LANGUAGE COMPREHENSION AND INTEGRATION

AIS 2023

TAKING STOCK

1. We have seen that people easily learn associations, and can use them for prediction about the future
2. People integrate information from the recent past to make predictions
3. We also know that human memory systems are optimized for storing 7+- 3 items, and that chunking is one of the main tools used to construct these items
4. All this suggests that our ability to use learned associations, keep track of the recent past, and make predictions would be fundamental for language comprehension as well.

1. The core question is: do we store language associations at all?
2. And if so, how do we use them for comprehension. These are two main questions underlying modern and less modern theories of language.

HOW DO WE LEARN ‘THE RULES’ OF LANGUAGE?

1. From experience, we memorize structures: which can be words or types of phrases that go together
2. We learn abstract rules that specify constraints on grammaticality
3. We acquire statistical knowledge on co-occurrence of language entities.

BEHAVIORAL PSYCHOLOGY: MEMORIZATION OF STRUCTURES.

Behavioral psychology research applied S-R models for language: Each word is thought of being a stimulus for the next, building up an overall structure out of local associative relations.

In language comprehension: the learned sequences of adjacent elements are internally represented as automatically characterizing a sentence as it is encountered. Allows to determine if it is grammatical.

We would say (today) that on behaviorism, knowledge of language is knowing which words/elements are followed by other words.

It doesn't explain how 'meaning' is made, but it explains how people produce valid sentences to communicate an idea: they know which words/elements can come after others.

ELEMENTS IN S-R APPROACHES TO LANGUAGE

S-R approaches claimed we learn which words come after each other

Horse -> *races* (relation between specific lexical items)

But they also allowed for some abstraction, which allowed for ‘slots’. E.g., patterns such as

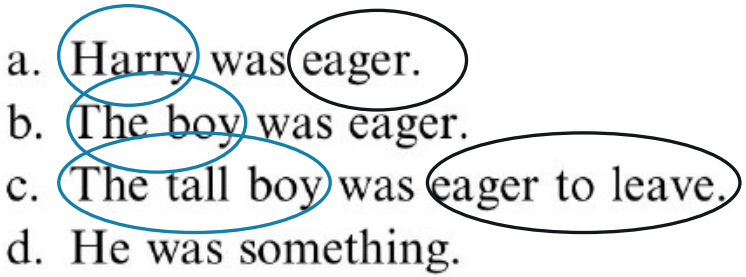
1. The -> *X* -> *Y*-es (relation between def article, noun [sing] verb [sing]
2. The -> Xes – Y_ (relation between def, plural noun, plural verb)

ABSTRACT ELEMENTS IN SENTENCE LEVEL

At certain point, Behaviorists accepted some level of abstraction.

Attempted to describe different types of phrases.

They accepted that longer phrases seemed to resolve into combinations of shorter ones, which in turn could resolve into single words.



MEANING IN S-R APPROACHES TO LANGUAGE

Behaviorists did not develop theories about mental phenomenon (e.g., theories of conceptual representation). They had several parallel ideas:

Substitution view: Words ('signs') substituted for the objects in the world. Objects in the world produce a real reaction, and the sign produces a similar reaction

Disposition view: Language produces a tendency (disposition) towards certain behaviors 'The meaning of a linguistic expression is therefore 'disposition towards response sequence' (Osgood, 1952)

*Other views: tried to explain how we understand words that refer to objects we have not directly experienced. Words associated with direct experience: Sign. Indirect experience: Assign. These obtain meanings by 'compositional effects' of lower level features (zebra = horse + stripes). Beginning of cognitive approach.

CHOMSKY'S CHALLENGES TO S-R LEARNING AND BEHAVIORISM: ABSTRACT RULES

Chomsky (1959) identified multiple weaknesses, including

1. inability to explain comprehension of sentences not heard previously,
2. Inability to handle hierarchical structure and long-distance dependencies

Example : The horse[s].....[long clause] is/are falling.

The ability of an algorithm to determine whether a new, never-hear-before expression sensible remains a strong benchmark for the validity of an NLP algorithm.

EVIDENCE FOR RULE (GRAMMAR) -DRIVEN COMPOSITIONALITY

When put (strung) together words in noise are perceived more clearly than when presented randomly (“Horses”, “eat”) : 25% accurate per word, but 50% when put together. “Acoustic representation” not mapped independently per word.

This opened the door to the question of *how does knowledge of sentences impact comprehension*, where ‘top-down’ explanation have been battling bottom-up ones for over 60 years.

And why are things clearer in sentences? They cannot be stored in a mental lexicon as words are, so how do they impact comprehension?

- The answer: The constraints are not from rote memory but from more abstract systems of rules:

So adults develop representations that cannot directly derived from prior experience.

CHOMSKY AND TRANSFORMATIONAL GRAMMAR

What is a sentence?

Chomsky (1957): A sentence is what the grammar describes it as a sentence.

The grammar was generative: it described the sentence structures of a language as a natural and creative part of human knowledge.

Grammar constitutes a theory of the speaker’s underlying linguistic knowledge. It is much less (=not) concerned with the (noisy) actual behavior or its importance for the study of language

SENTENCE MEANING IS CAPTURED BY ITS STRUCTURE

The structure is a configuration that is allowed by the language.

Structures ‘licensed’ by the grammar of a language are ones that can be produced by the language’s grammar (rewrite rules)

S -> NP VP
NP -> Det N
VP -> V NP
VP -> V PP

1. John is easy to please
2. John is eager to please

In (1), John is (unstated) object of pleasing

In 2, he is the subject who pleases

Rephrasing:

It is easy to please john

** it is eager to please john

THE TRANSFORMATION FROM KERNEL SENTENCES IS KEY

SENTENCES WITH VERY SIMILAR SURFACE STRUCTURES MAY HAVE VERY DIFFERENT KERNELS

Chomsky hierarchy

The tree model works something like this example, in which:

S - sentence,

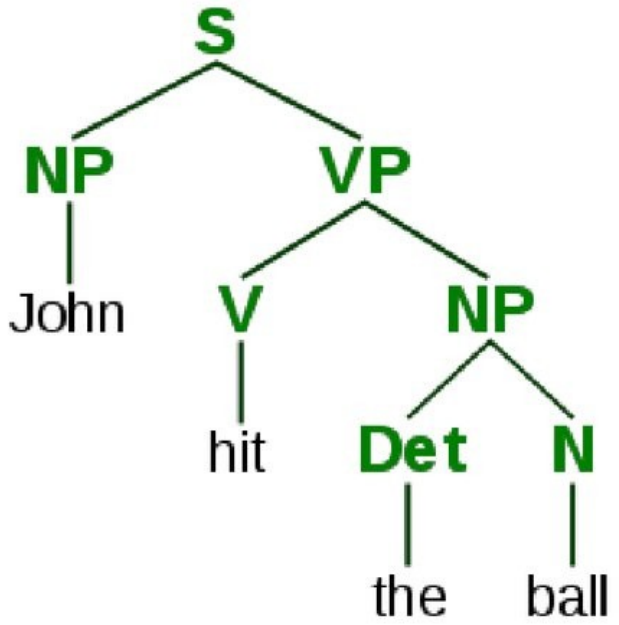
D - determiner,

N - noun,

V - verb,

NP - noun phrase,

VP - verb phrase.

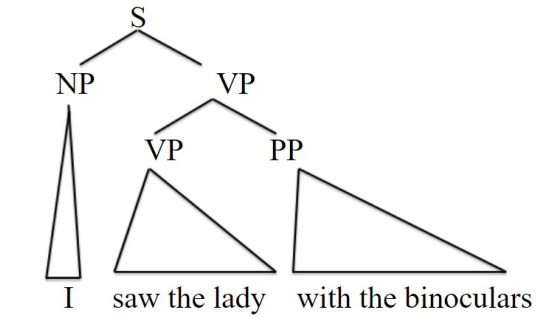


SENTENCE MEANING IS RELATED TO ITS STRUCTURE

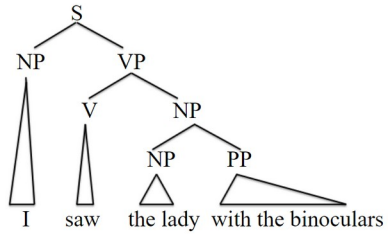
Structures are represented as trees.

SENTENCE MEANING IS RELATED TO ITS STRUCTURE

An ambiguous string of words means there are two possible meanings



I [saw the lady] with the binoculars



I saw [the lady with the binoculars]

DERIVATIONAL THEORY OF (PSYCHOLOGICAL) COMPLEXITY (I)

The first theory that tried to account for difficulty of sentence comprehension

Innovative in that it

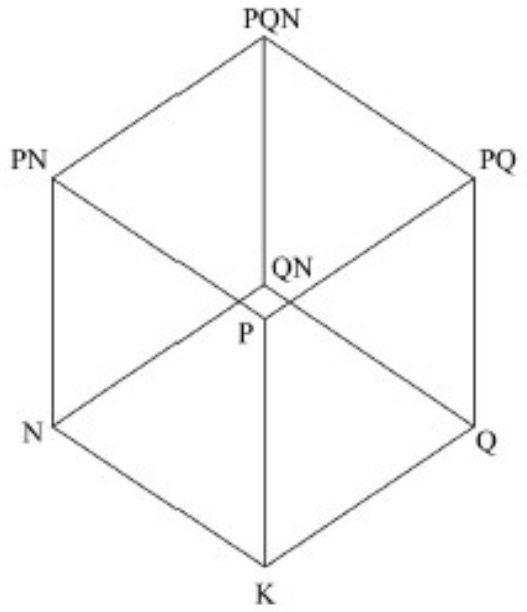
- Identified an objective metric on which complexity can vary
- Examined correlations between value of this metric and human behavior

The metric was the number of transformations needed to transform a kernel sentence to a given sentence.

