HDP-CCG

Yonatan Bisk

Outline

- Grammar Induction
- Combinatory Categorial Grammar
- Dirichlet Processes
- HDP-CCG

Grammar Induction

Corpus

I ate cookies

she drank juice

she ate quickly

I ate chocolate cake

Grammar

Adjectives before Nouns

Subject Verb Object

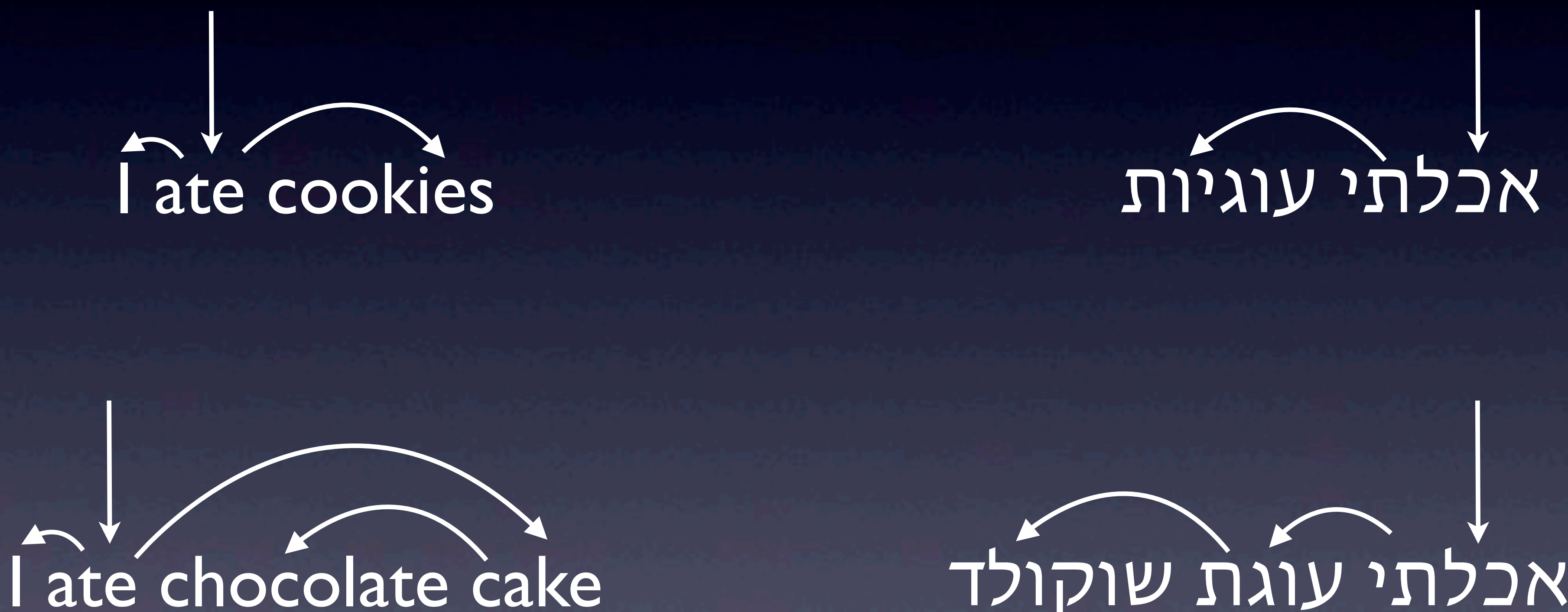
Adverbs after verbs

. . .

Grammar Induction

Corpus	Grammar
אכלתי עוגיות	?
היא שתתה מיץ	?
היא אכלה במהירות	?
אכלתי עוגת שוקולד	…

Dependencies



Questions

- How much initial knowledge does the system need?
- How much information exists in the text
- What’s the space of grammatical rules?

