LSTMs
vs
hierarchical structure in language

Marco Baroni

Facebook AI Research

In collaboration with

German Kruszewski
FAIR

Yair Lakretz
NeuroSpin

Dieuwke Hupkes
ILLC

Motivation

- RNNs, LSTMs in particular, are amazing
- Train them on (variants of) language modeling, get improvements in tasks that required big pipeline of knowledge-based NLP tools
- What does this tell us, as theoretical linguists?
  - Nature vs nurture debate, etc.
- Questions I am currently interested in:
  - Are LSTMs really learning something about deeper, latent aspects of grammar, such as hierarchical structure?
  - If they do, how?
  - How does LSTM language processing compare to human language processing? [not discussed here]

Outline

- LSTMs are sensitive to hierarchical structure in language
  - Study 1: reading tea leaves in Principle Component space
  - Study 2: predicting syntactic tree depth with LSTMs

- How do LSTMs track long-distance linguistic information?
  - Study 3: meet the grandmother cells of long-distance number agreement

LSTM recap

- $C_t = f_t C_{t-1} + i_t \hat{C}_t$
- $h_t = o_t \tanh(C_t)$

- $i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$
- $f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$
- $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$
- $\hat{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C)$

Study 1:
LSTM sensitivity to constituent structure

Method

- French LSTM language model:
  - trained on 100M words from frWaC
  - 2 layers, 500 units per layer, dropout 0.5
  - Vocabulary size: 100k words
- Trained model applied to corpus of artificially constructed sentences with controlled syntactic structures
- Extract all word-by-word hidden vectors, compute their PCA
- Visualize trajectories of hidden vectors of same-structure sentences in first two PCs

The syntactic structures corpus

- 10k artificially generated French sentences
- Controlling for size of NP and VP, PoS of words at NP/VP boundaries (adjective, noun, ..., auxiliary, verb, ...), tense
- E.g.:
  - 2/6; N/Aux; P [cette cousine] [a consulté une lettre de boulanger]
  - 4/4; N/Aux; F [les noix du jardin] [vont nourrir ces frères]
  - 6/2; N/Aux; P [ces citrons du marché du patelin] [ont pourri]
- 66 different constructions, with about 160 instances of each

An example construction

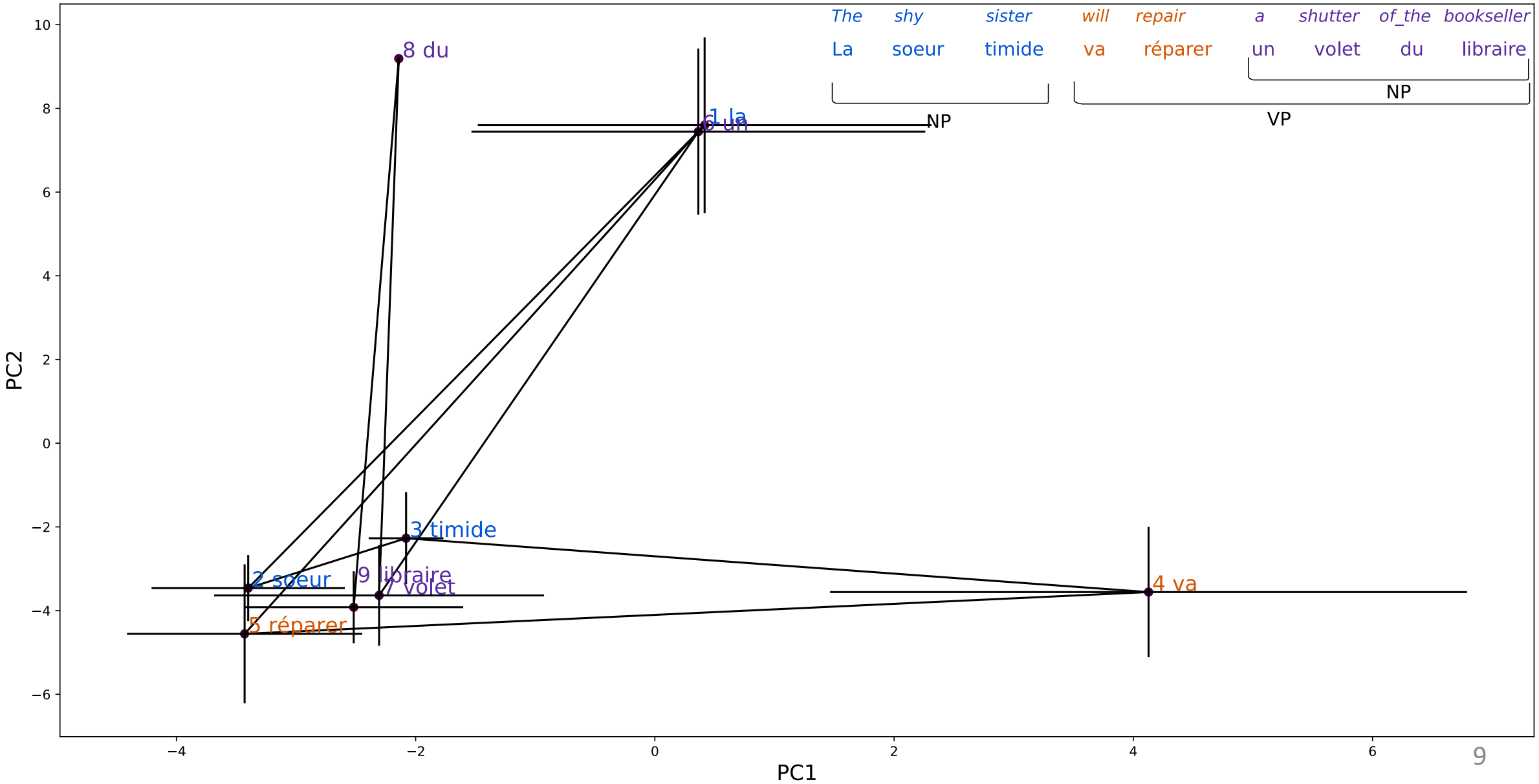
La	soeur	timide	va	réparer	un	volet	du	libraire
The	*shy*	*sister*	*will*	*repair*	*a*	*shutter*	*of_the*	*bookseller*
	...			...		...		
Un	cousin	nerveux	va	manger	ces	fruits	du	commerce
Des	frères	malades	vont	manger	des	pruneaux	du	verger
Ces	cousines	nerveuses	ont	bricolé	les	chaises	du	boulanger
DET	ADJ	NOUN	AUX	VERB	DET	NOUN	PREP	NOUN