The entire sentence was taken as a unit, and a set of pre-set rules (a grammar) was used to see if the given sentence can be derived from the kernel sentence. No incremental parsing

"Psychological validity" of the grammar as a mental model.

a. Mary hit Mark.	K
b. Mary did not hit Mark.	N
c. Mark was hit by Mary.	P
d. Did Mary hit Mark?	Q
e. Mark was not hit by Mary.	NP
f. Didn't Mary hit Mark?	NQ
g. Was Mark hit by Mary?	PQ
h. Wasn't Mark hit by Mary?	PNQ



PSYCHOLOGY AND TRANSFORMATION: PSYCHOLOGICAL VALIDITY? DERIVATIONAL THEORY OF COMPLEXITY

DERIVATIONAL THEORY OF (PSYCHOLOGICAL) COMPLEXITY (II)

DCT enjoyed some early success, particularly when using sentence - picture verification studies

Verification times
“The boy smelled the flower” <
“The flower was smelled by the boy”

DTC abandoned eventually because criticisms of method (construct validity) and findings suggesting these effects have to do with how the photo might be encoded into language.

PARSING BASED ACCOUNTS

Parsing is an incremental process by which we try assign a structure to what we have read/heard

In many cases the ambiguity is local and is disambiguated by later content.

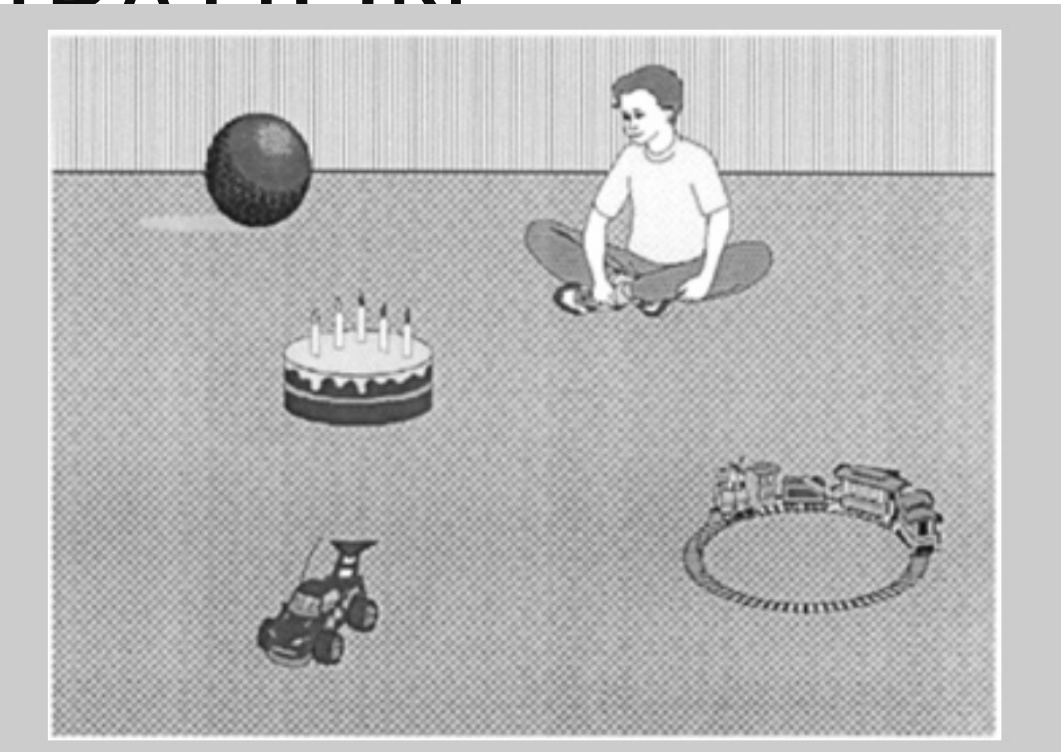
Parsing based accounts try to explain why some tentative interpretations are preferred and why some sentences are more difficult

Why is the following sentence difficult?

“The horse raced past the barn fell”

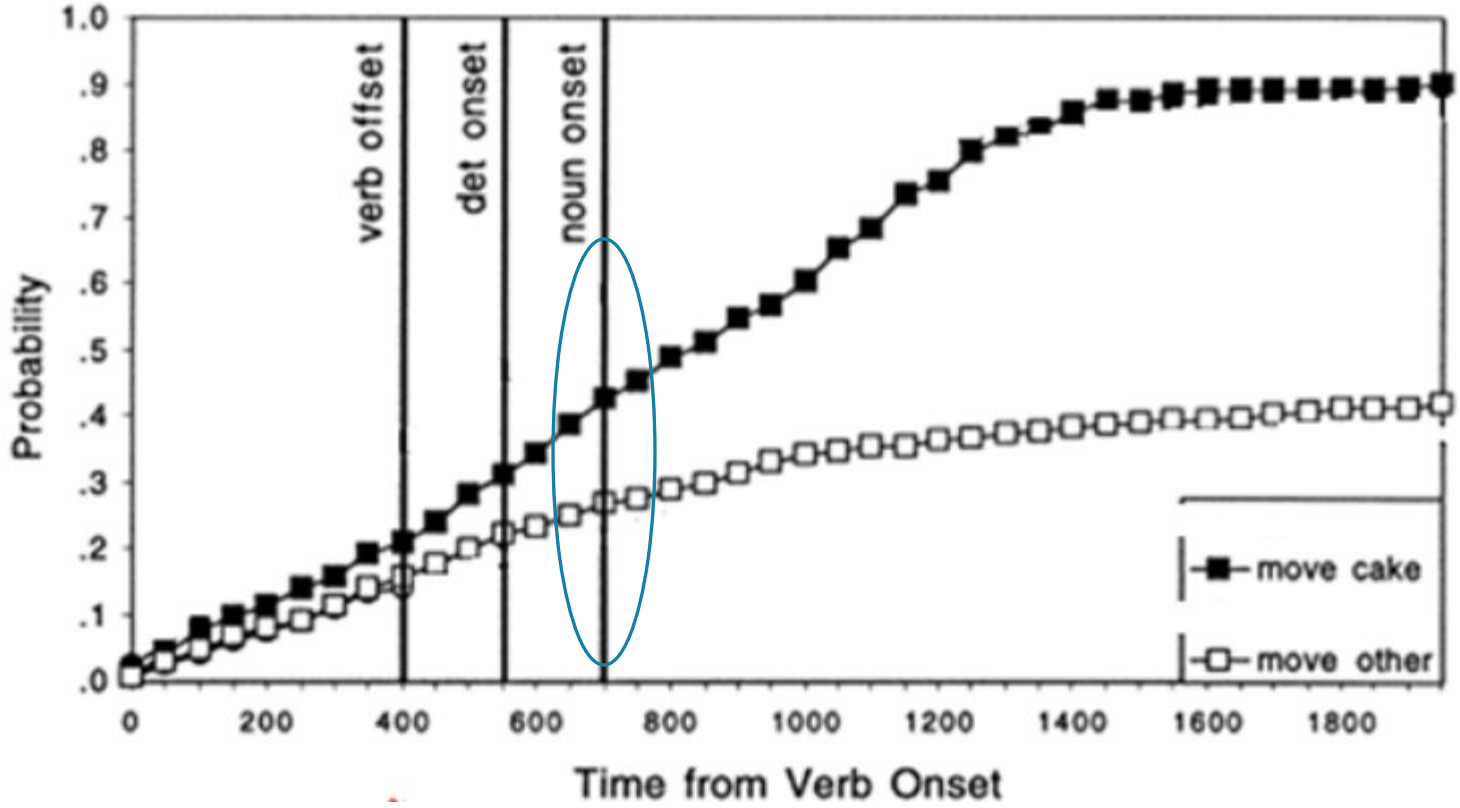
‘VISUAL WORLD PARADIGM’: EVIDENCE FOR ANTICIPATION

Look at the picture while hearing “the boy will move the cake”