Outline

- Grammar Induction
- Combinatory Categorial Grammar
- Dirichlet Processes
- HDP-CCG

Words as Functions

+

=

chocolate	cookie	chocolate(cookie)
N/N	N	N

Words as functions

 + eats +  = 

Obama | Ice cream | eats(Obama, Ice Cream)

N (S\N)/N N ⟶ S

CCG

A simple set of rules for combining grammatical structures

Intransitive S\N	am		ate	provides
Transitive (S\N)/N	am	threw	ate	provides
Ditransitive ((S\N)/N)/N		threw		provides

Word Probability

p(She | N)   p(ate | S\N)

$$\begin{array}{ccc} N & \quad & S\backslash N \\ \downarrow & & \downarrow \\ \text{She} & & \text{ate} \end{array}$$

Word Probability

p(ate | (S\N)/N)

p(She | N) ~~p(ate | S\N)~~ p(cake | N)

(S\N)/N

N ~~S\N~~ N

$\downarrow$ $\downarrow$ $\downarrow$

She ate cake

Distributions

How many words per category? **Infinite?**

p(She | N) p(cake | N)

How many categories produce the same word?

Infinite?

p(ate | (S\N)/N) p(ate | S\N)

Two Problems

How do you deal with an infinite lexicon?

p(She | N) p(cake | N) Dirichlet Process

Can you share knowledge between distributions?

p(ate | (S\N)/N) p(ate | S\N) Hierarchical Dirichlet Process

Outline

- Grammar Induction
- Combinatory Categorial Grammar
- Dirichlet Processes
- HDP-CCG

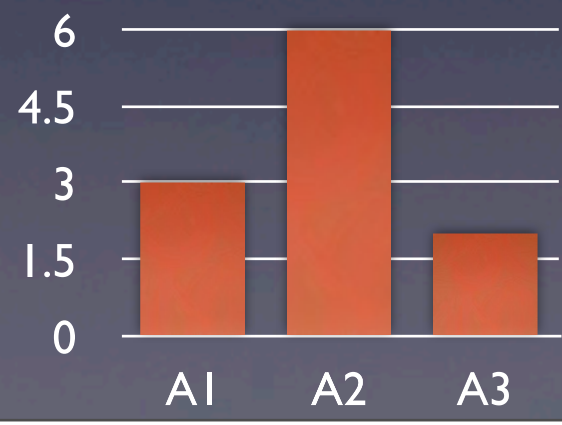
Dirichlet Distribution

Boxes	Finite measure
K = 3	$\alpha([x,y]) = y - x$

1 2 3 4 5 6 7 8 9 10 11 12

$\alpha(A_1) = 3$ $\alpha(A_2) = 6$ $\alpha(A_3) = 2$

$$(p_1, p_2, p_3) \sim$$



Dirichlet Process

Boxes	Finite measure
K = ∞	$\alpha([x,y]) = y - x$

1 2 3 4 5 6 7 8 9 10 11 12 …

$\alpha(A_1) = 3$ $\alpha(A_2) = 6$ $\alpha(A_3) = 2$

$$(p_1, p_2, p_3, \ldots) \sim$$

Dirichlet Process

Boxes

K = ∞

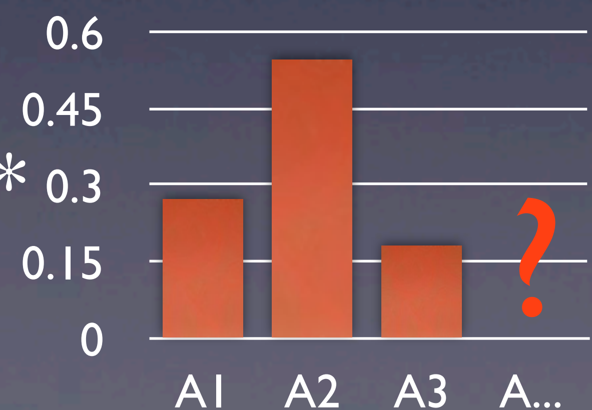
Finite measure

$\alpha([x,y]) = y - x$

1 2 3 4 5 6 7 8 9 10 11 12 •••

$\alpha(A_1) = 3$ $\alpha(A_2) = 6$ $\alpha(A_3) = 2$

$$(p_1, p_2, p_3, \ldots) \sim \alpha(X) *$$



Dirichlet Process

At any point, for any k, the data is Dirichlet distributed

For finite measure α on measure space X

$$(p_1, p_2, \ldots p_k) \sim Dir(\alpha G_0(A_1), \alpha G_0(A_2), \ldots \alpha G_0(A_k))$$