NP

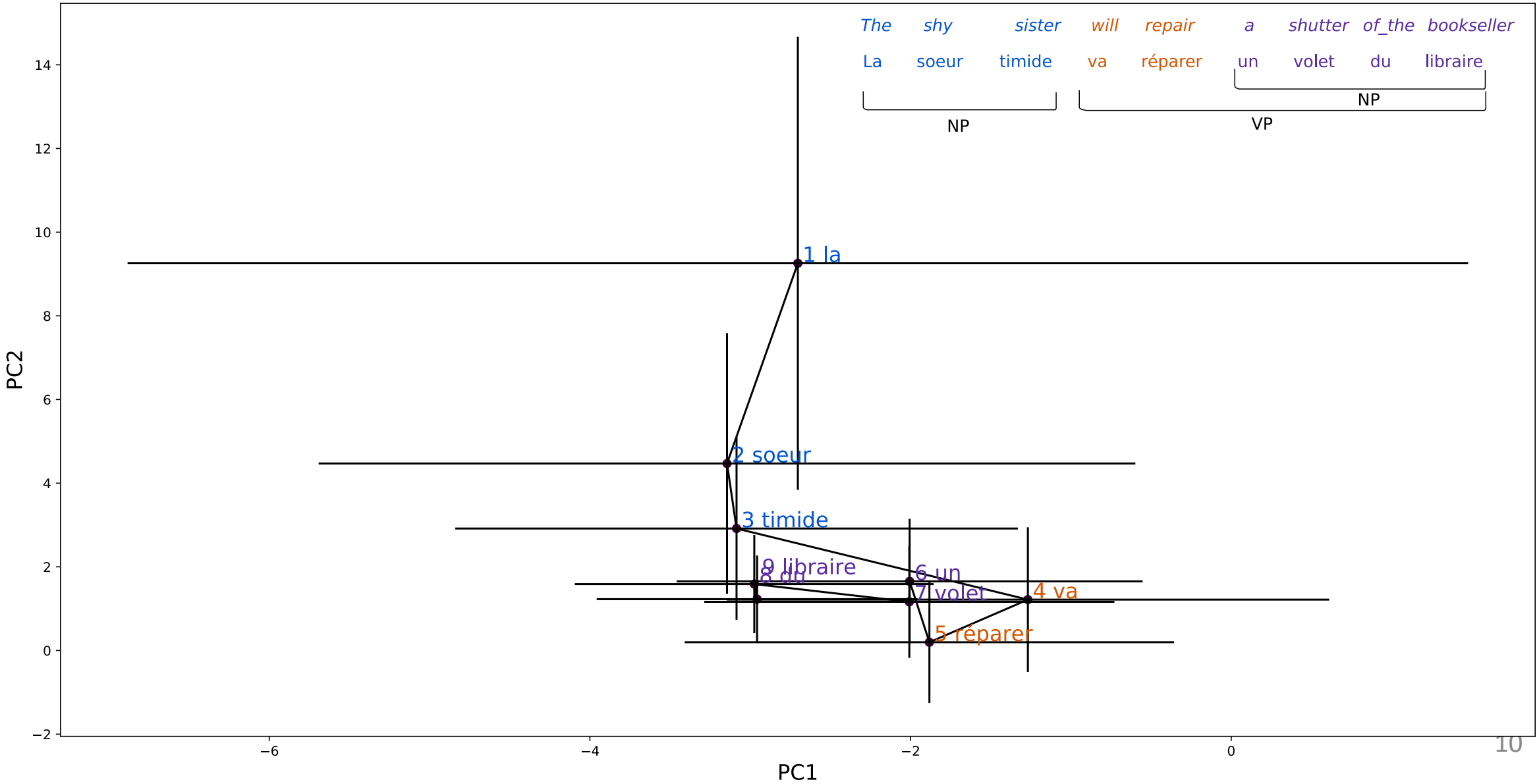
NP

VP

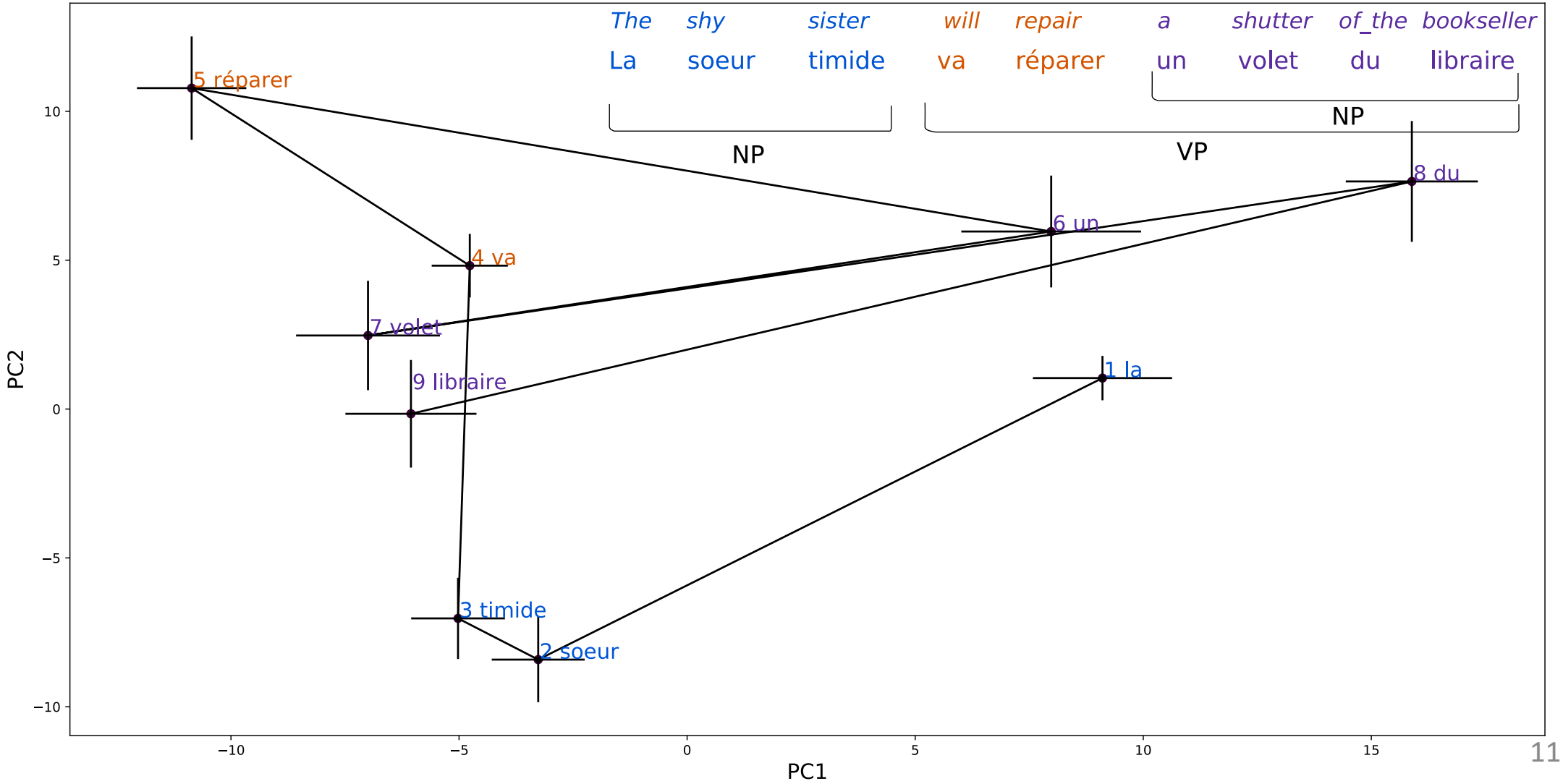
Word embeddings



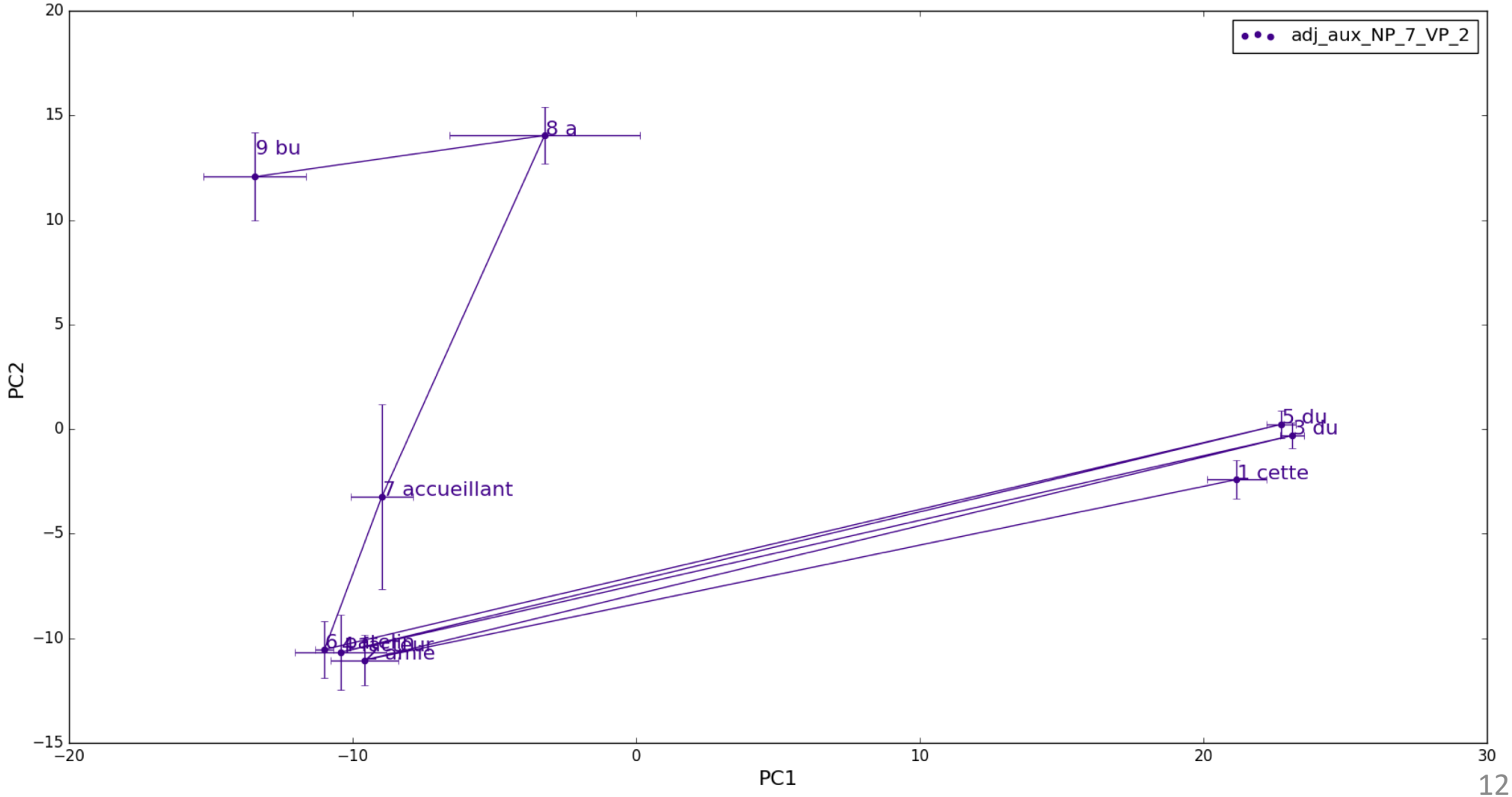
Bags of word embeddings



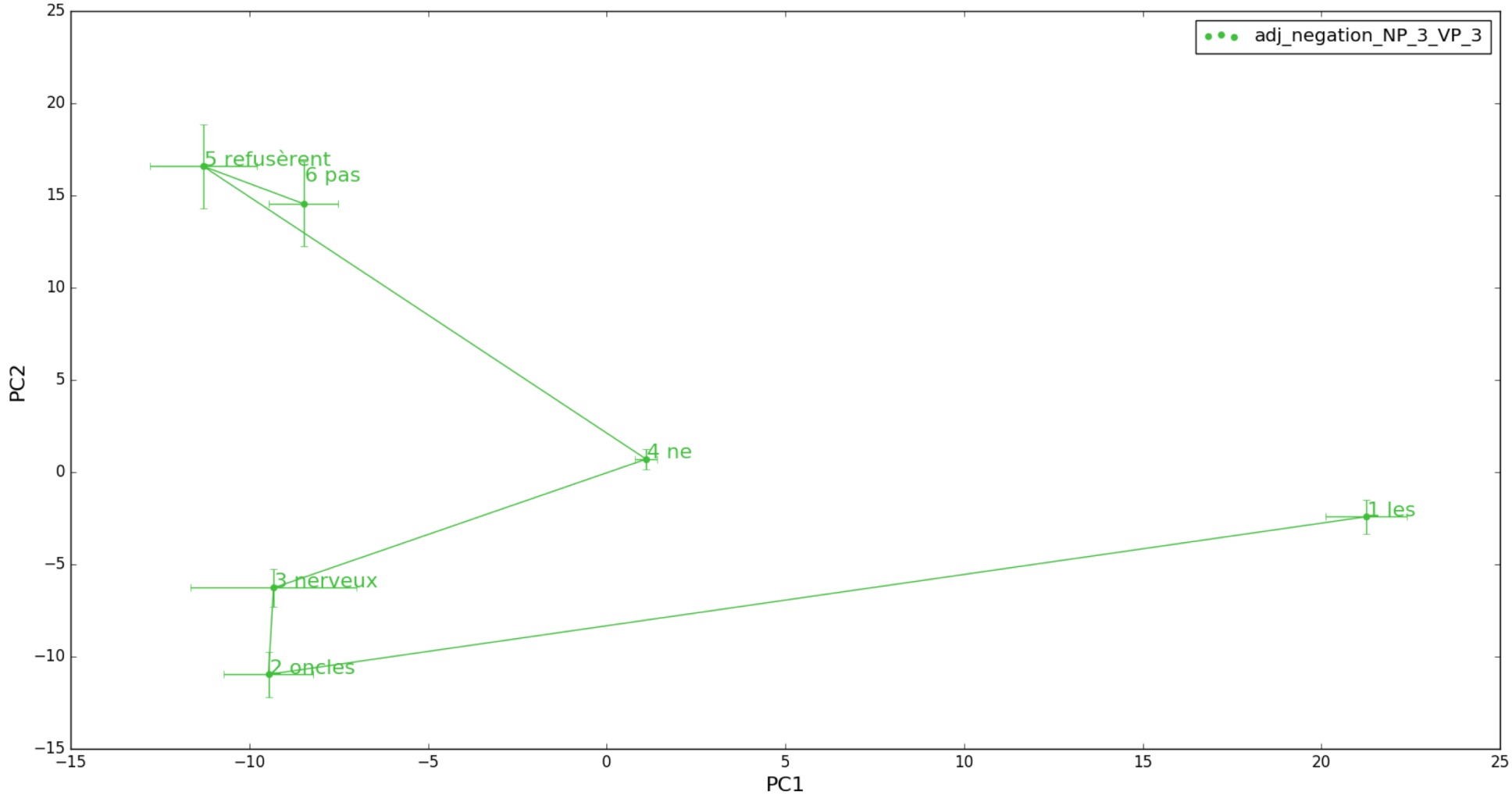
Hidden layer



Long NP, short VP

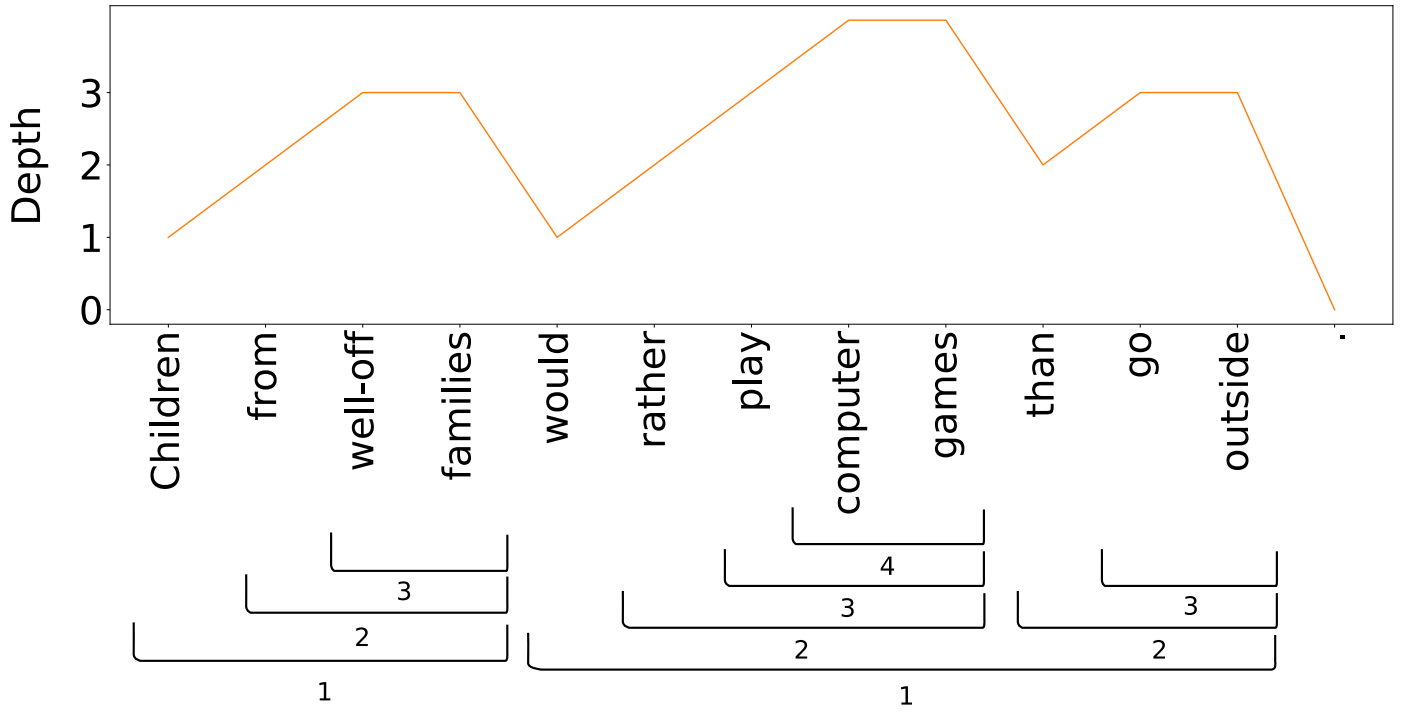


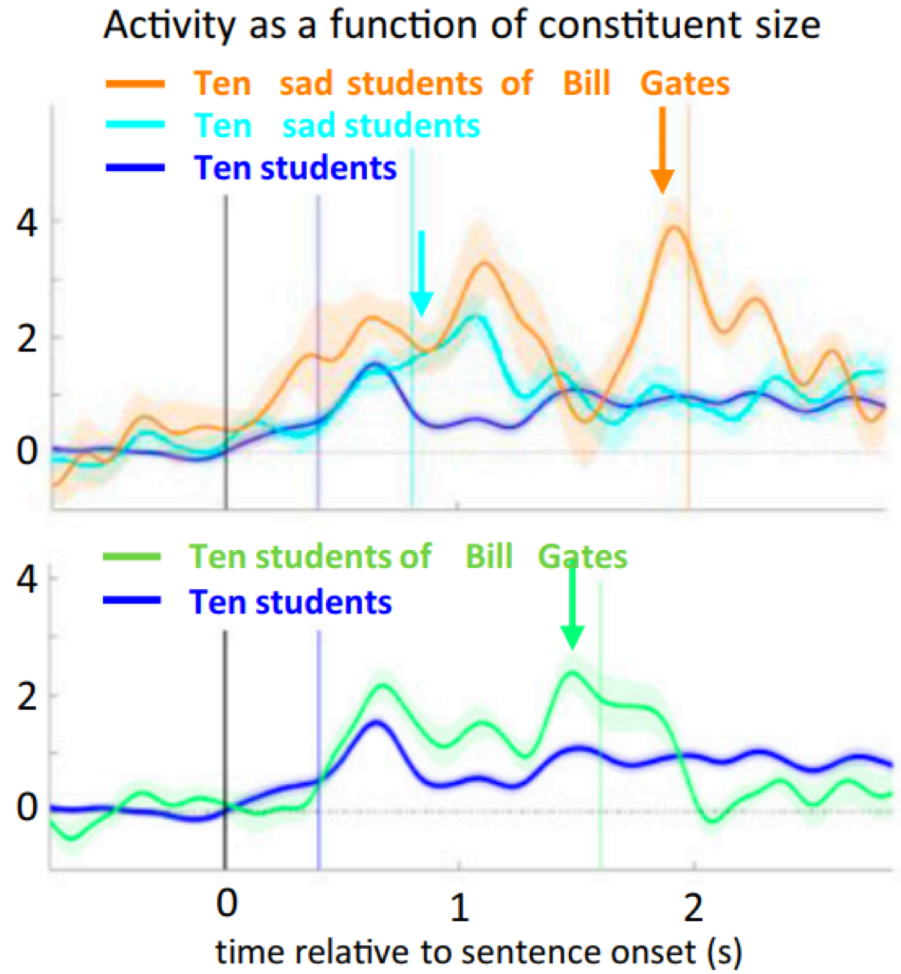
Negation


adj_negation_NP_3_VP_3
5 refusèrent
6 pas
4 ne
1 les
3 nerveux
2 oncles
PC1
PC2
25
20
