Weakly-Supervised Bayesian Learning of a CCG Supertagger

Dan Garrette, Chris Dyer, Jason Baldridge, Noah A. Smith

Type-Level Supervision

Type-Level Supervision

- Unannotated text
- Incomplete tag dictionary: word ↦ {tags}

Type-Level Supervision

Used for POS tagging for 20+ years

[Kupiec, 1992]
[Merialdo, 1994]

Type-Level Supervision

Good POS tagger performance
even with low supervision

[Das & Petrov 2011]
[Garrette & Baldridge 2013]
[Garrette et al. 2013]

Combinatory Categorial Grammar (CCG)

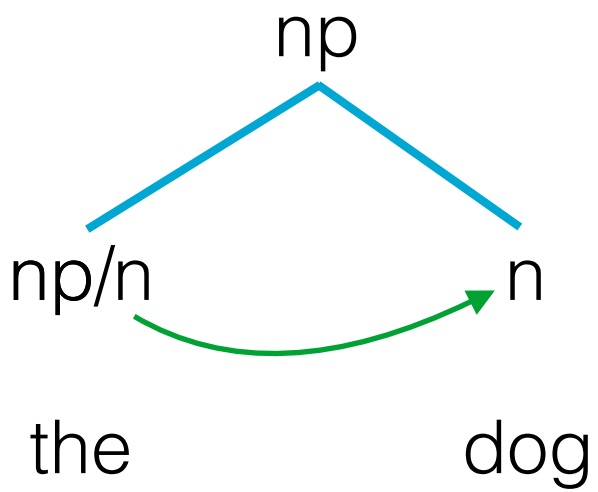
CCG

Every word token is associated with a **category**

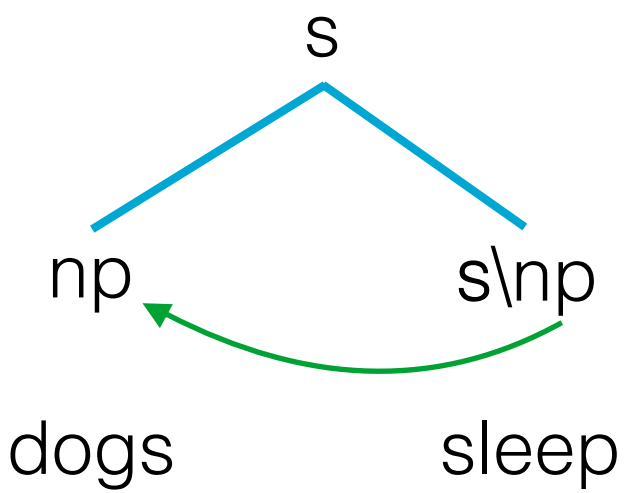
Categories **combine** to categories of constituents

[Steedman, 2000]
[Steedman and Baldridge, 2011]

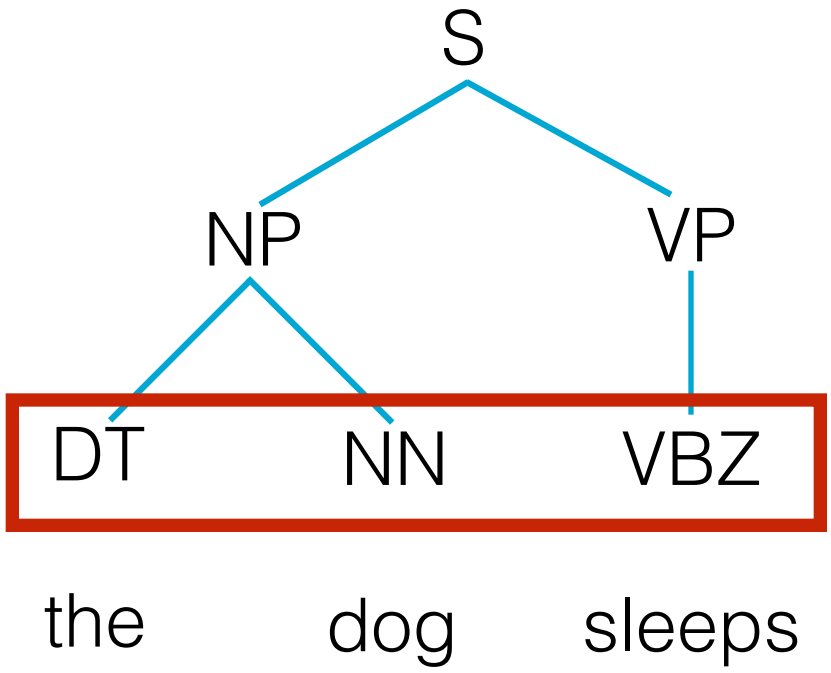
CCG

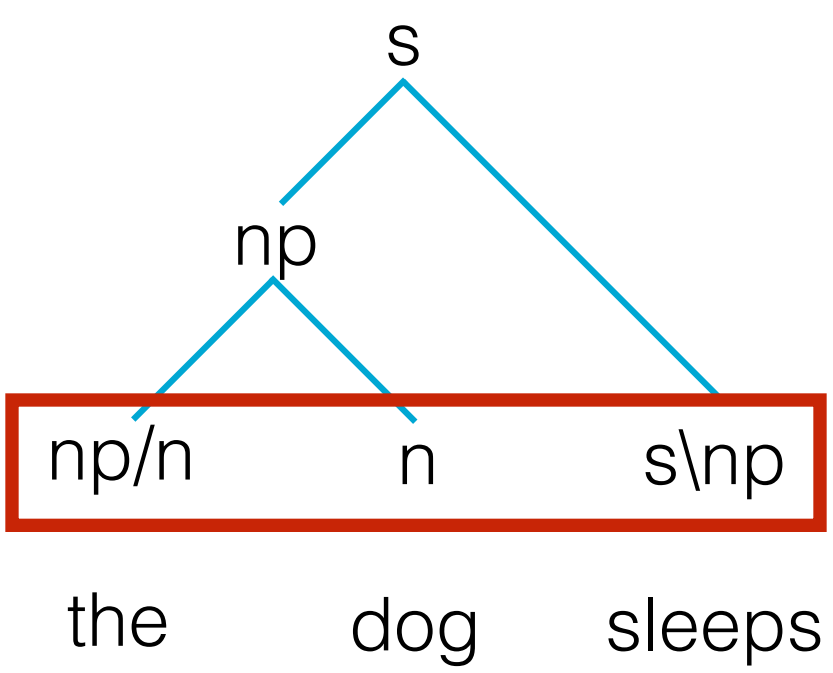


CCG



POS vs. Supertags





Supertagging

Type-supervised learning for supertagging
is much more difficult than for POS

Penn Treebank POS

48 tags

CCGBank Supertags

1,239 tags

CCG

The grammar formalism *itself*
can be used to guide learning

CCG Supertagging

CCG Supertagging

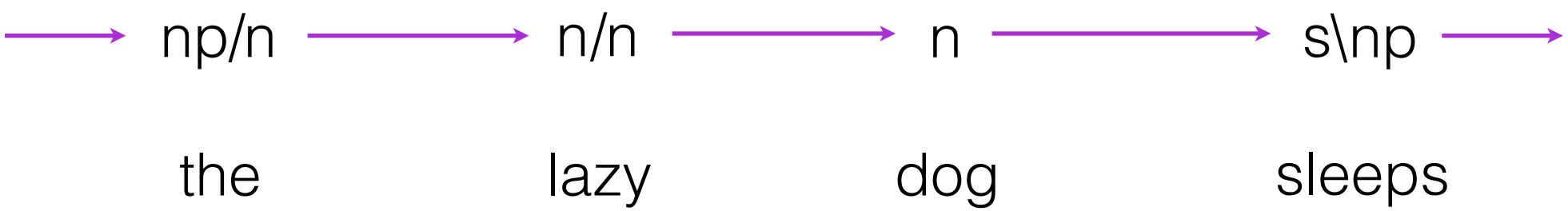
- Sequence tagging problem, like POS-tagging
- Building block for grammatical parsing