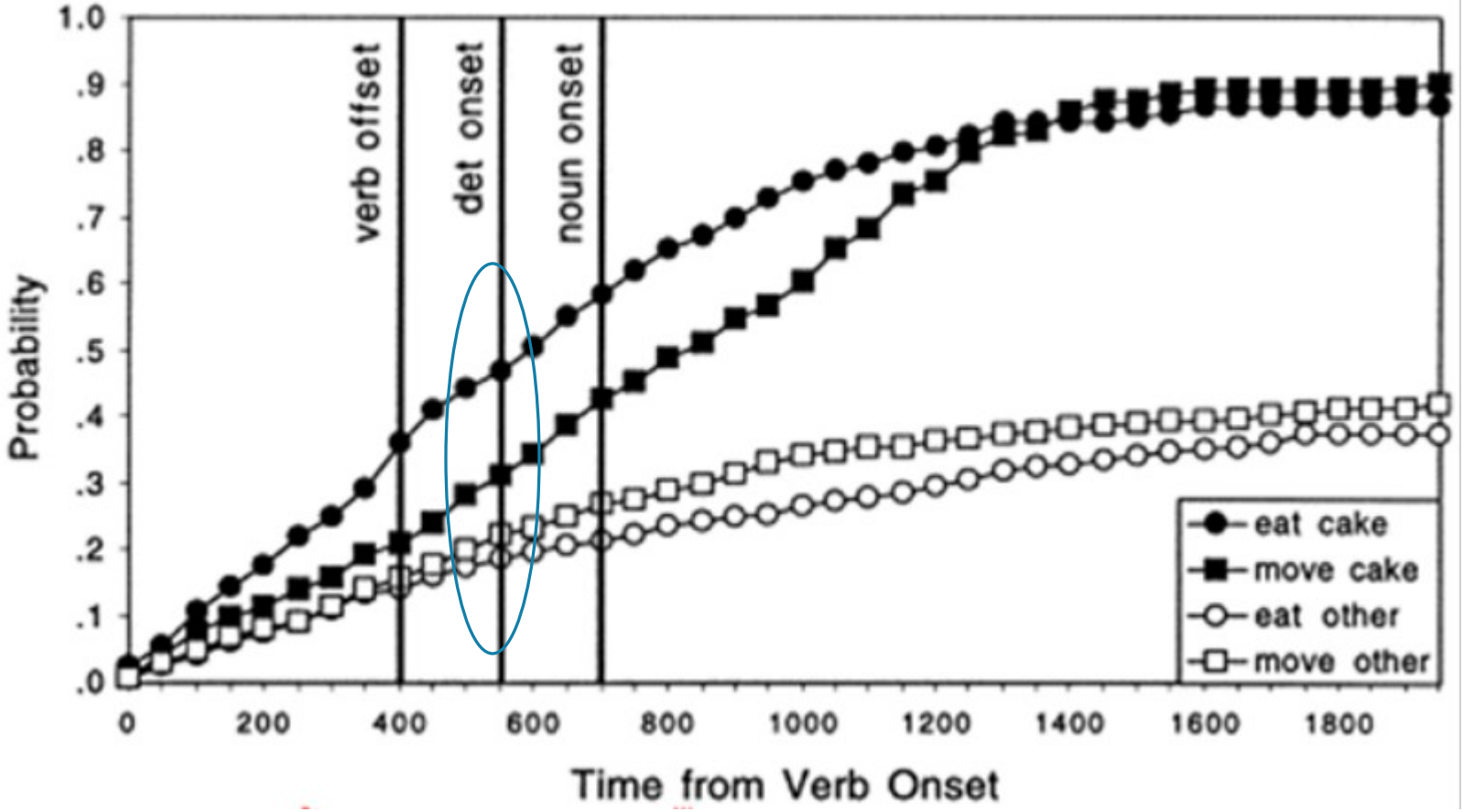


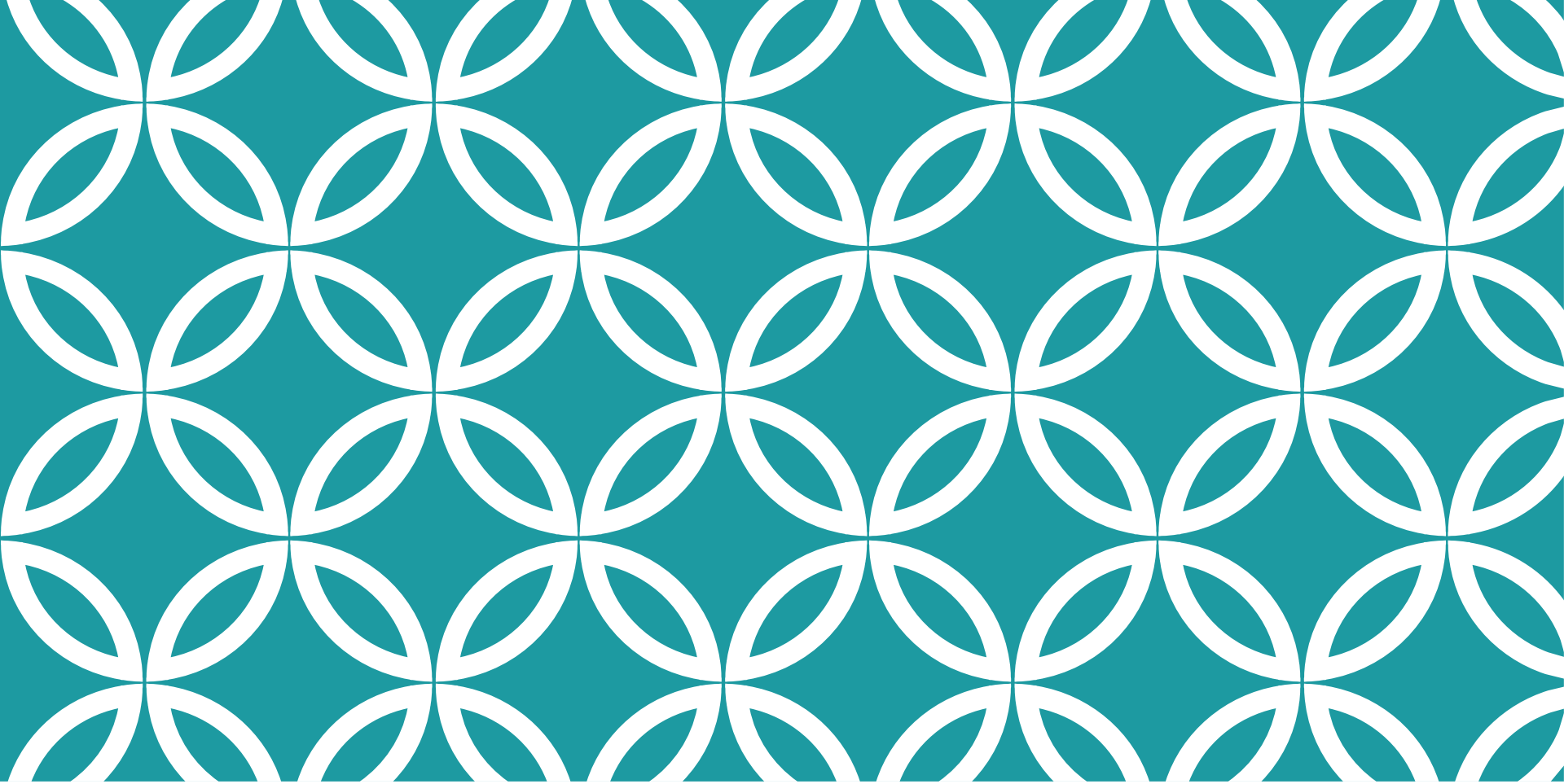
Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition, 73, 247-264

VISUAL WORLD PARADIGM: READING RESULTS. % TIME LOOKING AT 'CAKE' AND 'OTHER'



VISUAL WORLD PARADIGM: EVIDENCE FOR PREDICTION AT VERB





SUPPORT FOR STRUCTURE-PRUNING / SURPRISAL

And a few words on PCFGs

FRANK AND BOD'S CHALLENGE: DIFFICULTY = PROBABILITY WITHOUT SYNTAX (STATISTICAL KNOWLEDGE)

Rather than assume that comprehension is based on structure building, they test the idea by comparing 3 models of language

1. Probabilistic PSG
2. Markov process (N-gram)
3. Echo state network

Model quality == how accuracy of predicting human behavior.

Frank, S. L., & Bod, R. (2011). Insensitivity of the Human Sentence-Processing System to Hierarchical Structure. Psychological Science, 22(6), 829–834. https://doi.org/10.1177/0956797611409589

PSG VS. SEQUENTIAL REPRESENTATION

In ‘a’: assigning a structure from knowledge of grammar. Representing a hierarchy is implicit in this knowledge.

In ‘b’ representation of a linear sequence of elements

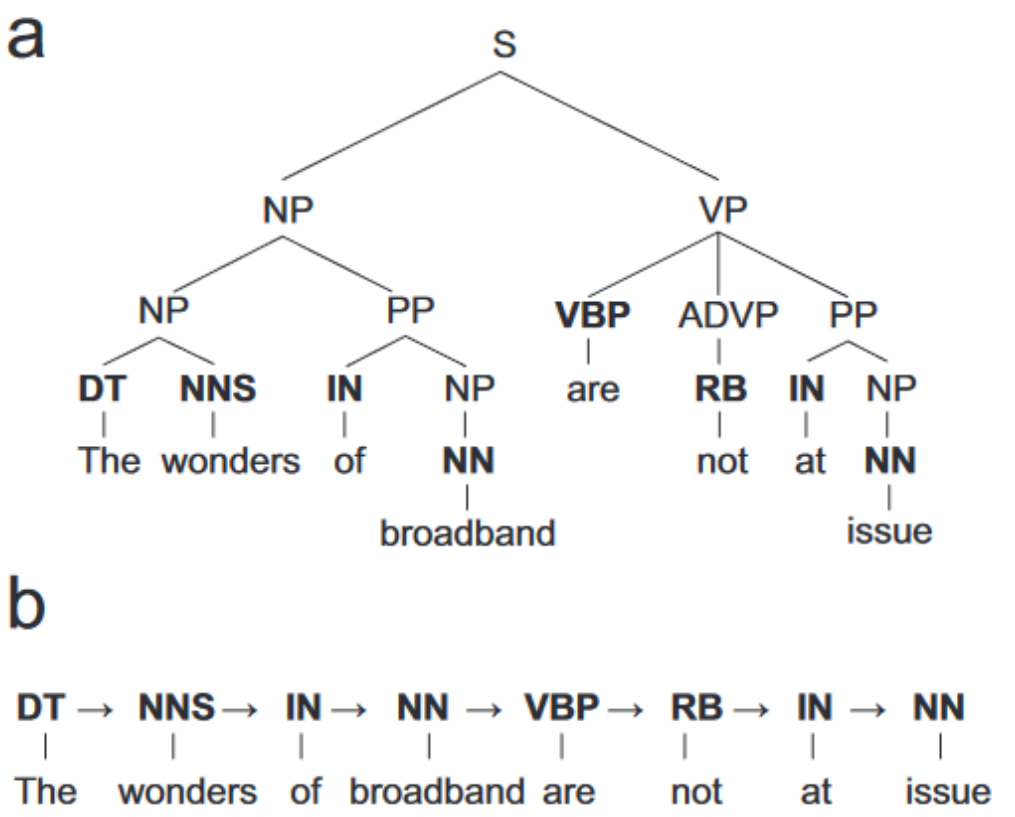


Fig. 1. Example (a) hierarchical phrase structure and (b) sequential structure for a sentence from the Dundee corpus (Kennedy & Pynte, 2005). Part-of-speech tags are shown in boldface (DT = determiner, NNS = plural noun, IN = preposition, NN = singular noun, VBP = present-tense verb, RB = adverb). Phrasal labels are shown in regular font (S = sentence, NP = noun phrase, PP = prepositional phrase, VP = verb phrase, ADVP = adverbial phrase).

TRAINING SET

‘Wall Street Journal’ corpus: 49K sentences

- This corpus has a syntactic-structure tree annotation for each sentences. Ideal for training Probabilistic PSG.