15
10
5
0
−5
−10
−15
−15
−10
−5
0
5
10
15
20
25

Study 2:
Looking for the footprints of a tree

Looking for the footprints of a tree





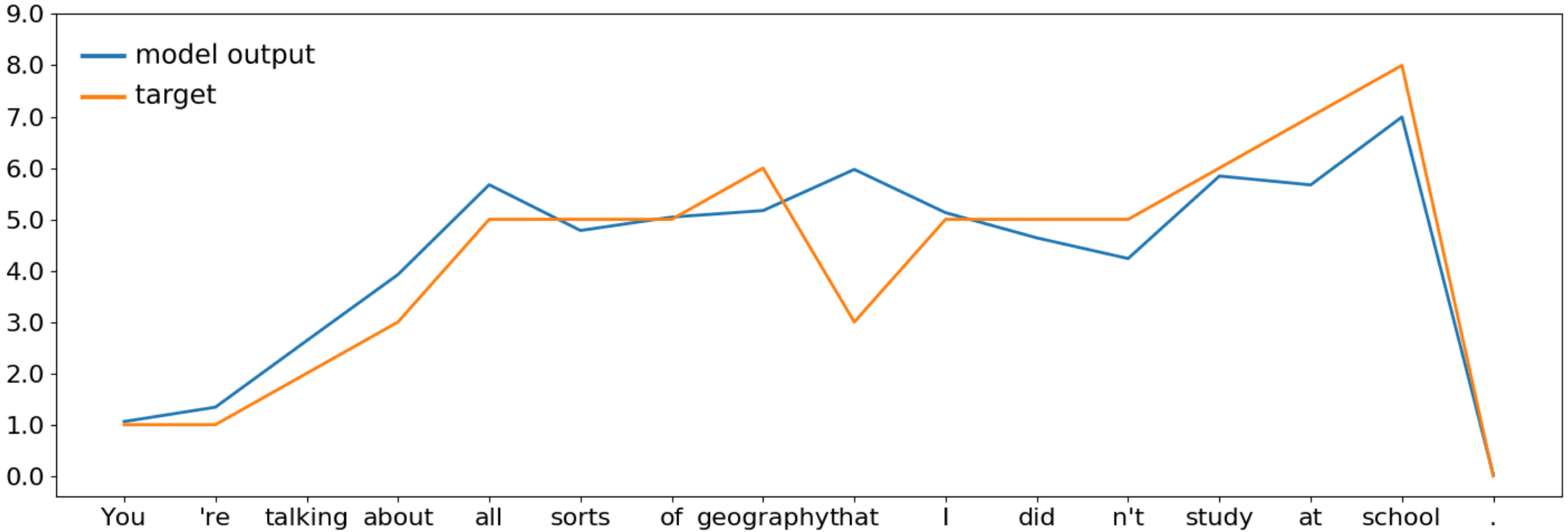
Nelson et al. (2017)

Looking for the footprints of a tree

- Parsed BNC corpus with Stanford PCFG parser
- Extracted 18k sentences where depth was minimally correlated with position
- Pre-trained Gulordava language model:
  - trained on 90M words from Wikipedia