Supertagging

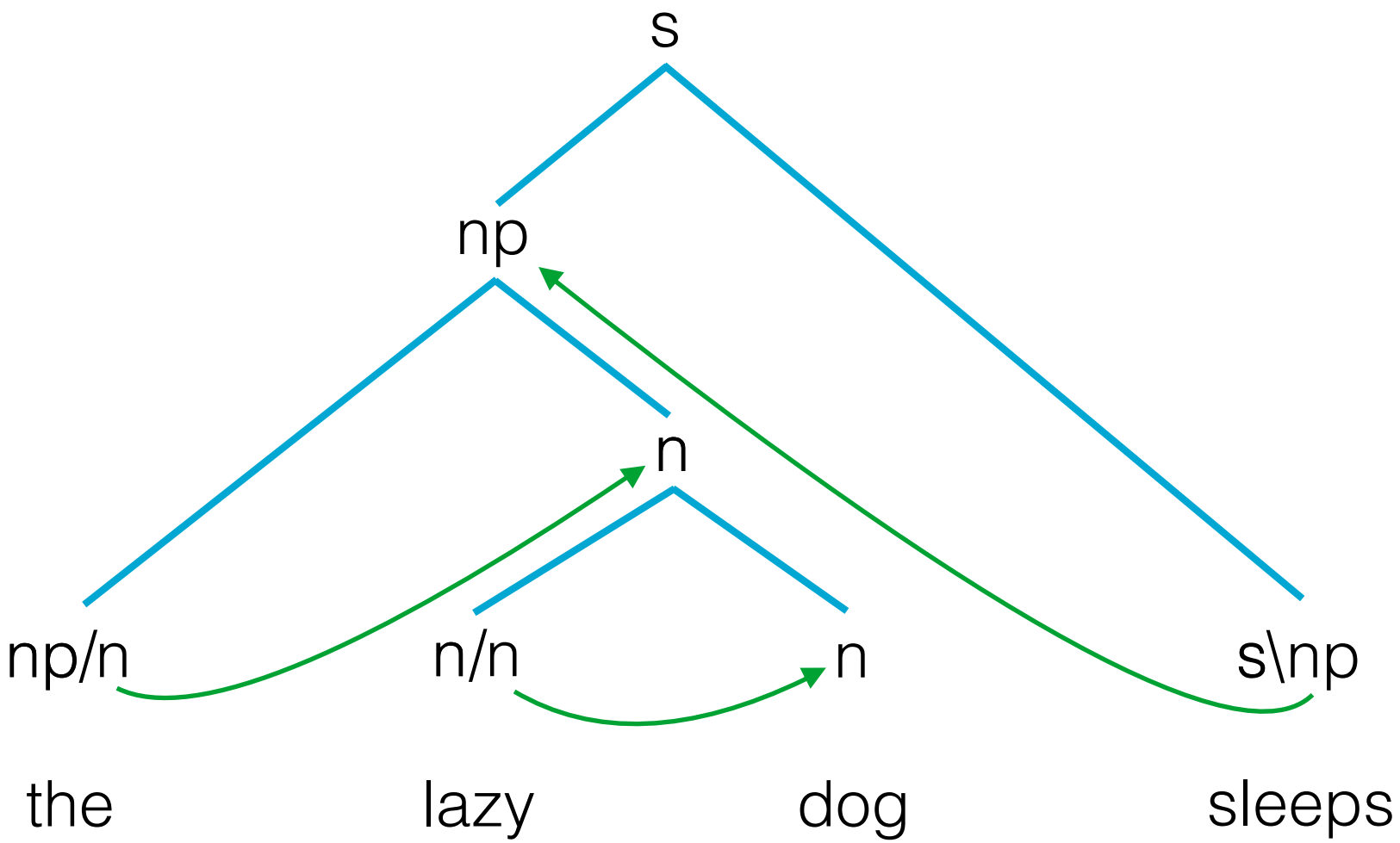
“almost parsing”

[Bangalore and Joshi 1999]

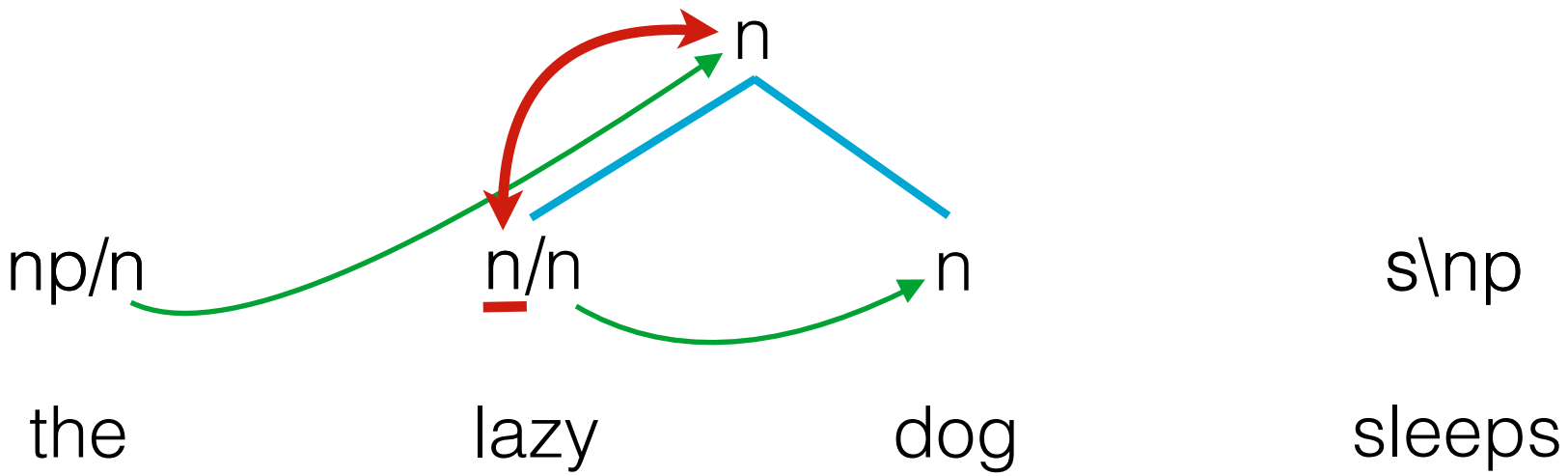
Why Supertagging?