probability of region is based on a region's size

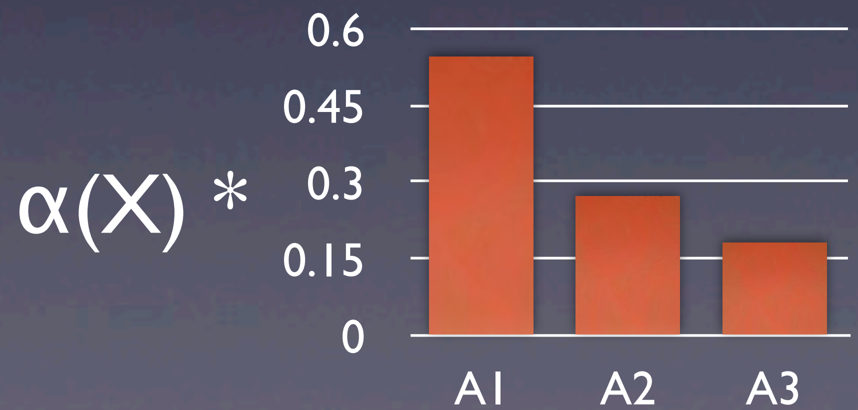
$$p_i = \frac{\alpha(A_i)}{\alpha(X)}$$

Constructing a Sample

Labels: $A_1, A_2, \ldots$ are meaningless so let's order them

1 2 3 4 5 6 7 8 9 10 11 12

$\alpha(A_2) = 3$ $\alpha(A_1) = 6$ $\alpha(A_3) = 2$



Constructing a Sample

Labels: $A_1, A_2, \ldots$ are meaningless so let's order them

$\underbrace{1\ 2\ 3\ 4}_{\alpha(A_2) = 3}\ \underbrace{5\ 6\ 7\ 8\ 9\ 10}_{\alpha(A_1) = 6}\ \underbrace{11\ 12}_{\alpha(A_3) = 2}\ \ldots$

$\alpha(X)$ *

? < A_3

Stick Breaking Construction

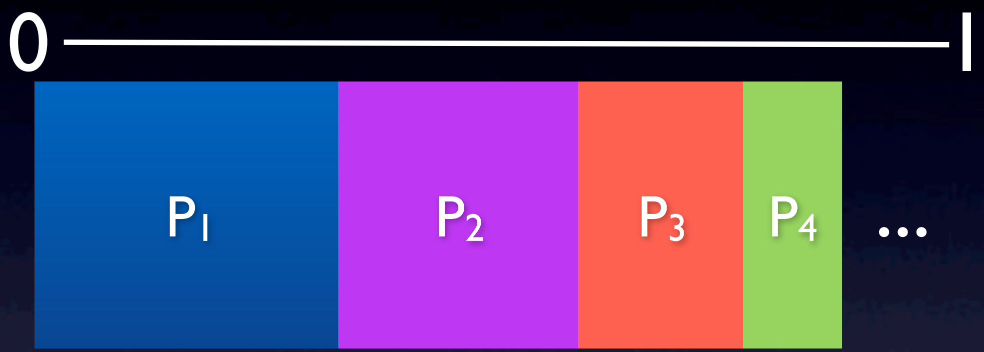
1. Create a diminishing sequence

$$\lim_{i \to \infty} \mathrm{E}[\mathrm{p}_i] = 0 \qquad \sum_{i}^{\infty} \mathrm{p}_i = 1$$

2. Attach weights to regions ($A_1, A_2, \ldots$)

$$\underbrace{1 \quad 2 \quad 3 \quad 4}_{A_2} \quad \underbrace{5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10}_{A_1} \quad \underbrace{11 \quad 12}_{A_3}$$