  - 2 layers, 650 units per layer, dropout 0.2
  - vocabulary size: 50k words
- (Lasso) regression to predict depth from hidden representations

Results:

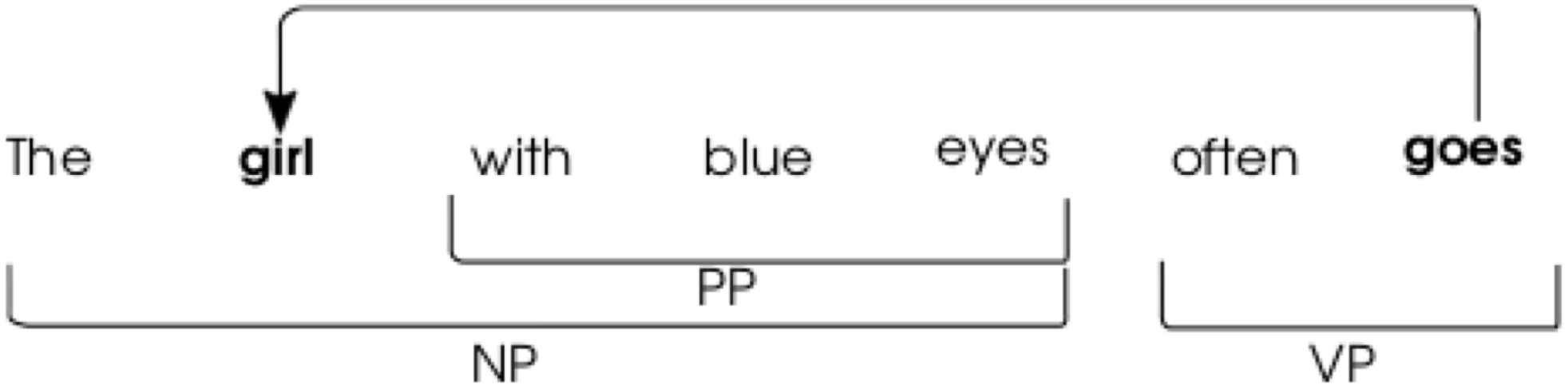
input representation	r^2
word embeddings	0.27
bag of embeddings	0.45
h_t	0.66
C_t	0.70


model output
target
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
You
're
talking
about
all
sorts
of
geographythat
I
did
n't
study
at
school
.