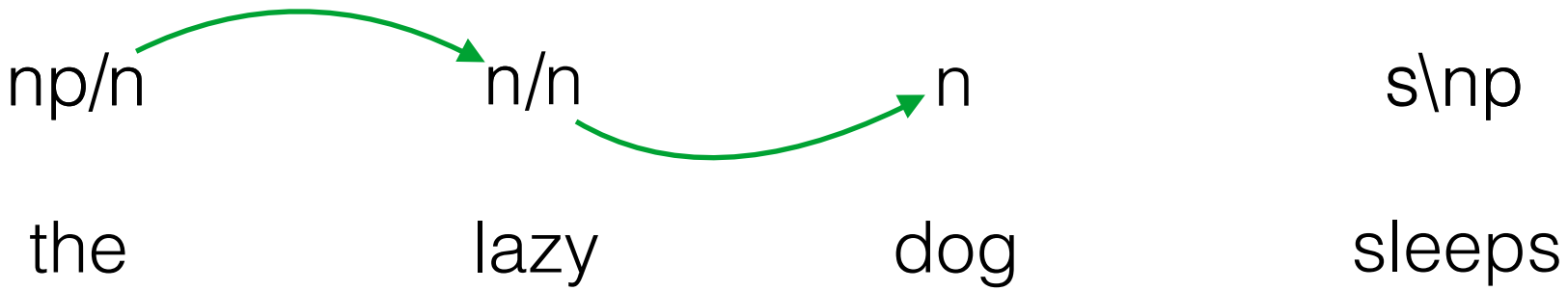
Why Supertagging?