For Markov models and ESNs, F&B replaced each word its Part of Speech category

- The models learn to predict P.O.S not specific word. Improves accuracy for small datasets.
- Also, no need for an algorithm to learn to predict highly specific semantics, as syntactic p.o.s is the target for prediction.

PROBABILISTIC PSG (I)

They experiment with 4 types of PSG.

- A classic which just considers what continuations can follow each phrase (memory-less)
- Nth order versions, which consider which continuations can follow combinations of phrases.

Classic PSG	Conditionalized on 2nd level
NP -> N (P = 0.3)	VP NP -> N (P = 0.3)
NP -> Det NP (P = 0.7)	VP NP -> Det NP (P = 0.7)
	PP NP -> N (0.1)
	PP NP -> Det NP (P = 0.9)

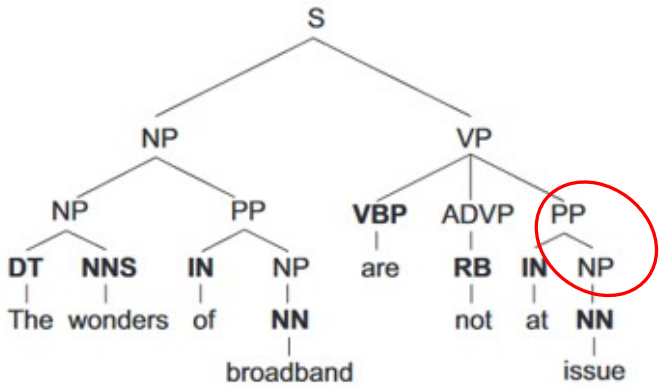
"In a standard probabilistic context-free grammar, the probability that a parent node will produce a particular set of children is conditional only on the parent node, but it is well known that parsing accuracy can be improved by taking the grandparent node (e.g., in Fig. 1a, VP, or verb phrase, is the grandparent of the rightmost IN and NP) into account (Johnson, 1998). Likewise, a rule's probability can be made conditional on information from even higher up in the parse tree"

PROBABILISTIC PSG (II)

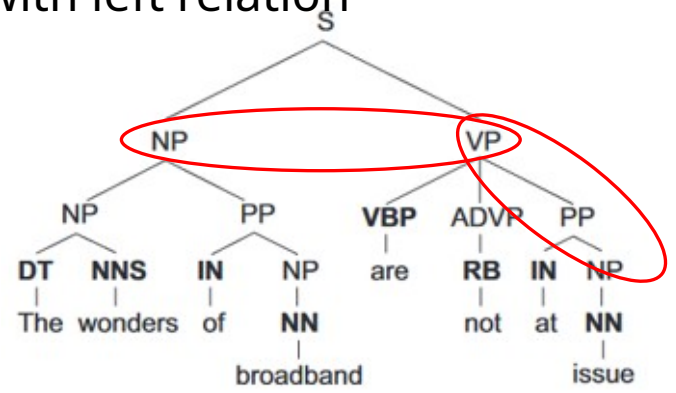
They make an effort of capturing as much structural knowledge as possible.

1. we obtained four different PSGs by varying the levels in the tree from which conditioning information was obtained: from only Level 1 (i.e., a standard probabilistic context-free grammar) up to Level 4, at most.
2. In addition, we induced four more PSGs, in which conditioning information was taken not only from ancestor nodes (e.g., grandparent nodes) but also from the ancestors' left siblings (e.g., in Fig. 1a, the left sibling of VP is NP), again varying the maximum number of levels up in the tree from one to four. In this manner, we obtained highly structurally sensitive syntactic models.

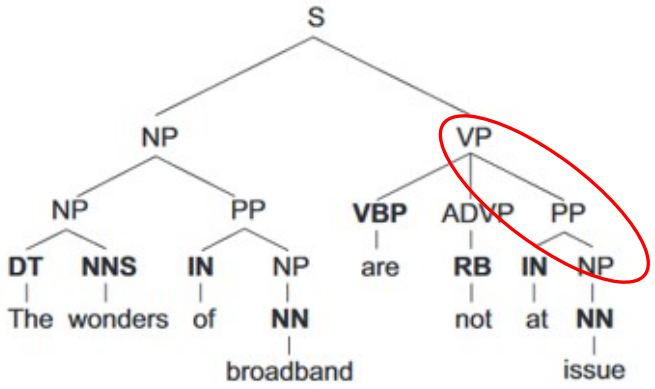
Regular probabilities



Nth =2,3,4 order probabilities with left relation



Nth =2,3,4 order probabilities



TYPES OF PSG LEARNED

Markov models. We used Markov models of first, second, and third orders (ns = 1, 2, and 3, respectively) in our experiments. In a Markov model of a given order (n), a symbol's probability depends on the context of n previous symbols only; that is, $P(w_t|w_{1\ldots t-1})$ is taken to equal $P(w_t|w_{t-n\ldots t-1})$. In our experiments, the probabilities of the sequences $w_{t-n\ldots t-1}$ and $w_{t-n\ldots t}$ were estimated from their occurrence frequencies in the training data. The raw frequency counts were smoothed because

MARKOV MODELS

MARKOV MODELS

First order Markov Model

	V	Det	N
V	0%	70%	30%
N	40%	40%	20%

2nd order Markov Model

	V	D	N
NV	0%	60%	40%
VN			
DN			
NN			

ECHO STATE NETWORKS

A recurrent architecture that learns sequential relations between input elements without strict limitation on temporal integration window (long range relations as well)

Learns to predict a P.O.S

Output activations treated as probabilities over upcoming P.O.S tag, and the correct tag. True P.O.S of next word used as true target.

MODEL EVALUATION

From each of the computational systems they derive a measure of Surprisal for the next word (p.o.s). All 3 approaches produce p(w) in context.

They evaluate Linguist Accuracy: How Surprising is the p.o.s for each system. Presumably, the better a system learns stats of language, the less surprised it is by the future

They evaluate Psychological Accuracy: how well do the surprise ratings produced by each model predict word reading times

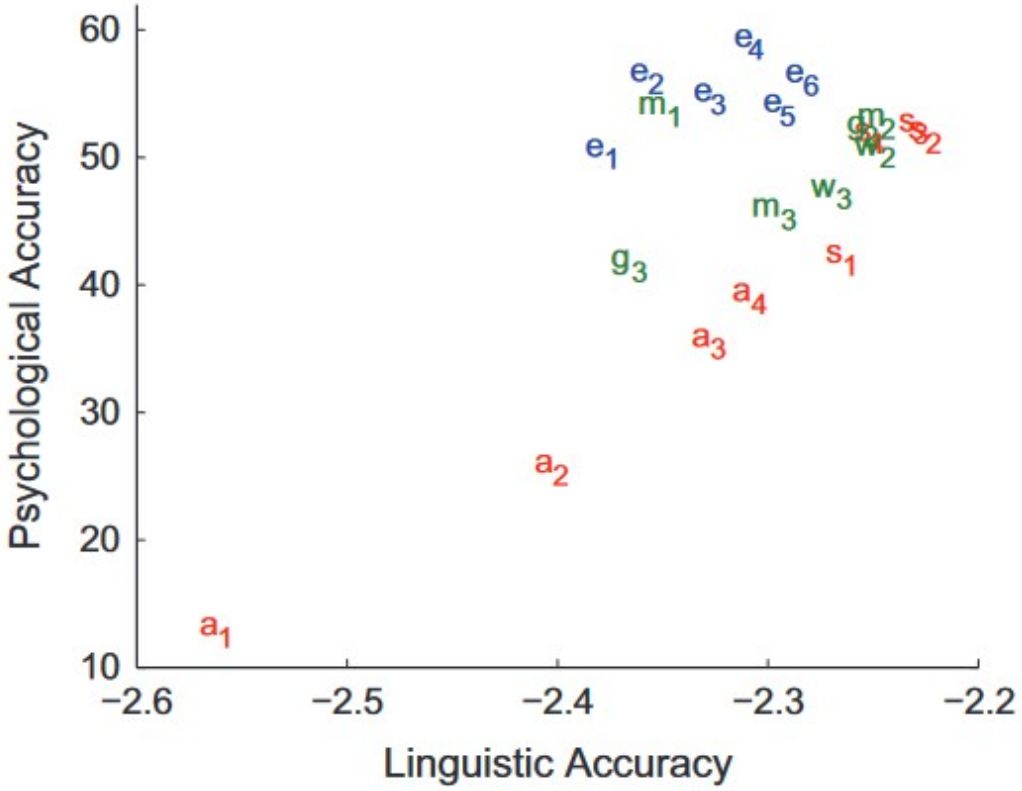
- Reading times obtained with eye tracker for >2K sentences
- Quantified by variance reduced when Surprisal is added as explanatory variable to a baseline model including just word-frequency and word-length

RESULTS

A's: PSGs with different constraint strength [Ss take left siblings)

M's Markov processes with different strength

E's echo-state networks



MODERN APPROACHES FOR WORDS IN CONTEXT AND APPLICATION TO BEHAVIOR

Modern computational approaches to language extend the Markov model and ESN to compute sophisticated quantities reflecting the meaning of a word in context

When the meaning of word-in-context is specified, it is possible to determine integration-difficulty in a sentential context.

- It is also possible to evaluate the computational model as a cognitive or neurobiological model

We first will discuss how to describe the meaning of a word

THE DISTRIBUTIONAL HYPOTHESIS

The distributional hypothesis

- words that occur in the same contexts tend to have similar

Meanings (Harris, 1954)

- “You shall know a word by the company it keeps” (Firth, 1957)

Examples:

- Cucumber, sauce, pizza, ketchup → Tomato
- Soundtrack, lyrics, sang, duet → Song

REPRESENTING WORDS AS VECTORS

Simple option

- Each word is represented as a vector whose length is *V*: the size the entire vocabulary (unique entries)
- The vector capture the co-occurrence of the target word with each other word in the Vocabulary, where co-occurrence is defined as adjacency within a certain distance window (e.g., x words away)

	song	cucumber	meal	black
tomato	**0**	**6**	**5**	**0**
book	2	0	2	3
pizza	**0**	**2**	**4**	**1**

We expect words with similar meanings to have similar vectors

Given this representation it is simple to compute word similarity

The granularity of 'adjacency' is flexible: can be sentence/paragraph/document etc'

REPRESENTING WORDS AS VECTORS

Co occurrence frequencies using raw counts are intuitive, but problematic: some extremely frequent words can dominate the rows and make all words appear quite similar (‘dominate’) the representation.