Study 3:
How do LSTMs store long-term syntactic information?

The long-distance agreement task

- Linzen et al. (2016), Gulordava et al. (2018)



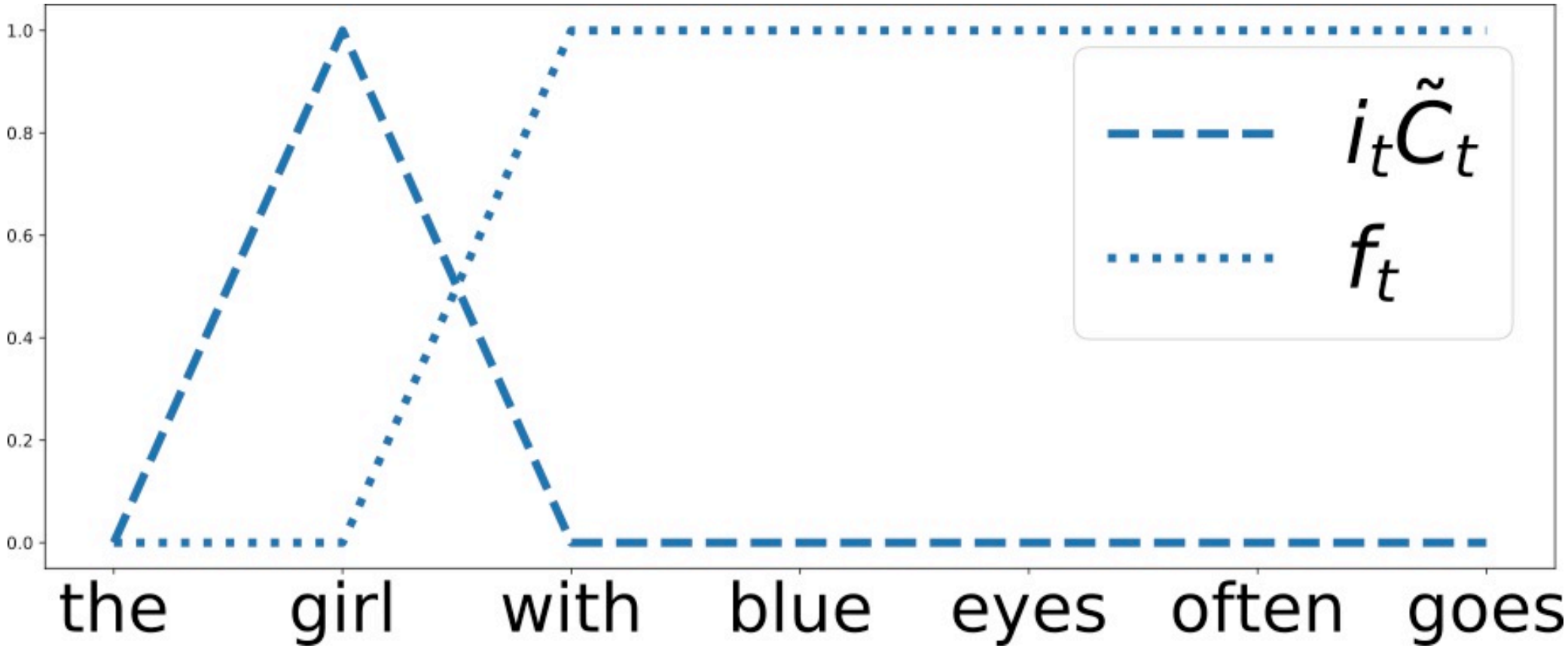
- 3,156 corpus-extracted examples in our version
- How do LSTMs carry morphosyntactic features across the embedded constituents?

LSTM re-recap

- $C_t = f_t C_{t-1} + i_t \hat{C}_t$
- $h_t = o_t \tanh(C_t)$

- $i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$
- $f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$
- $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$
- $\hat{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C)$

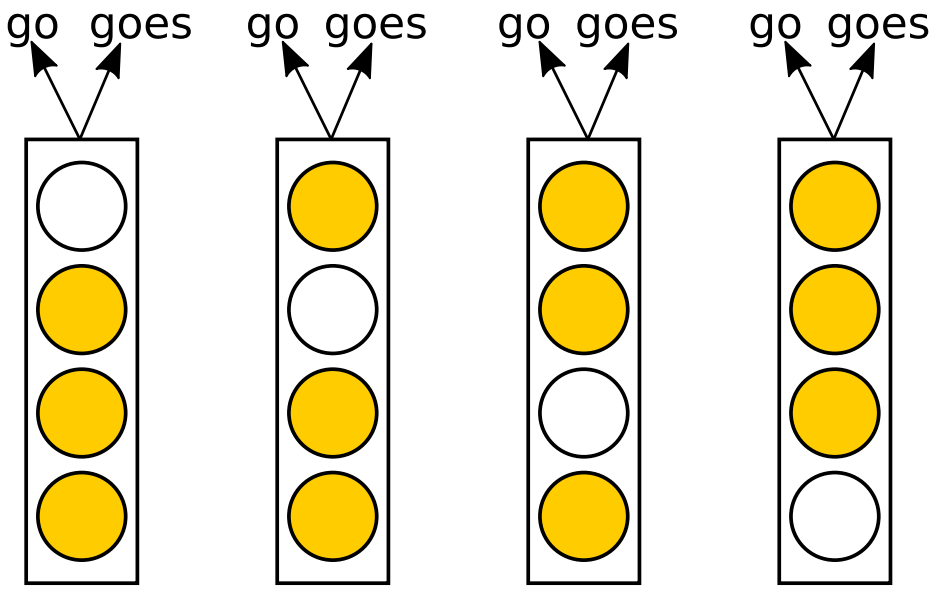
How might LSTMs carry morphosyntactic information through time?