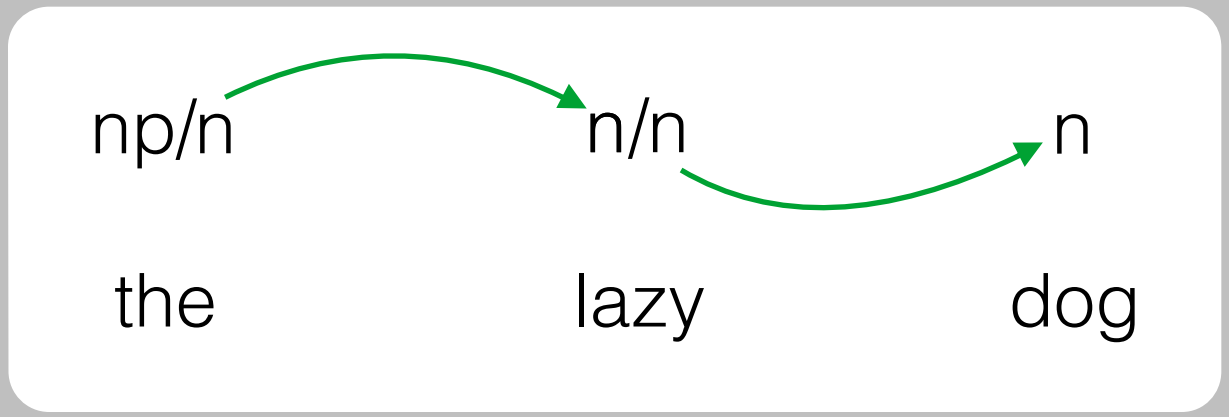
CCG Supertagging



CCG Supertagging

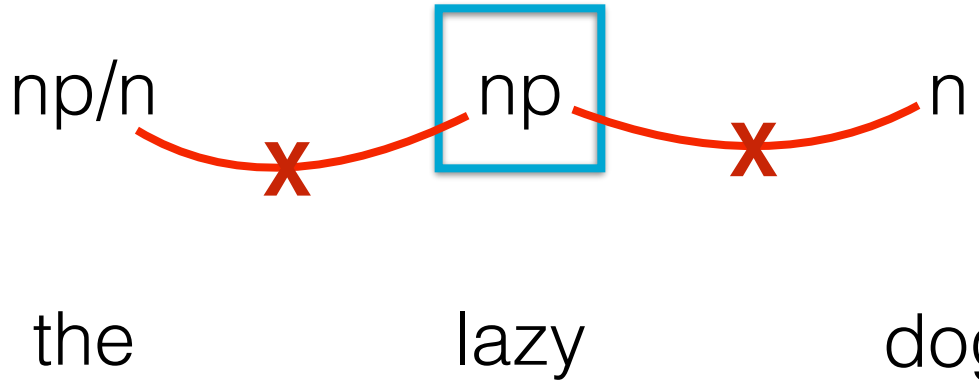


CCG Supertagging



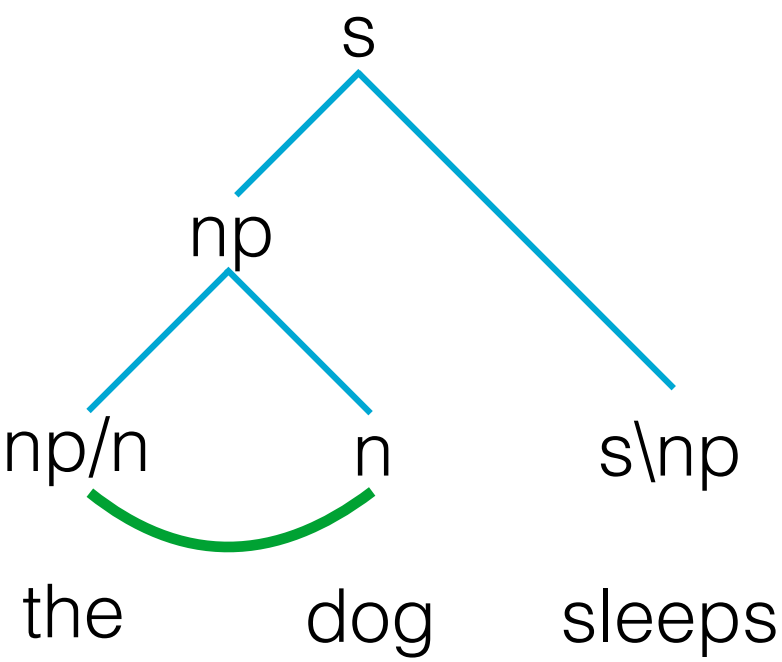
CCG Supertagging

Principle #1

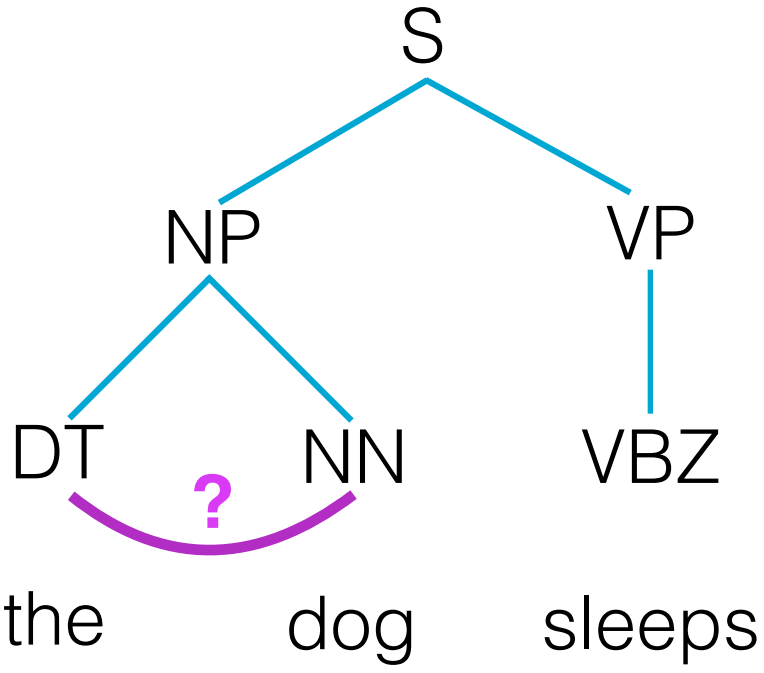


Prefer Connections

Supertags vs. POS

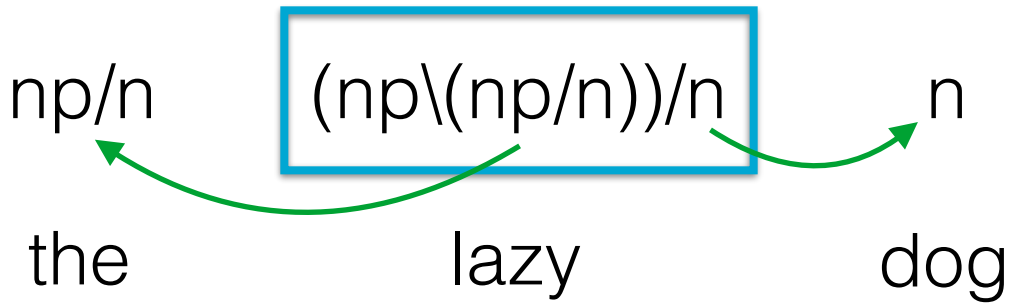


universal, intrinsic
grammar properties



all relationships
must be learned

Principle #2



Prefer Simplicity

Prefer Simplicity

buy := $(s_b\backslash np)/np$ appears **342** times in CCGbank

e.g. “Opponents don't **buy** such arguments.”

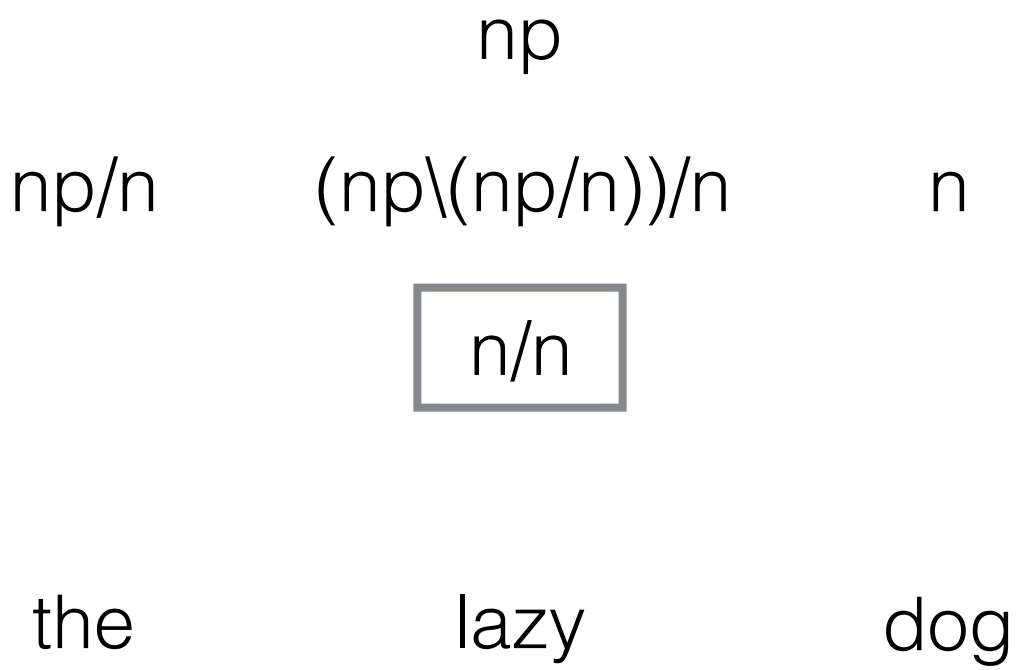
buy := $(((s_b\backslash np)/\mathbf{pp})/\mathbf{pp})/np$ appears **once**

“Tele-Communications agreed to **buy** half of Showtime Networks from Viacom (pp) for $ 225 million. (pp)”

Weighted Tag Grammar

a	$\longrightarrow$	{s, np, n,…}	$p_{atom}(a) \times p_{term}$
A	$\longrightarrow$	B / B	$\overline{p_{term}} \times p_{fwd} \times p_{mod}$
A	$\longrightarrow$	B / C	$\overline{p_{term}} \times p_{fwd} \times \overline{p_{mod}}$
A	$\longrightarrow$	B \ B	$\overline{p_{term}} \times \overline{p_{fwd}} \times p_{mod}$
A	$\longrightarrow$	B \ C	$\overline{p_{term}} \times \overline{p_{fwd}} \times \overline{p_{mod}}$

CCG Supertagging



HMM Transition Prior

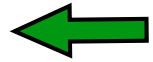
$$P(\mathbf{t} \rightarrow \mathbf{u}) = \lambda \cdot \underbrace{P(\mathbf{u})}_{\text{simple is good}} + (1-\lambda) \cdot \underbrace{P(\mathbf{t} \rightarrow \mathbf{u})}_{\text{connecting is good}}$$

Type-Supervised Learning

unlabeled corpus

tag dictionary

same as
POS tagging

universal properties of the CCG formalism 

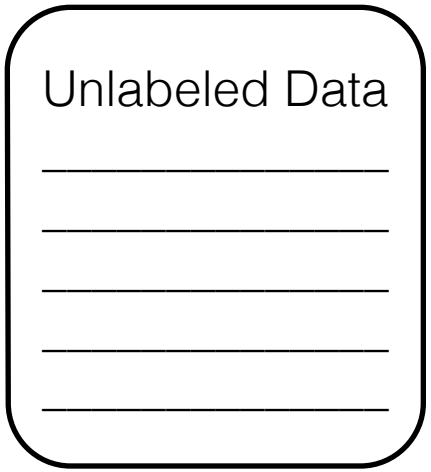
Training

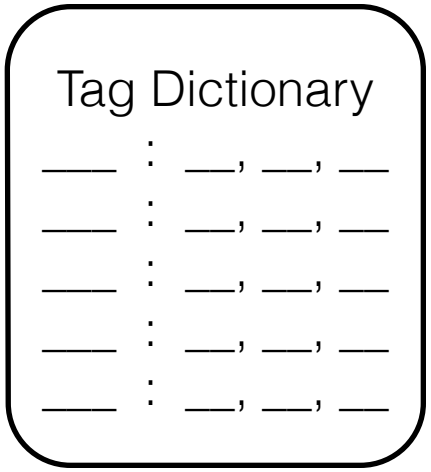
Posterior Inference

Forward-Filter Backward-Sample (FFBS)

[Carter and Kohn, 1996]