So rather than raw counts we need a measure that quantifies how frequently 2 measure appear ‘together’, but normalized by how probable their co-occurrence would be if they were independent.

For this we use PMI: joint probability, divided by product.

$$PMI = \log \frac{p(x,y)}{p(x)p(y)}$$

	song	cucumber	meal	black
tomato	**0**	**6**	**5**	**0**
book	2	0	2	3
pizza	**0**	**2**	**4**	**1**

COMPRESSING THE WORD VECTORS

There are two main problems with the word-vector matrices

1. They are massive (> 100K rows/columns)
2. They have many entries that are 0 because some words will not appear next to others: these are called sparse matrices

How do we compress this representation?

3. We can use dimensionality-reduction techniques like SVD.
    1. From a 100K^2 matrix, we get a [100K x D] matrix with D < 100K, and D reflecting the latent dimensions on which each word receives a score.
4. Alternatively, we can use methods that compute directly the lower-dimensional vector: Word-Embedding methods

PRINCIPLES OF WORD-EMBEDDING

1. A lower-dimensional space is used to represent all words (e.g., 300-500 Dimensions).
2. All words are embedded into the same space
3. It ends up that: similar words have similar vectors

WORD2VEC: COMPUTING EMBEDDINGS FROM ADJACENCY

1. Principle: Words that appear in similar contexts should have similar vector representations
2. Word2Vec trains a system that predicts a TargetWord from the contexts in which it appears (e.g., the 2 words +- before/after it anywhere in the text)

WORD2VEC: COMPUTING EMBEDDINGS FROM ADJACENCY

The input: one-hot vectors

the: (1,0,0,0,0); cat: (0,1,0,0,0); sat: (0,0,1,0,0); on: (0,0,0,1,0); floor: (0,0,0,0,1)

A training item: The words before and after the target word

Example: the cat **sat** on floor

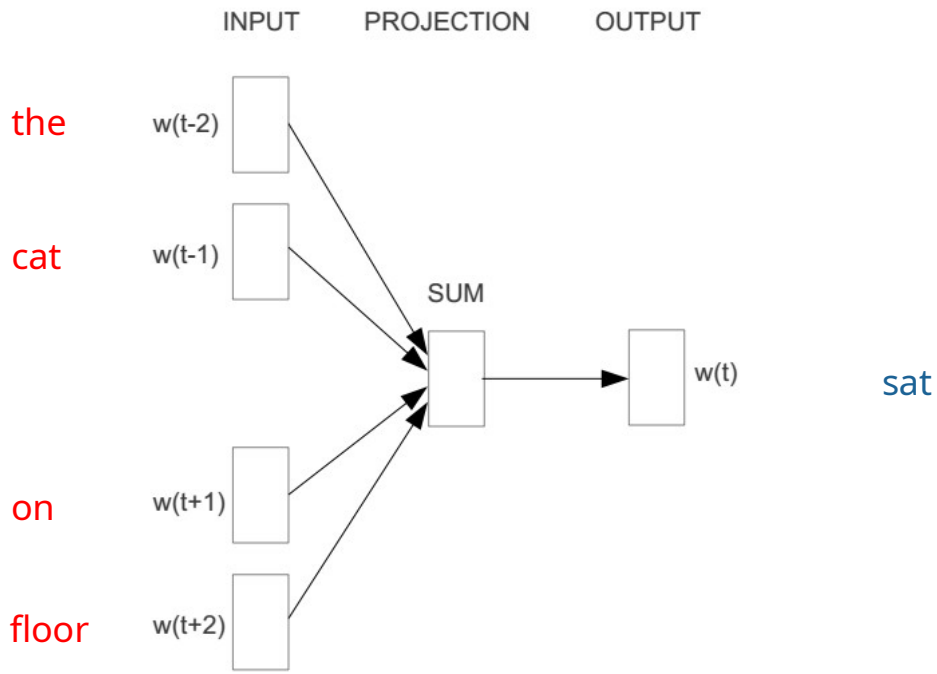
'sat' will be the word to be predicted;

This approach is called Continuous Bag of Words (CBOW). It ignores word order (shuffled context will give same result) but there's sufficient info in the context to infer a miss word.