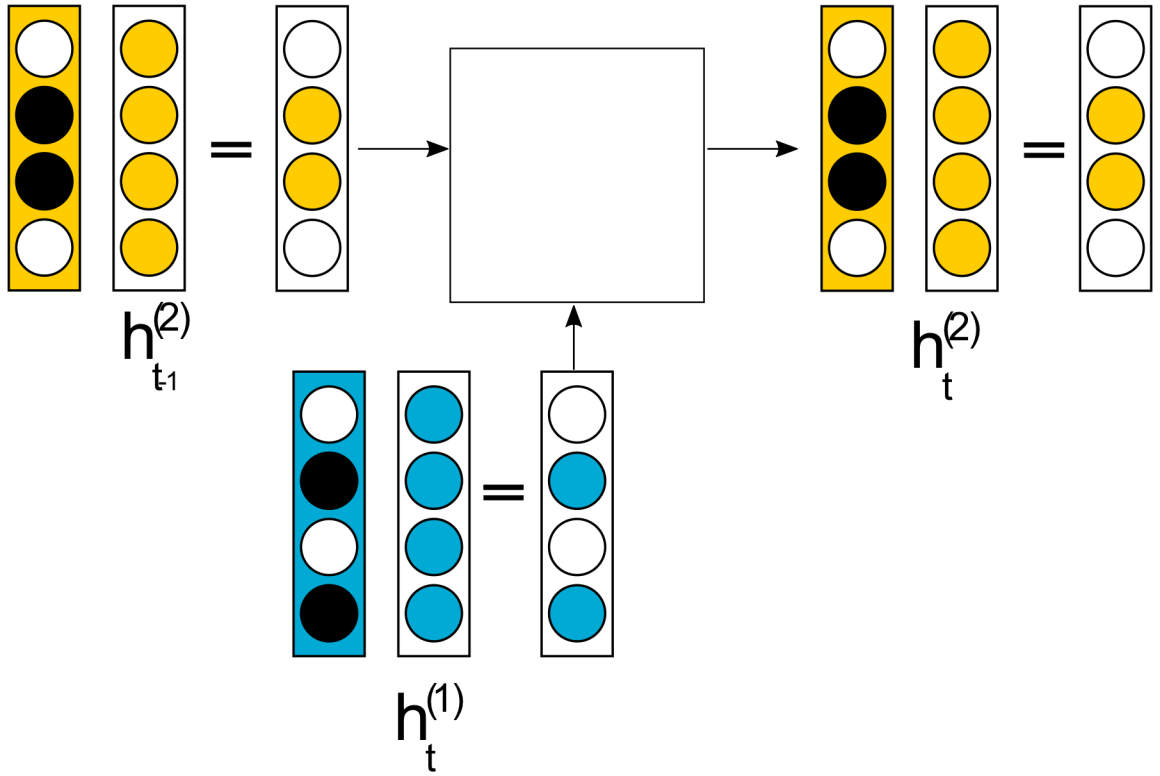
$$C_t = f_t C_{t-1} + i_t \hat{C}_t$$

Searching carry-through cells by ablation

- Look for units that impact performance on the Linzen data-set



single top layer cells

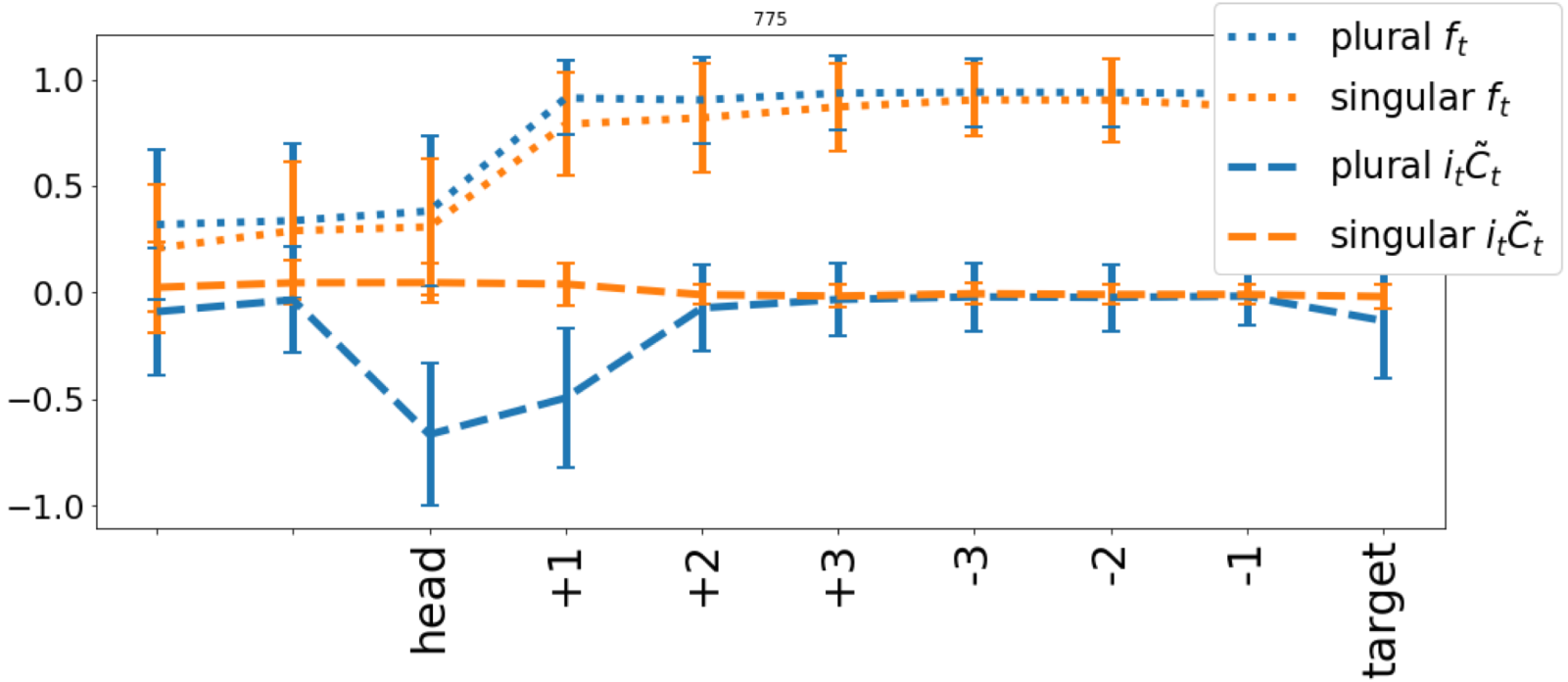


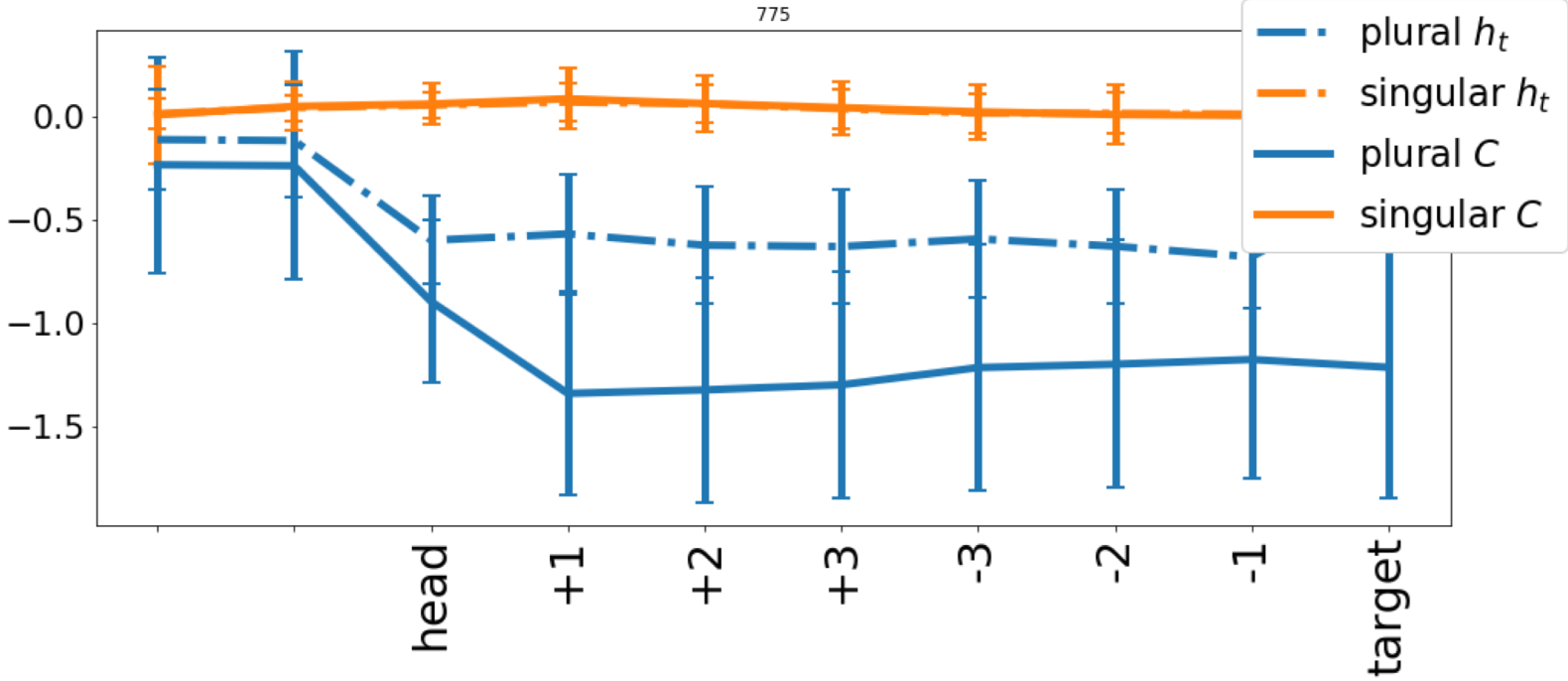
distributed patterns

Accuracy after ablation