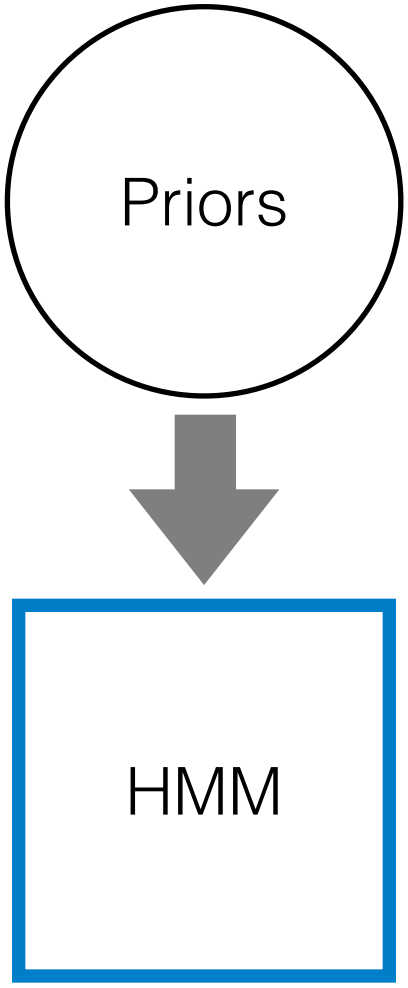
Posterior Inference





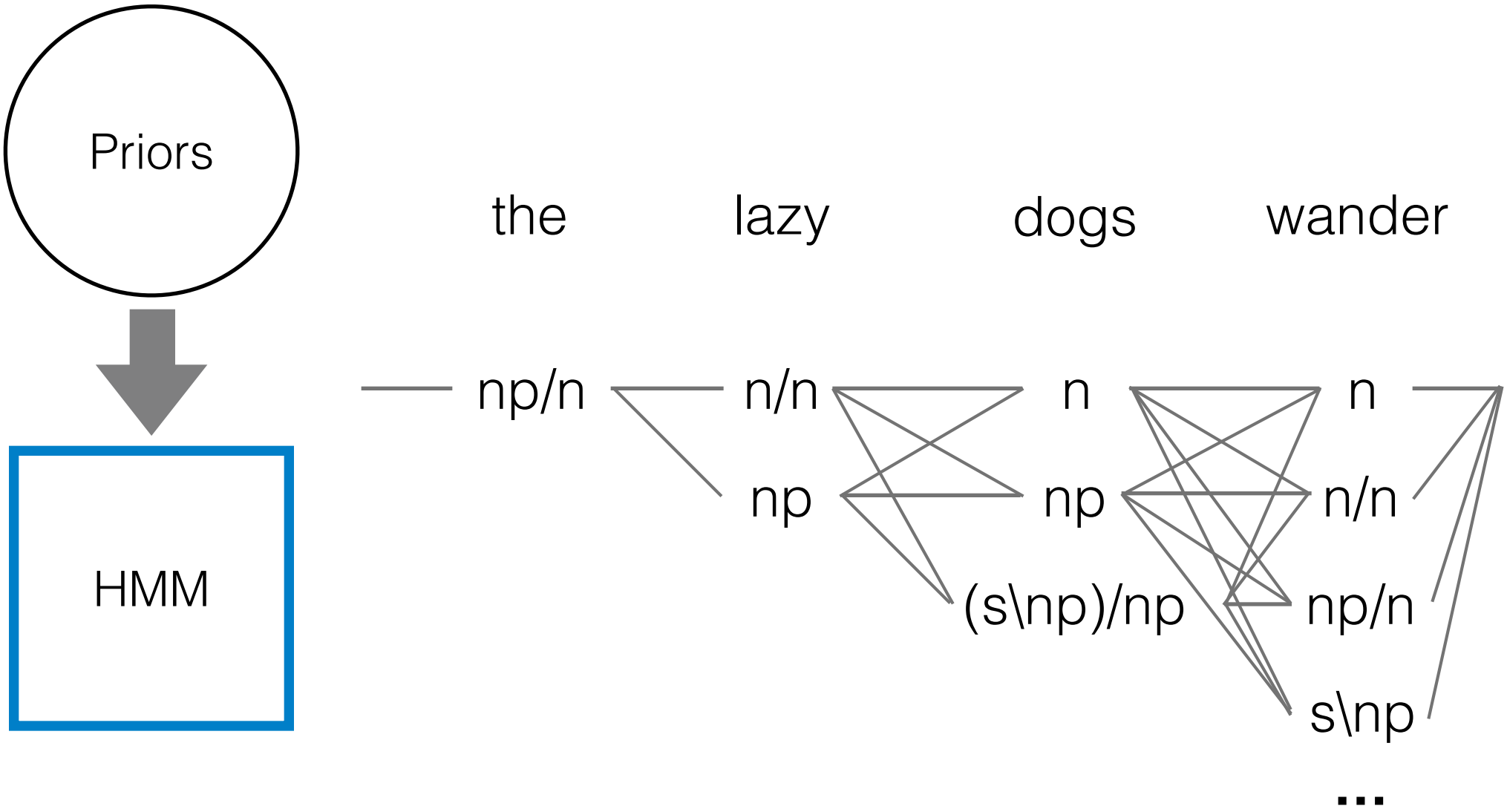
the	lazy	dogs	wander
np/n	n/n	n	n
	np	np	n/n
		(s\np)/np	np/n
			s\np
			...

