form.
Journal of Theoretical Biology, 364:220–230.

Sadrzadeh, M., Kartsaklis, D., and Balkır, E. (2018).

Sentence entailment in compositional distributional semantics.
Annals of Mathematics and Artificial Intelligence, 82(4):189–218.

Socher, R., Huval, B., Manning, C., and A., N. (2012).

Semantic compositionality through recursive matrix-vector spaces.
In *Conference on Empirical Methods in Natural Language Processing 2012*.

Vilnis, L. and McCallum, A. (2014).

Word representations via gaussian embedding.
arXiv preprint arXiv:1412.6623.

Vulić, I., Gerz, D., Kiela, D., Hill, F., and Korhonen, A. (2017).

Hyperlex: A large-scale evaluation of graded lexical entailment.
Computational Linguistics, 43(4):781–835.