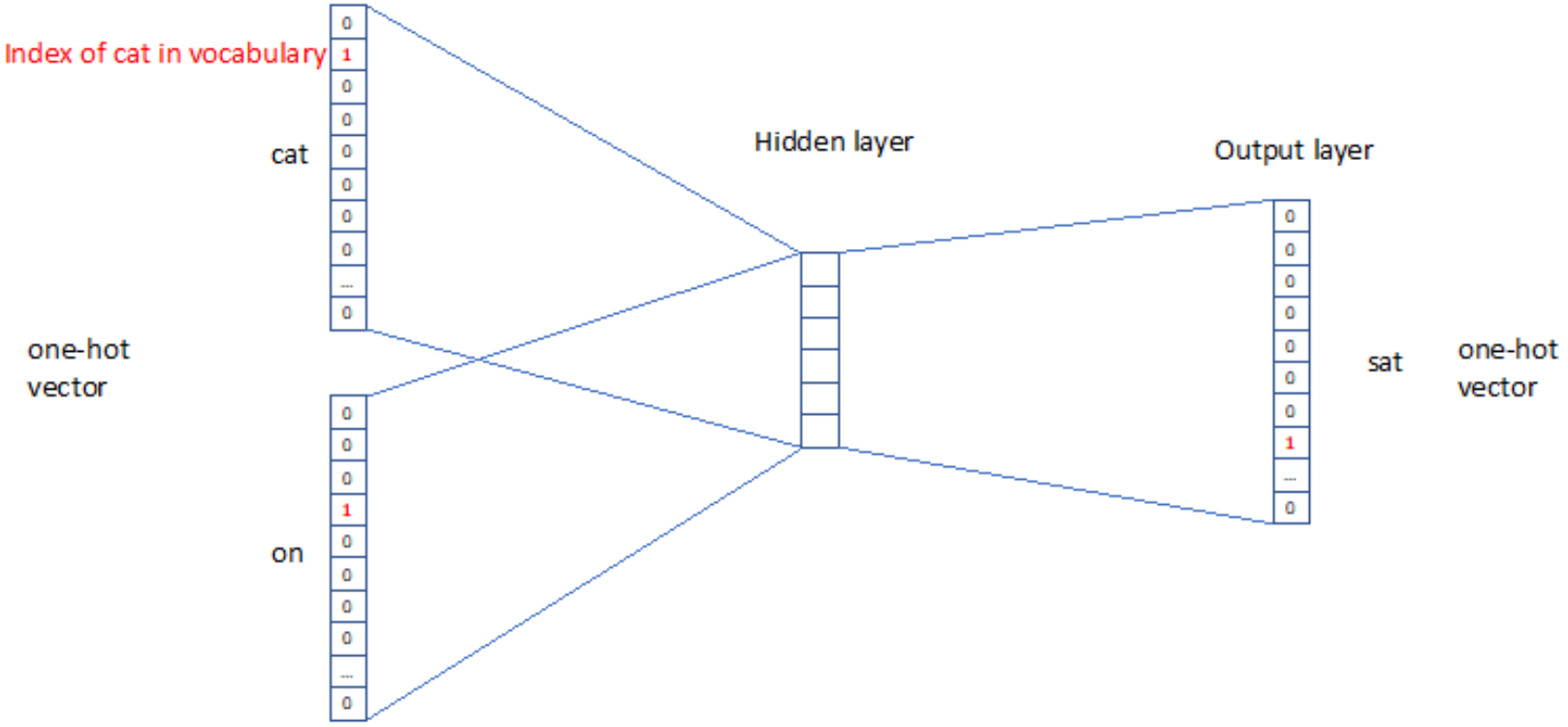
WORD2VEC – CONTINUOUS BAG OF WORD

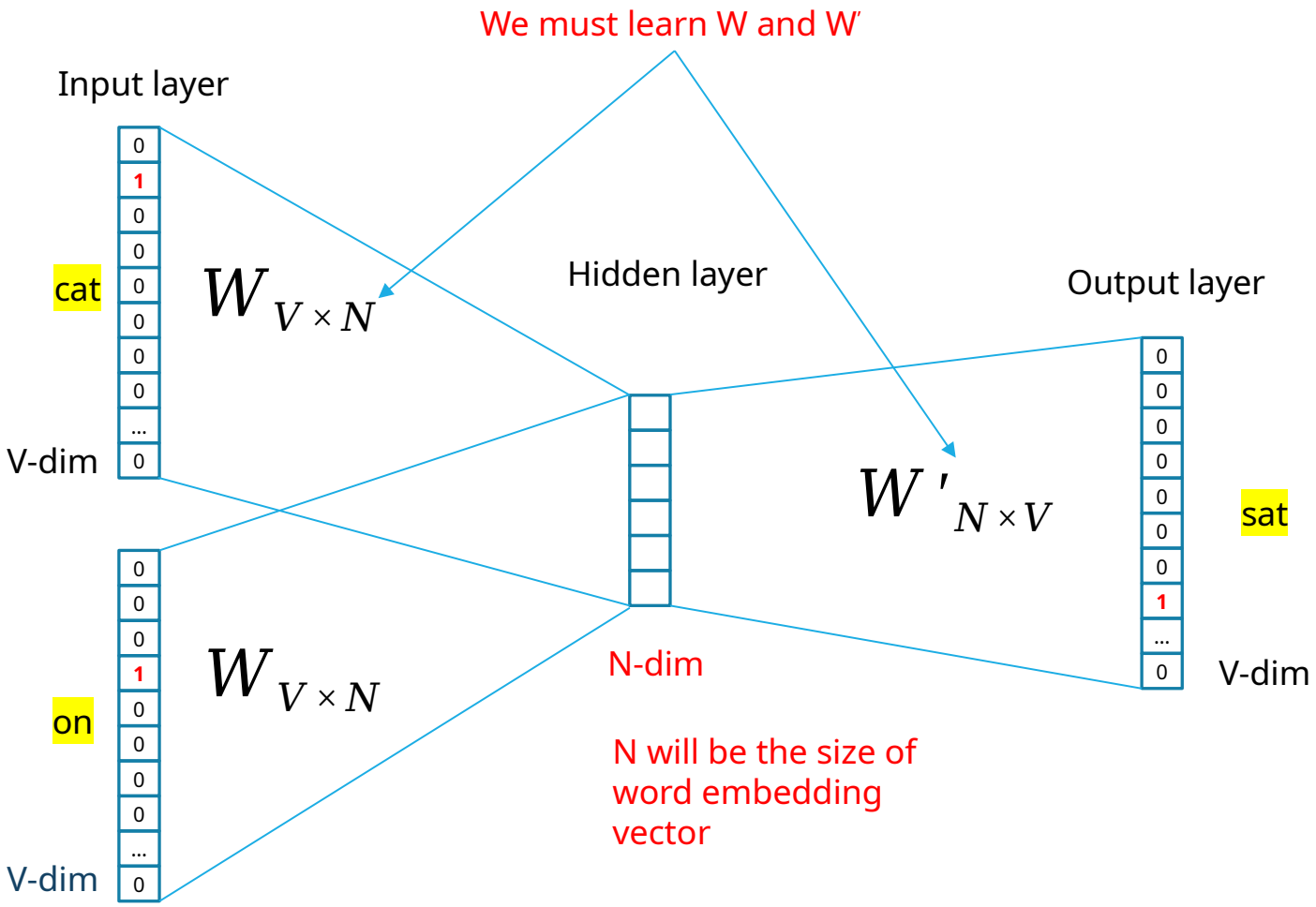
E.g. “The cat sat on floor”

- Window size = 2

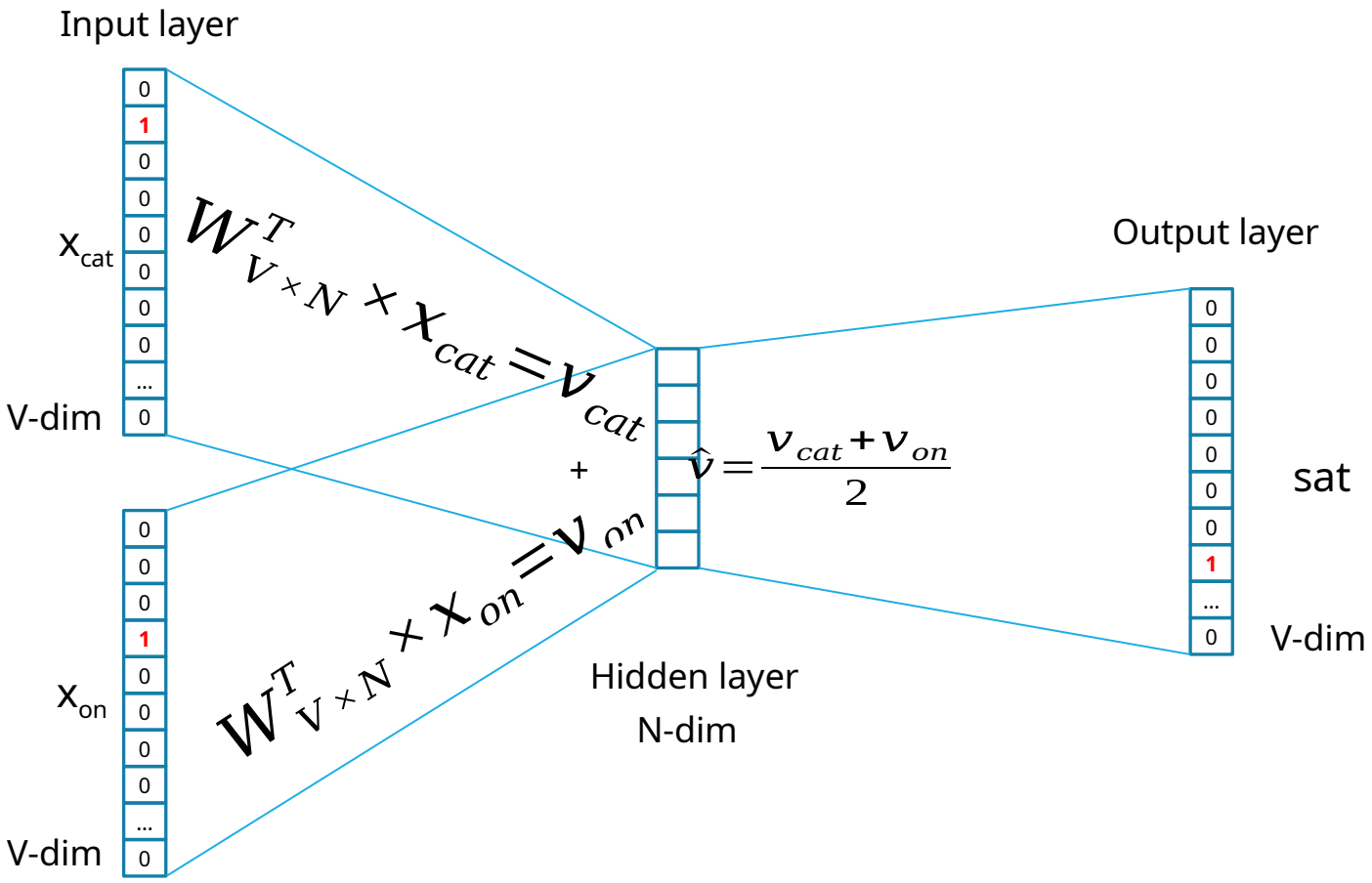


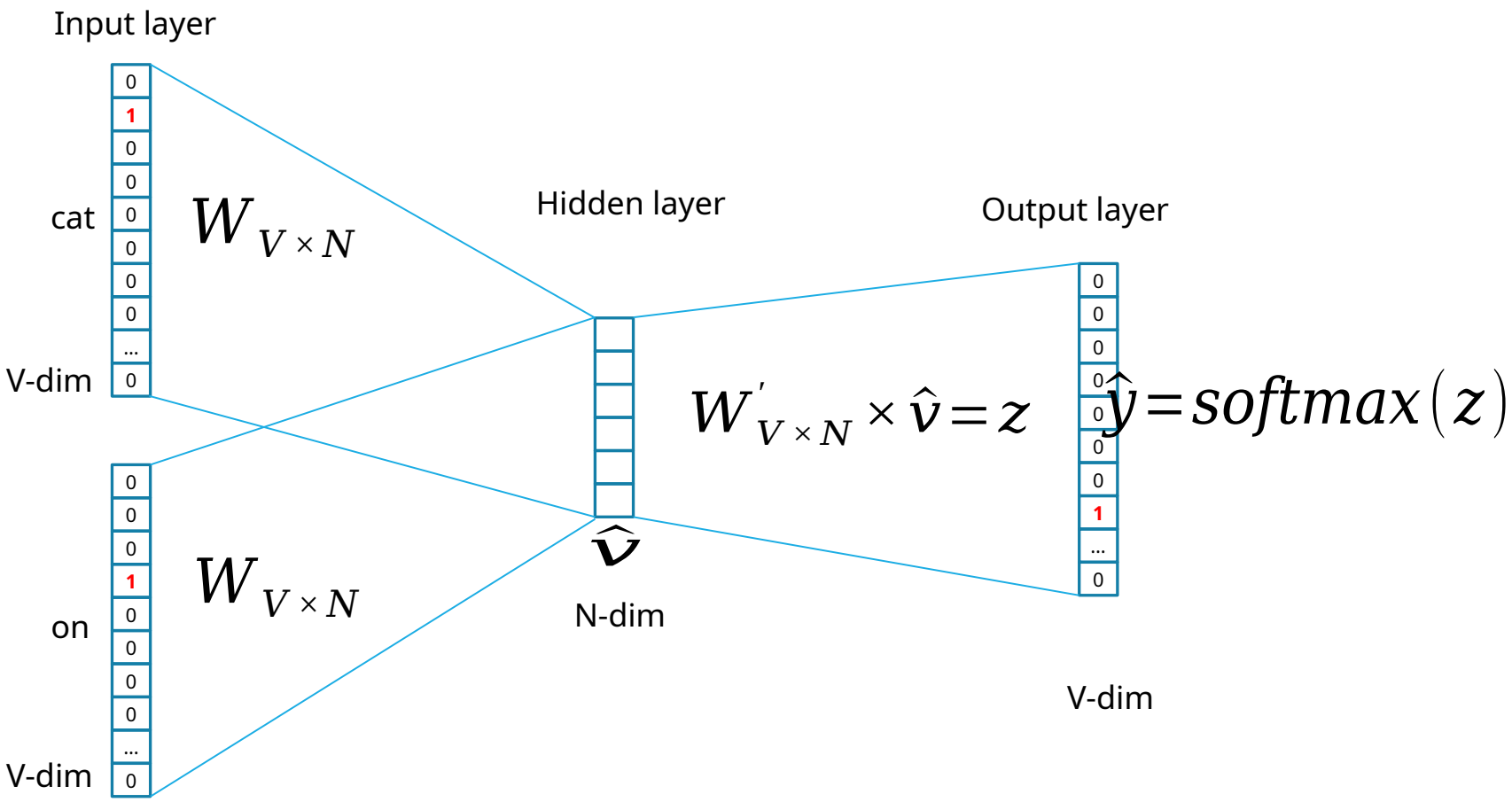
www.cs.ucr.edu/~vagelis/classes/CS242/slides/word2vec.pptx