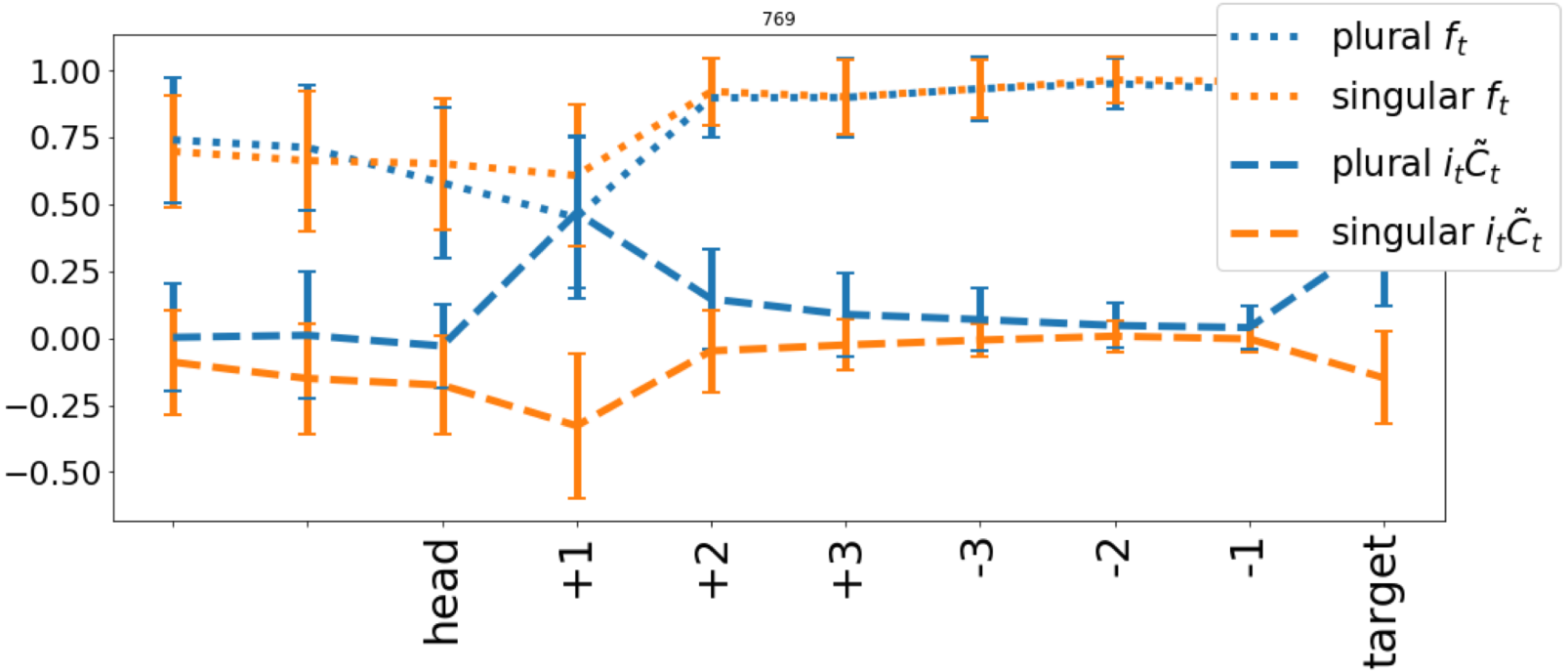
ablated unit(s)	accuracy	impact
none	94%	0%
775	85%	10%
769	90%	4%
775+769	66%	30%
any other unit	93%	<1%
majority baseline	63%	33%

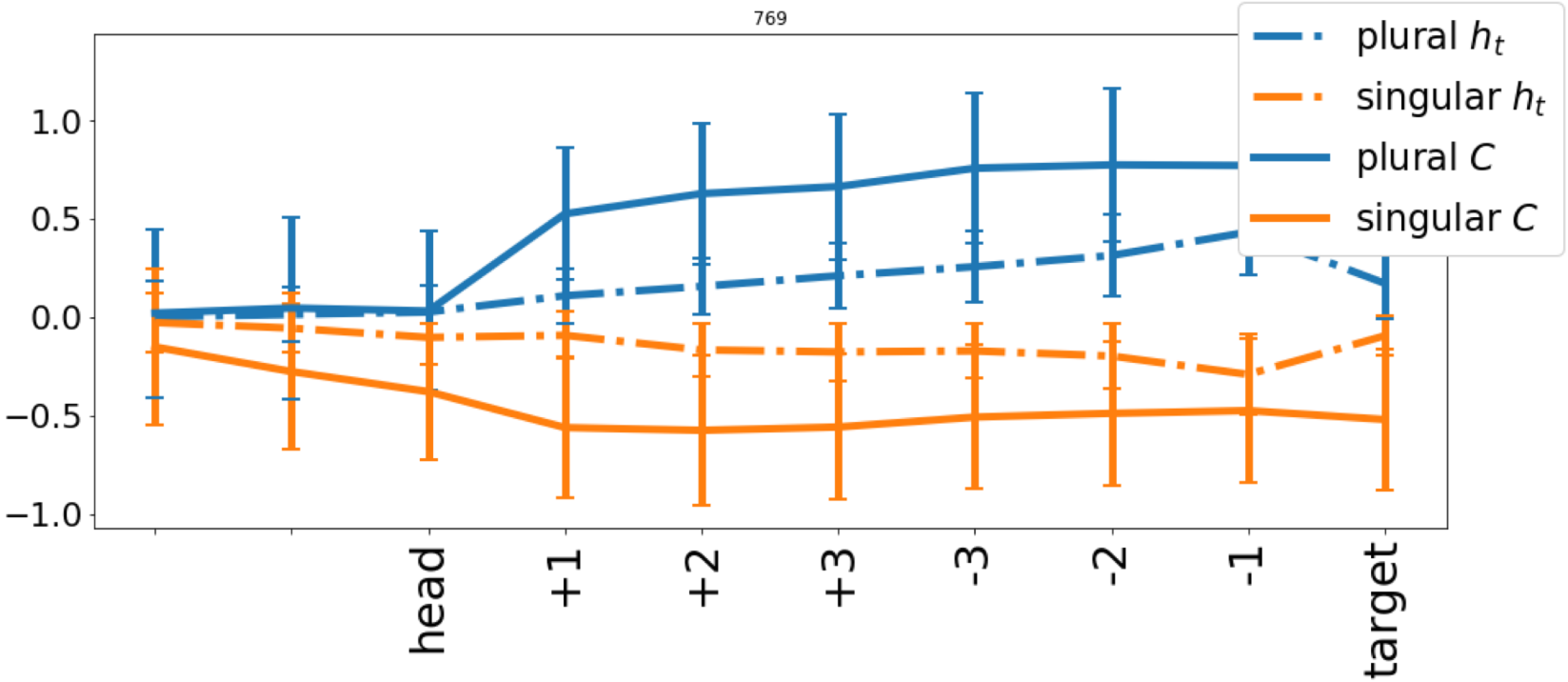
Unit 775


775
plural f_t
singular f_t
plural $i_t\tilde{C}_t$
singular $i_t\tilde{C}_t$
1.0
0.5
0.0
−0.5
−1.0
head
+1
+2
+3
-3
-2
-1
target