Posterior Inference

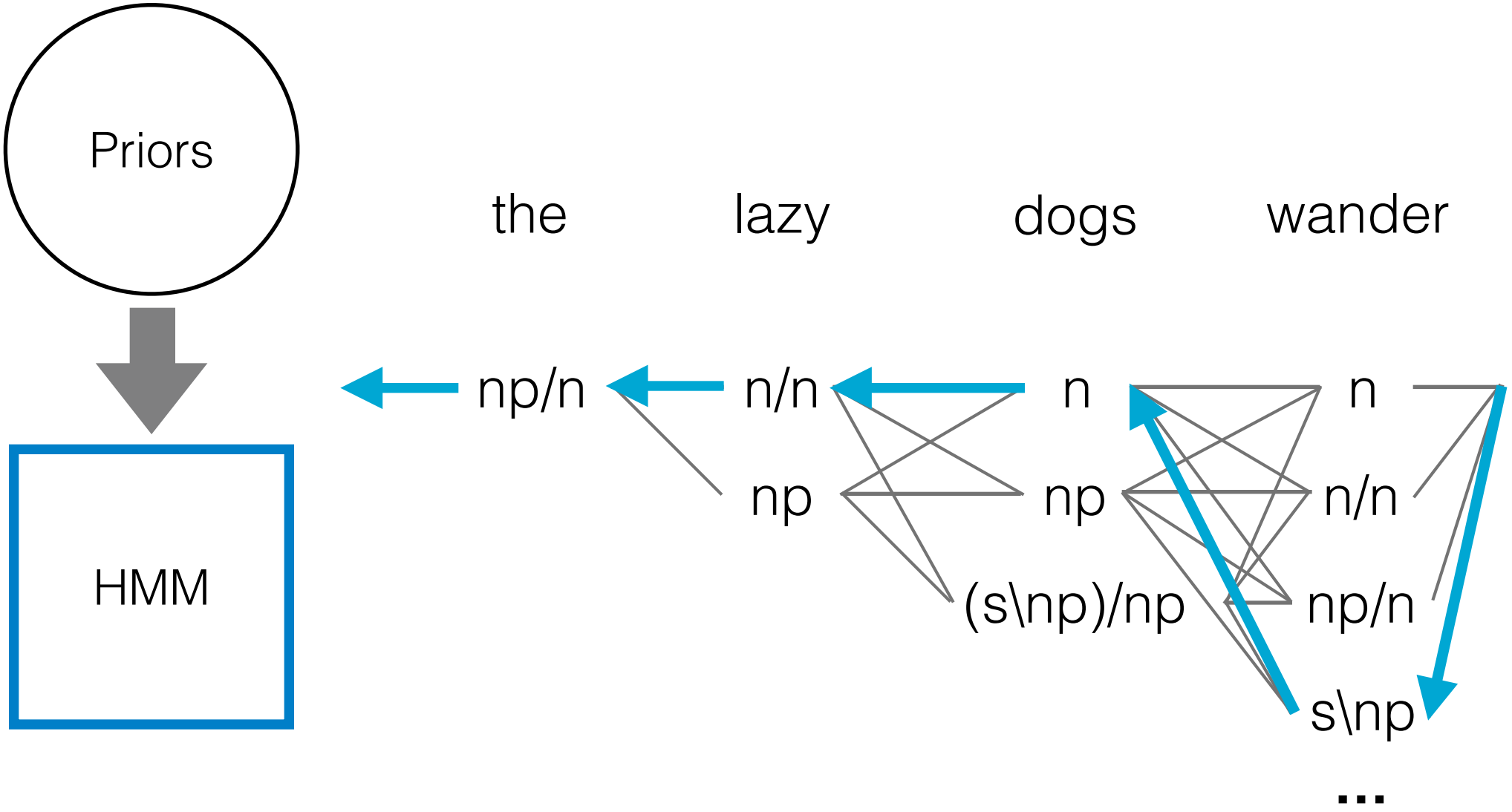


the	lazy	dogs	wander
np/n	n/n	n	n
	np	np	n/n
		(s\np)/np	np/n
			s\np
			...