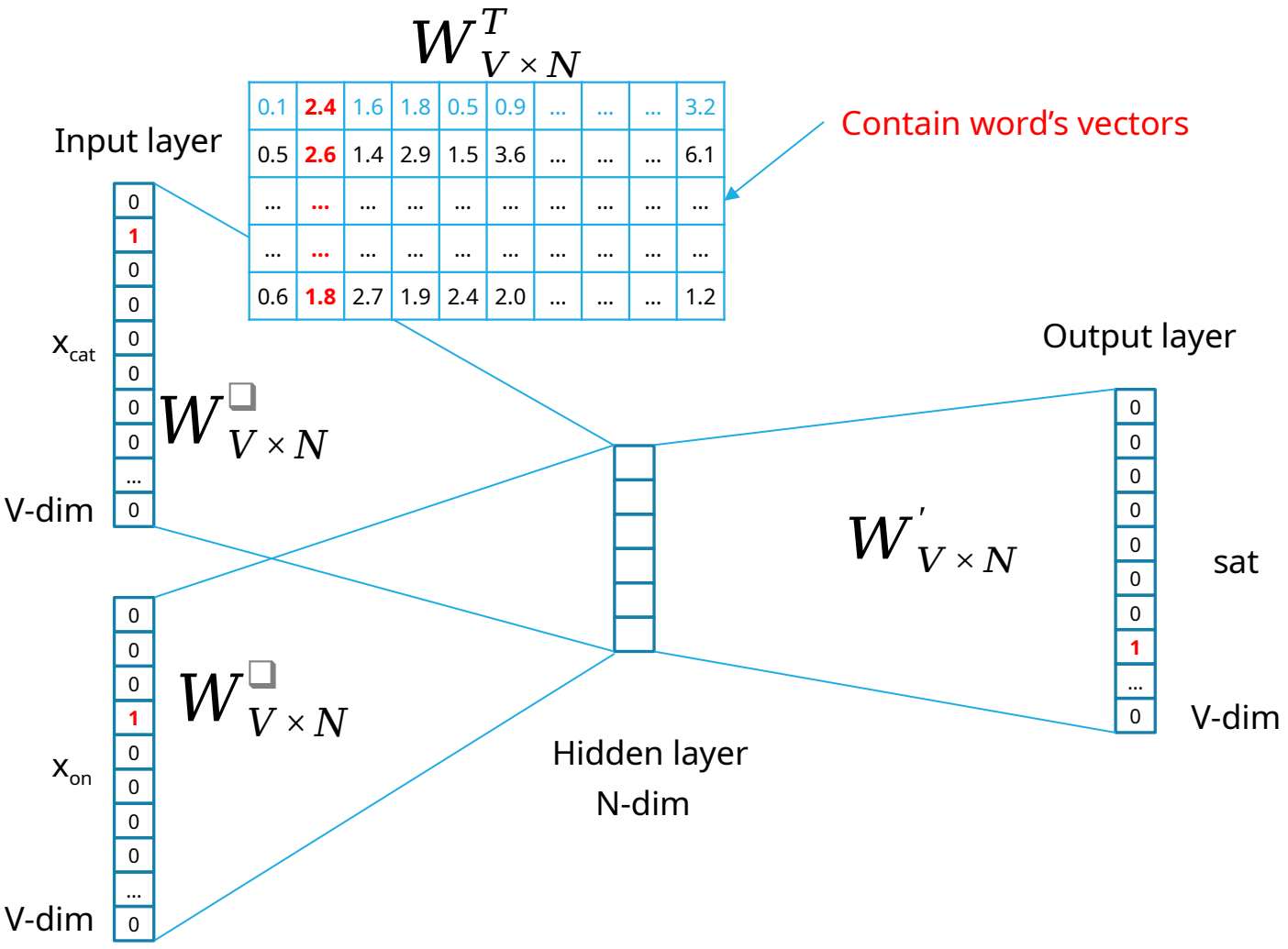

Index of cat in vocabulary
cat
one-hot vector
on
Hidden layer
Output layer
sat
one-hot vector



www.cs.ucr.edu/~vagelis/classes/CS242/slides/word2vec.pptx


Input layer
x_{cat}
V-dim
x_{on}
V-dim
$W^T_{V\times N} \times x_{cat} = v_{cat}$
$W^T_{V\times N} \times x_{on} = v_{on}$
+
$\hat{v} = \frac{v_{cat} + v_{on}}{2}$
Hidden layer
N-dim
Output layer
sat
V-dim


Input layer
cat
V-dim
on
V-dim
$W_{V \times N}$
$W_{V \times N}$
Hidden layer
$\hat{v}$
N-dim
$W'_{V \times N} \times \hat{v} = z$
Output layer
$\hat{y} = softmax(z)$
V-dim



We can consider either W or W' as the word's representation. Or even take the average.

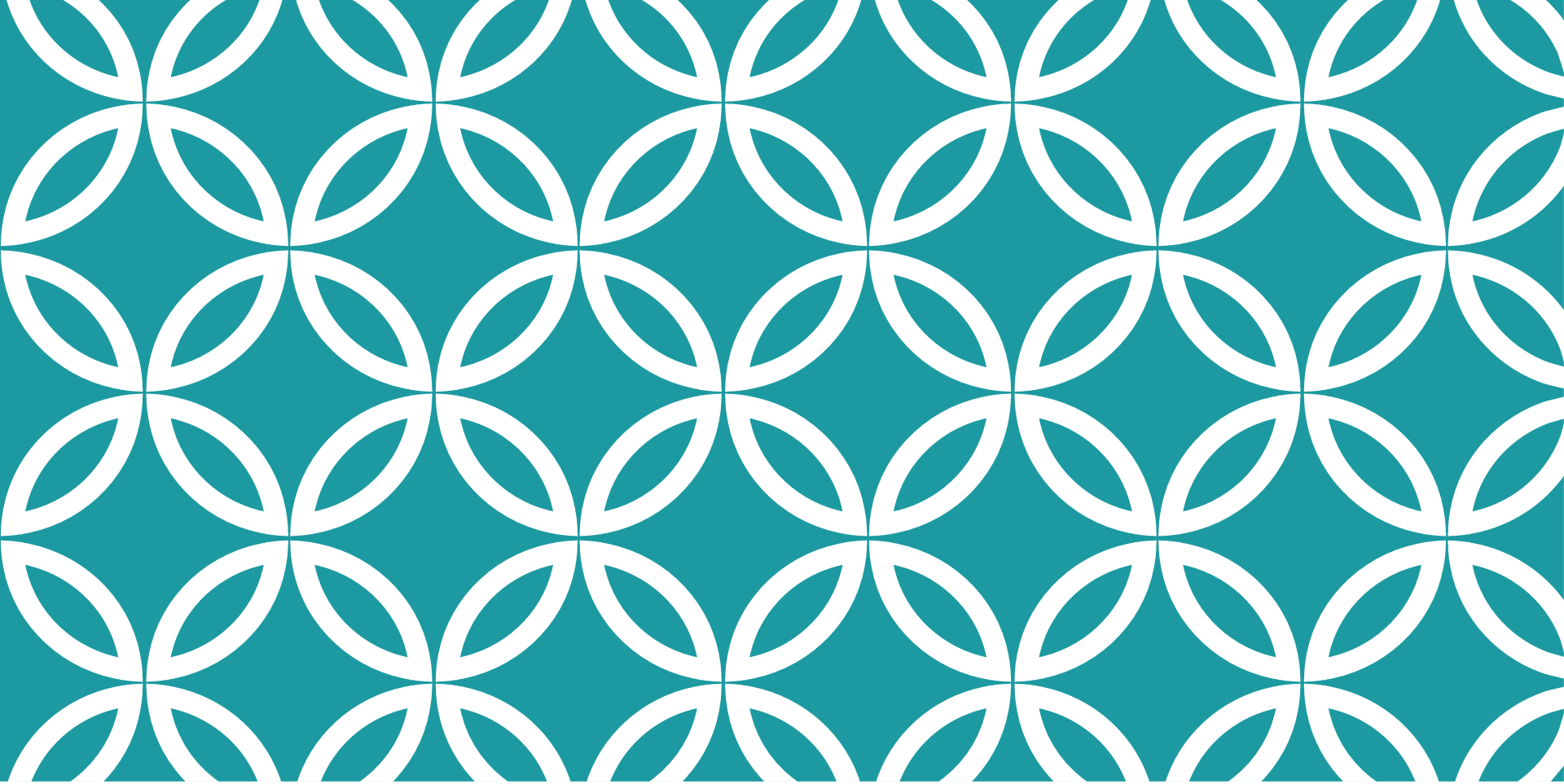
SAMPLE RESULTS WORD2VEC; DIFFERENT WINDOW SIZES: WORD WALK

Window size = 3

Word	Cosine distance
go	0.488083
snipe	0.464912
shoot	0.456677
fly	0.449722
sit	0.449678
pass	0.442459
climbs	0.440931
walked	0.436502
ride	0.434034
stumble	0.426750
bounce	0.425577
travelling	0.419419
walking	0.412107
walks	0.410949
trot	0.410418
leaping	0.406744
sneak	0.401918
climb	0.399793
move	0.396715
wait	0.394463
going	0.391639
shouted	0.388382
roam	0.388073
thrown	0.384087
get	0.383894

Window size = 30

Word	Cosine distance
walking	0.486317
walked	0.430764
walks	0.406772
stairs	0.401518
go	0.399274
sidewalk	0.385786
stand	0.380480
cortege	0.371033
wheelchair	0.362877
strapped	0.360179
hollywood	0.356544
carousel	0.356187
grabs	0.356007
swim	0.355027
breathe	0.354314
tripped	0.352899
cheer	0.352477
moving	0.350943
inductees	0.347791
walkway	0.347164
shout	0.346229
pounding	0.340554
blvd	0.339121
crowd	0.338731
levada	0.334899



SURPRISAL & WORD EMBEDDINGS

As models of behavior

CAN MODEL-PREDICTIONS FROM SURPRISAL AND WORD EMBEDDINGS PREDICT BEHAVIOR?