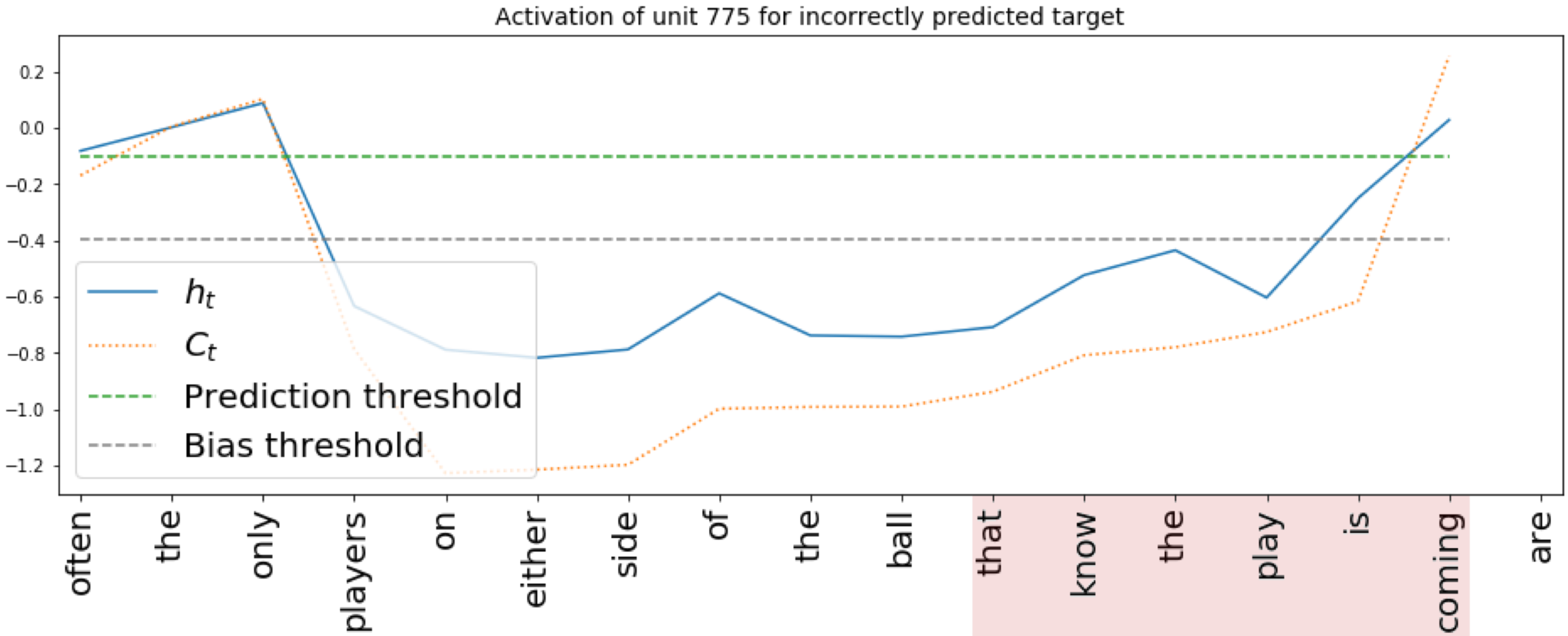

775
plural h_t
singular h_t
plural C
singular C
0.0
−0.5
−1.0
−1.5
head
+1
+2
+3
-3
-2
-1
target

Unit 769

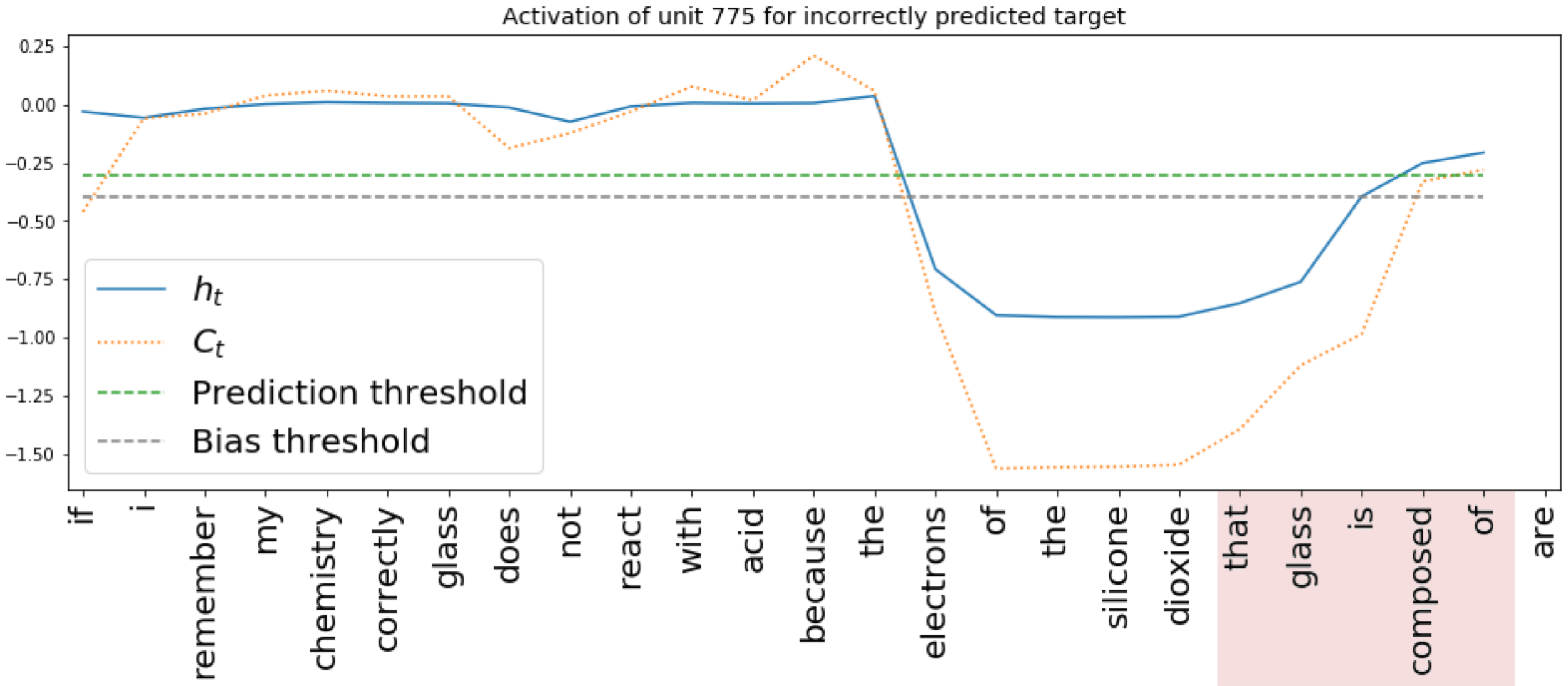

769
plural f_t
singular f_t
plural $i_t\tilde{C}_t$
singular $i_t\tilde{C}_t$
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
head
+1
+2
+3
-3
-2
-1
target


769
plural h_t
singular h_t
plural C
singular C
1.0
0.5
0.0
−0.5
−1.0
head
+1
+2
+3
-3
-2
-1
target

Failure examples



Failure examples



Tracking multiple dependencies: The case of non-subject relatives

- Subject relative:
  - The **boy** [that **likes movies**] **is**

- Non-subject relative:
  - The **boy** [that **girls like**] **is**

- 100 subject relatives, 14 non-subject relatives extracted from Linzen dataset with high-precision, low recall heuristics

Accuracy by sentence type

sentence type	accuracy
all	94%
subject relatives	96%
non-subject relatives	71%
majority baseline	63%

Generating controlled relative clauses
500 examples per structure

- The **father** that **fears** the **girl** definitely **admires**
- The **father** that **fears** the **girls** definitely **admires**
- The **fathers** that **fear** the **girl** definitely **admire**
- The **fathers** that **fear** the **girls** definitely **admire**

- The **father** that the **girl fears** definitely **admires**
- The **father** that the **girls fear** definitely **admires**
- The **fathers** that the **girl fears** definitely **admire**
- The **fathers** that the **girls fear** definitely **admire**

main verb accuracy

relative verb accuracy

subject relatives

The **father(s)** that **fear(s)** the **girl(s)** definitely **admire(s)**

	relative noun			
main noun	*singular*		*plural*	
singular	**80%**	**87%**	**60%**	**87%**
plural	**93%**	**100%**	**99%**	**100%**

object relatives

The **father(s)** that the **girl(s) fear(s)** definitely **admire(s)**

	relative noun			
main noun	*singular*		*plural*	
singular	**46%**	**82%**	**27%**	**100%**
plural	**89%**	**50%**	**92%**	**100%**

Inconclusions

- Language-modeling-trained LSTMs discover quite a bit about hierarchical structures in language
- Long-distance feature percolation is performed in a localist fashion by a small number of grandmother cells
  - How about structure tracking?
- Plenty of unsolved puzzles
  - Most interesting to me: does the LSTM keep track of multiple agreement patterns? if so, how?
- English sucks for this sort of studies