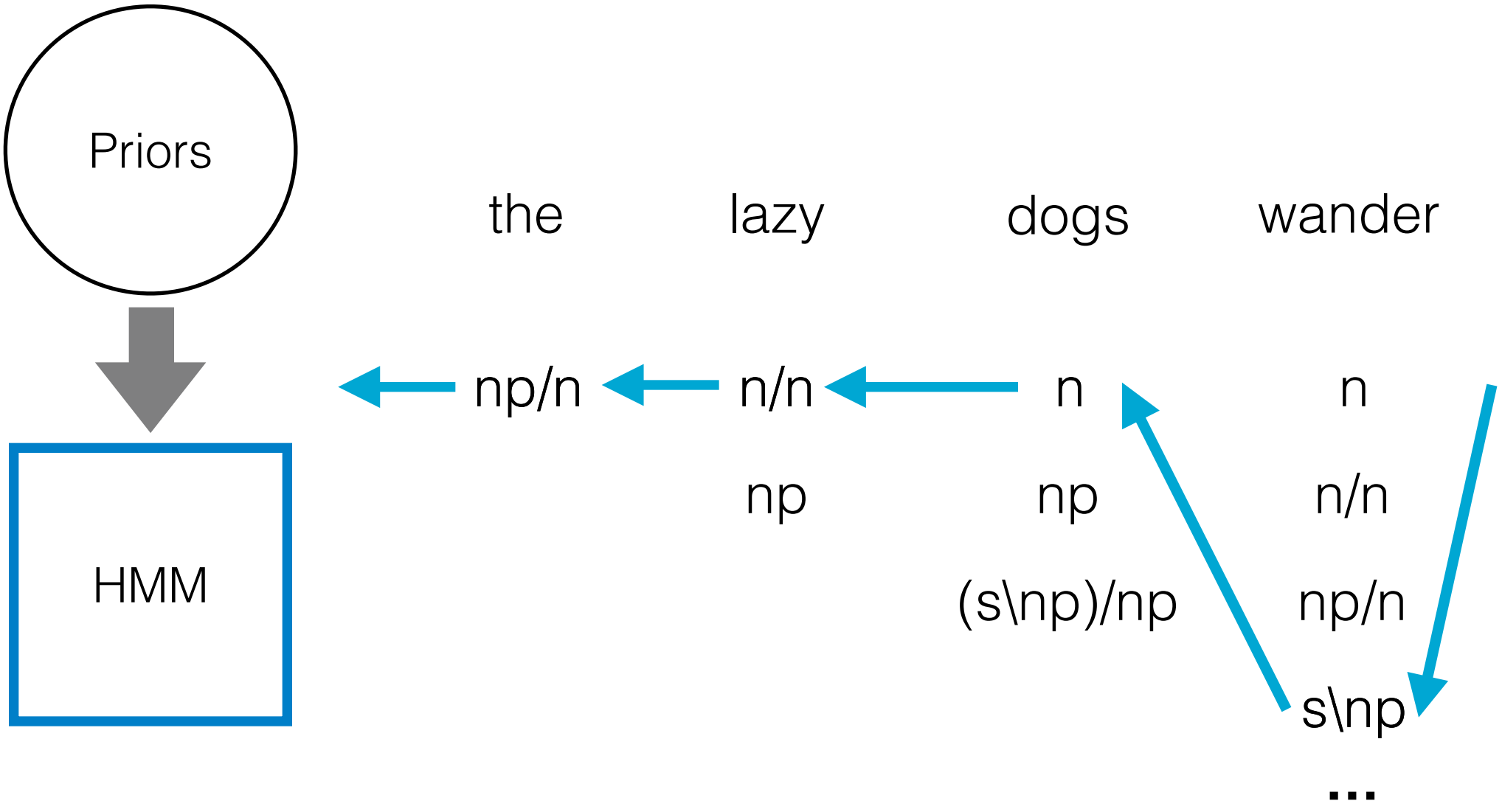
Posterior Inference



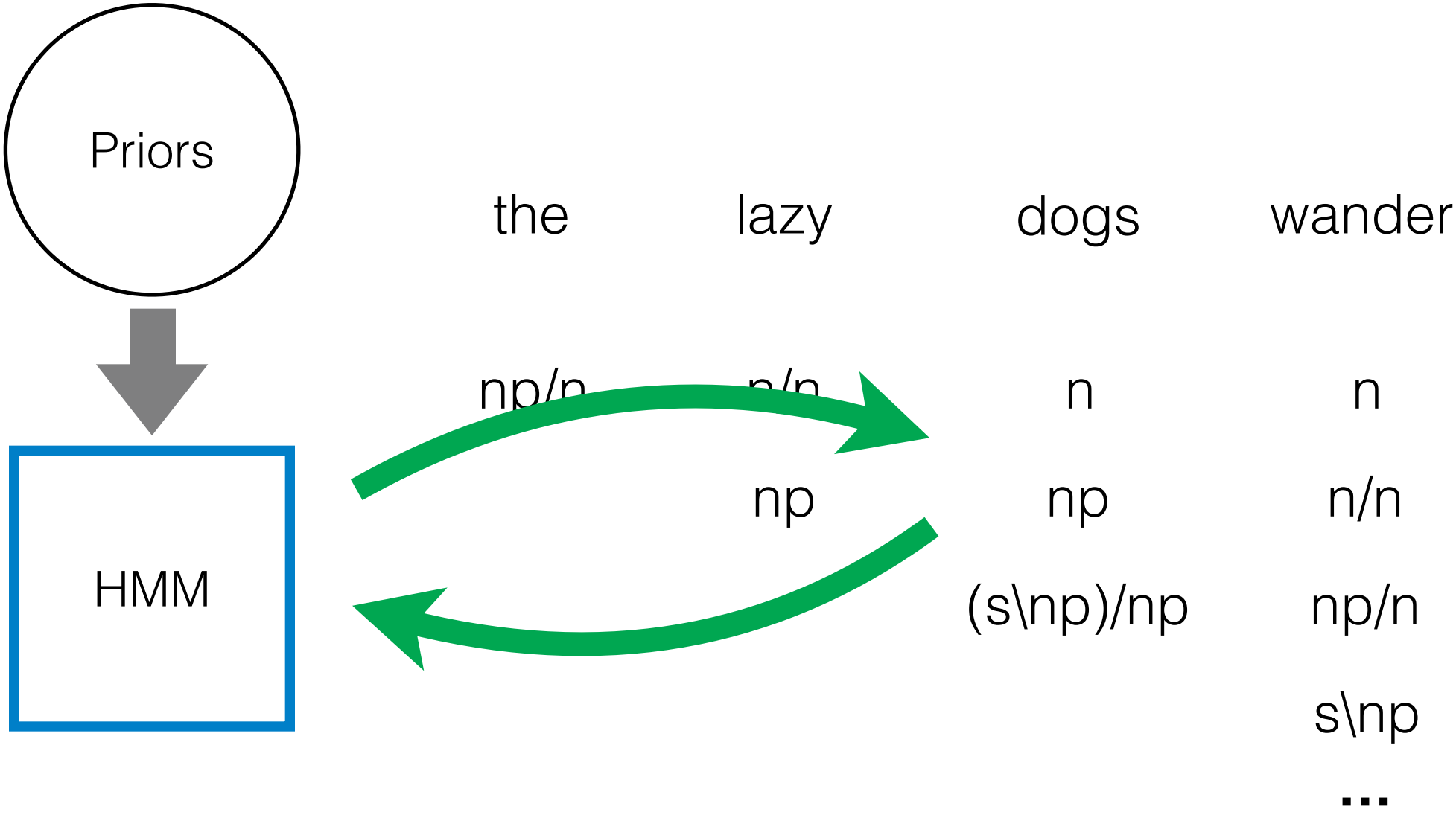
Posterior Inference



Posterior Inference



Posterior Inference



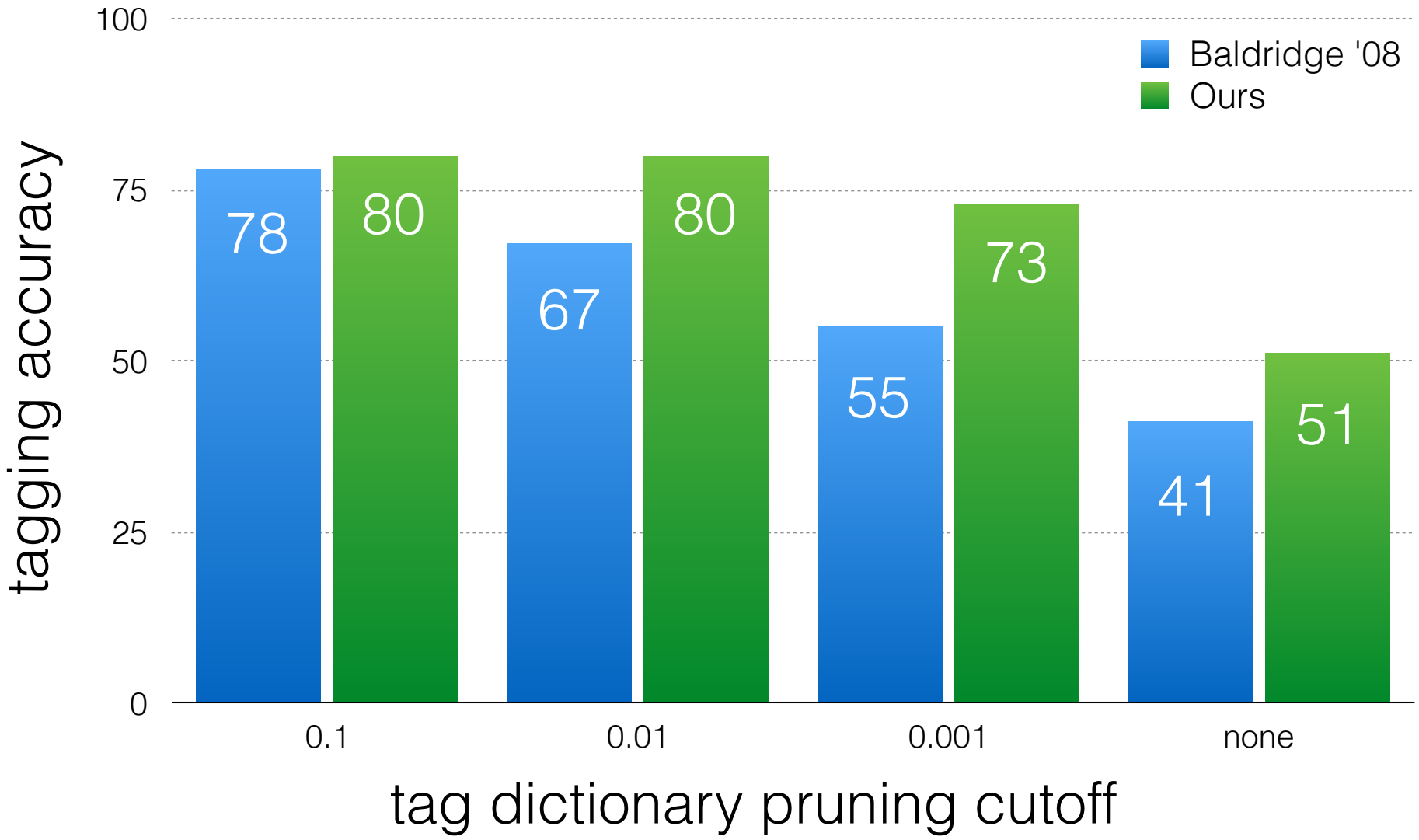
Experiments

Baldridge 2008

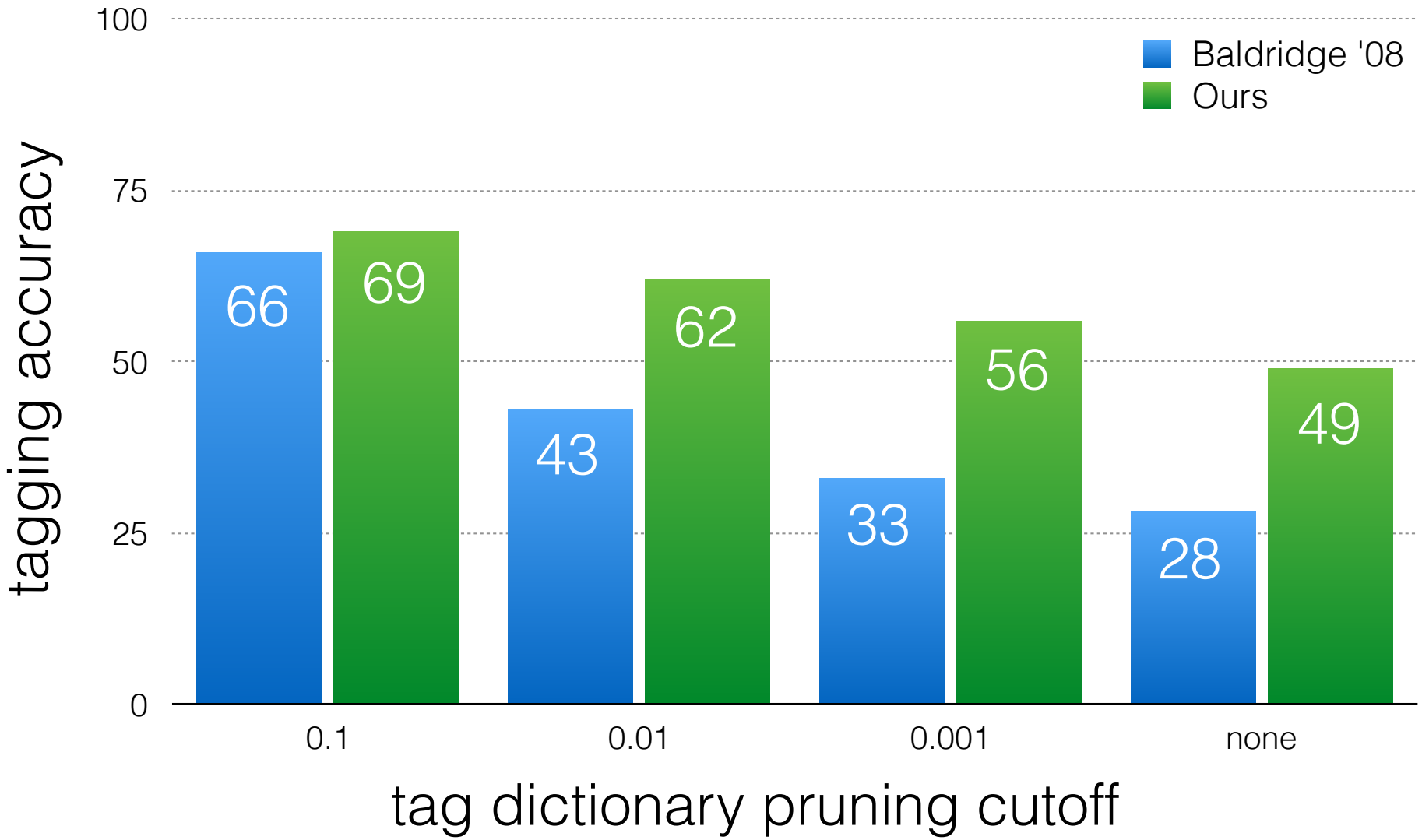
Use universal properties of CCG to initialize EM