Russo et al. 2020

1. Presented an audio narrative (in Italian) to 31 listeners
2. Compute surprisal for each word in narrative from probabilistic knowledge model
3. Computer vector-space representation (word embedding)

DEF. SURPRISAL

$$\text{surprisal}(t) = -\log_{10}(\mathrm{P}((W_t|W_1 \ldots., W_{t-1}))$$

$W_N\}$ the probability of observing the word W_t given the whole left-side context $P(W_t \mid W_1,..., W_{t-1})$ can be approximated to $P(W_t|W_{t-2}, W_{t-1})$ (i.e. the probability of observing the word W_t given only the two preceding words). Surprisal values estimated by trigram models have been used

DEF. VECTOR SPACE (WORD EMBEDDING) SEMANTICS

Define ‘word meaning’ in terms of adjacency of the word to each of 1000 ‘anchor’ words (similar to Mitchell).

The i-th component (1=1:1000) of the word vector w (c_i) is estimated as the ratio between the conditional probability of the context word c_i given the word w and the (unconditional) probability of the context word c_i

$$v_i = p(c_i|w)/p(c_i)$$

The adjacency window is set to 10 words

Semantic similarity between 2 words is quantified via cosine similarity.

DEF. SEMANTICS-WEIGHTED SURPRISAL (I)

Developed originally by Mitchell and Lapata (2009)

Integrates the probabilities obtained by the trigram model and the semantic similarity calculated from the vector space model:

the trigram probability (2nd Order Markov Estimate) is scaled by a positive factor depending on the semantic similarity of the current word with its recent history

- Positive factor between 0 and INF.
- i.e., upscales or downscales the trigram probability

DEF. SEMANTICS-WEIGHTED SURPRISAL (II)

1. Markov-based surprisal of
2. Semantic weight factor for

$$SwS = p(W_t|W_{t-1}, W_{t-2})\Sigma_i w_i h_i p(c_i)$$

The semantic weight factor sums over all features *i* , where

is value of feature i for

is the semantic vector capturing the context {t-3 : t-6}. Obtained by multiplication of the semantic vectors of the four content words preceding the lower bound of the trigram, but also weighing the semantic contribution of each context word by its frequency:
scaling factor

FMRI INVESTIGATION

REGRESSION MODEL

Predicting variables are descriptors associated with each words learned

Regressors of no theoretical interest:

- Raw lexical frequency
- Word duration
- RMS (Energy) amplitude of word sound

Regressors of interest:

- Surprisal
- Semantic weighted surprisal

Control condition: Reversed speech stimuli with words coded in the exact same way.

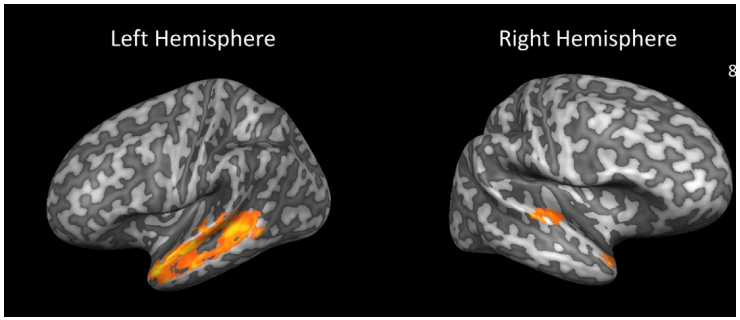
MAIN RESULTS

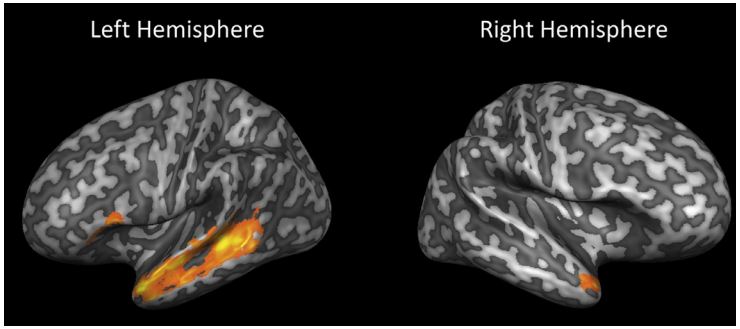
Whole-brain, voxel-based fMRI data analysis

LS (LEFT) and SwS (RIGHT) predictors identify largely similar brain systems:

- In all significant clusters BOLD is positively correlated to LS or SWS.

Model-fit analysis suggests that in many of these areas, SwS is a more accurate model.



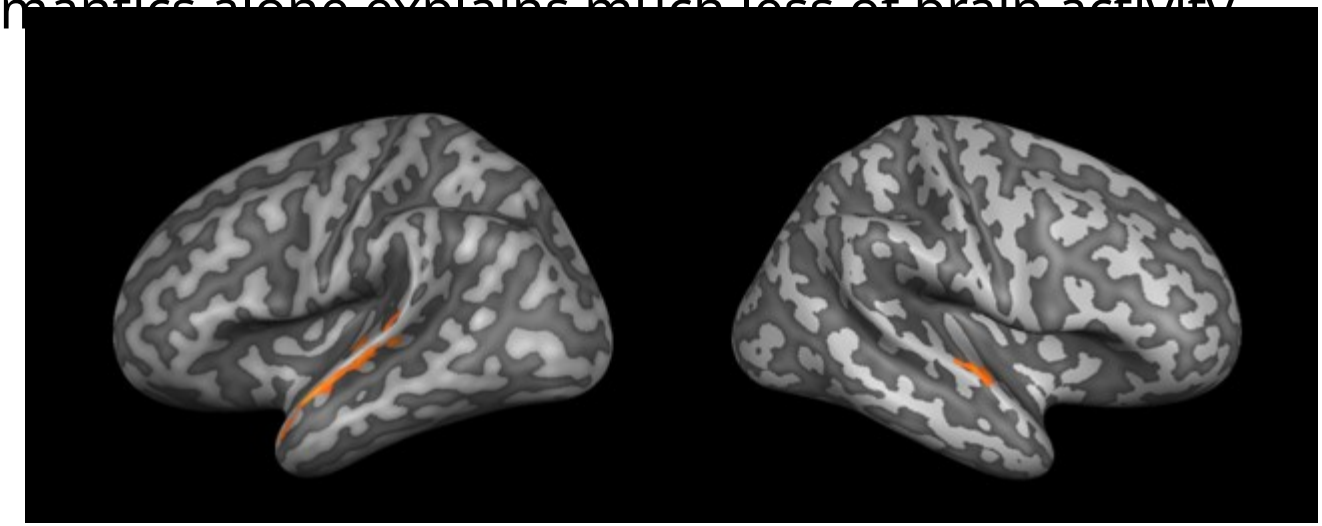


DO WE NEED PROBABILITY AT ALL?

$$SwS = p(W_t|W_{t-1}, W_{t-2}) \Sigma_i w_i h_i p(c_i)$$

What happens if we just take the semantic component as predictor?

Semantics alone explains much less of brain activity

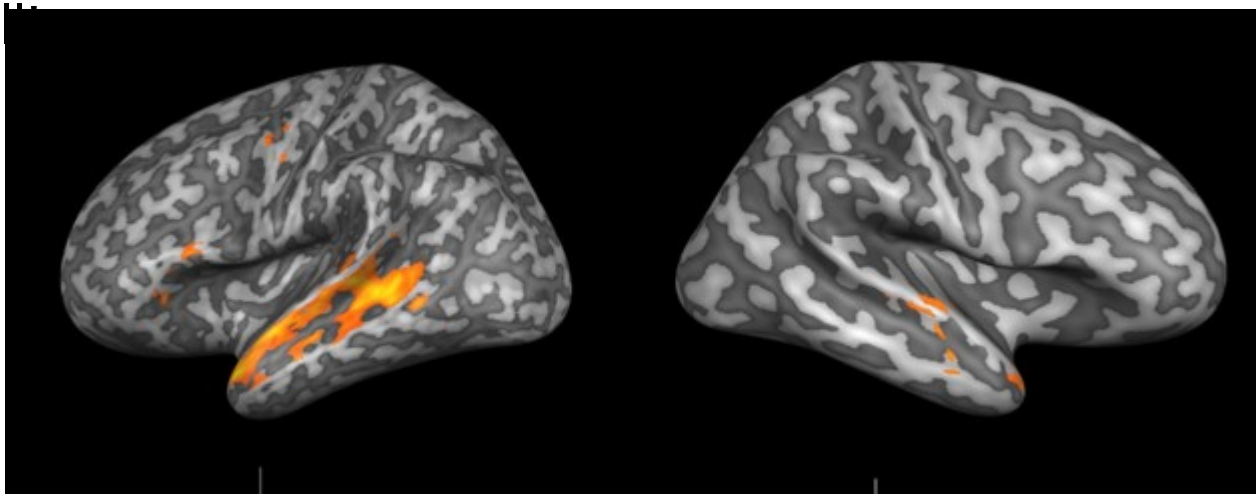


DO WE NEED CONDITIONAL PROBABILITY?

$$SwS = p(W_t|W_{t-1}, W_{t-2}) \Sigma_i w_i h_i p(c_i)$$

Words that are less predictable given W-1; W-2 could just be less frequent in language.

Prediction from Lexical Frequency (Raw) does good job as well



CONSIDERATIONS IN DISCUSSION

Higher activity may reflect prediction errors, which are a signal to other brain areas to update based on errors

It may be difficult to dissociate different surprise models, as they may be highly correlated:

- SwS and LS predictors were correlated at r = + 0.77.

Only SwS activated a cluster in the L-IFG

SUMMARY OF LANGUAGE MODULE

We covered three approaches for explaining the mental capacity for language:

1) memorizing which words or phrases go together;

2) (implicit) knowledge of abstract rules referred to as transformations;

3) statistical knowledge about co-occurrence of linguistic elements, where the main debate is on what these elements are: grammar-based structures as captured by probabilistic grammars, or parts of speech for which linear transitions are learned.

We have discussed ways in which words can be represented in computational systems via word embeddings

We have discussed a study that shows how contextual information can be modeled and used when accounting for human behavior, merging contextual semantic fit and contextual probability.