- Simpler definition of category complexity
- No corpus-specific information

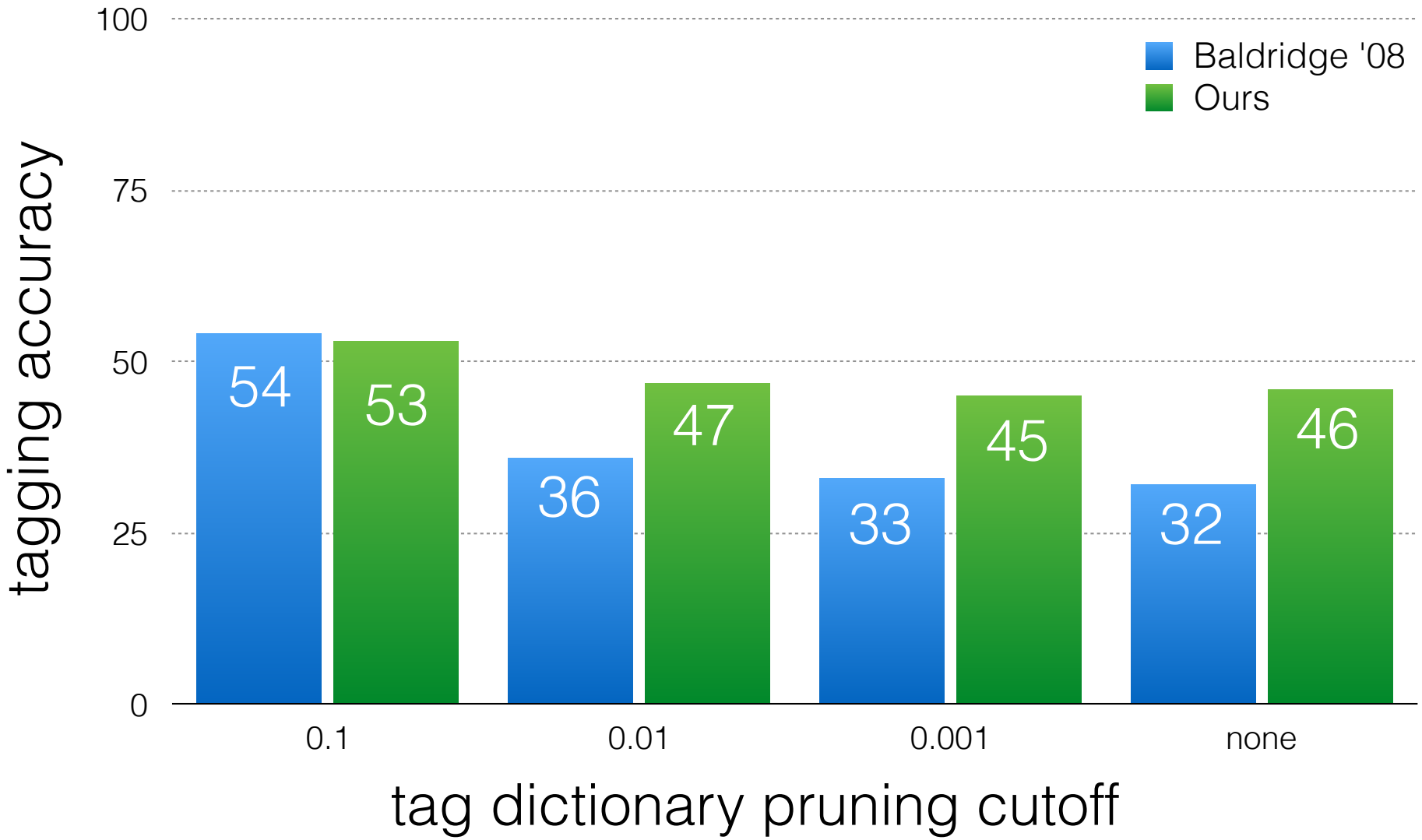
English Supertagging



Chinese Supertagging



Italian Supertagging



Code Available

GitHub repository
linked from my website

Conclusion

Combining annotation exploitation with
universal grammatical knowledge
yields good models
from weak supervision