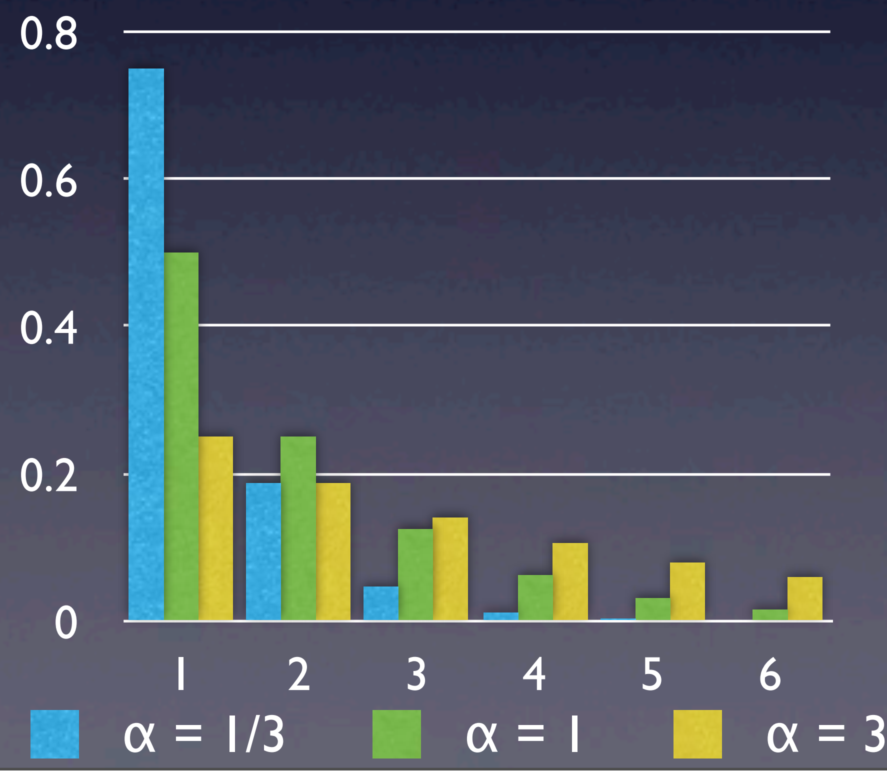
Constructing a sequence



$$E[x] = \frac{1}{1+\alpha}$$

$$V_i \sim \mathrm{Beta}(1, \alpha)$$

$$p_i = V_i \prod_{j<i} (1 - V_j)$$



Stick Breaking Construction

1) we know how to make this sequence now

$$\mathrm{P}(\cdot) = \sum_{k=1}^{\infty} \mathrm{p}_k \delta_{A_k}(\cdot)$$

2) Indicator variable/label assignment

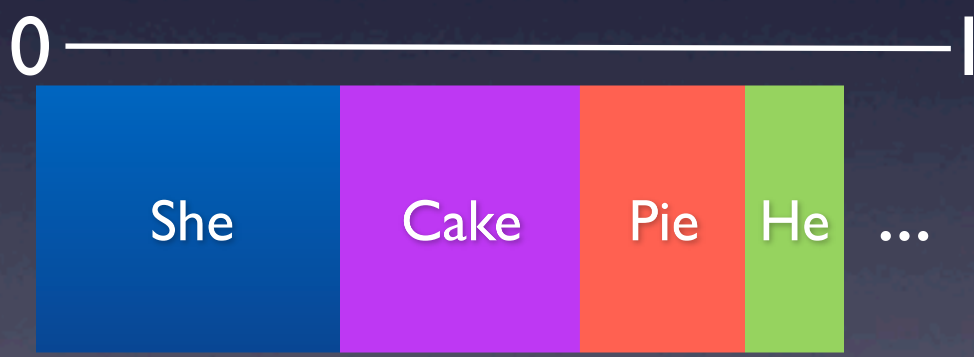
$$A_k \sim \frac{\alpha(\cdot)}{\alpha(X)}$$

DP Summary

- We can construct a sequence of weights
- We can attach them to points in our space
- The distribution of weights is intimately related to the size of our space (α)

Infinite Words

Probability of a word given N:

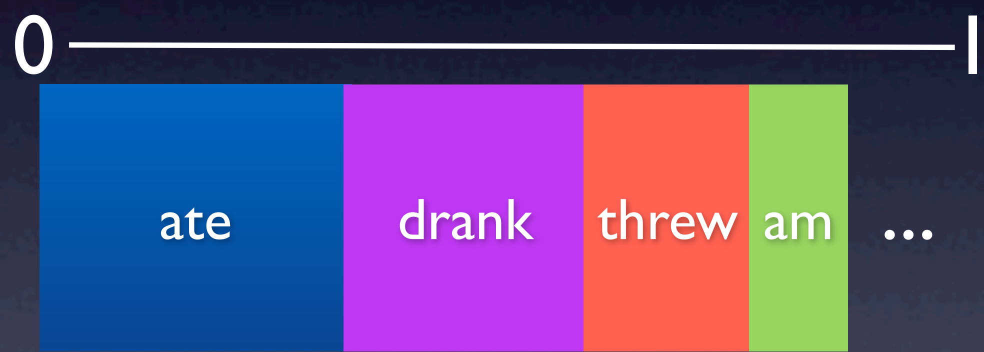


Sharing Knowledge

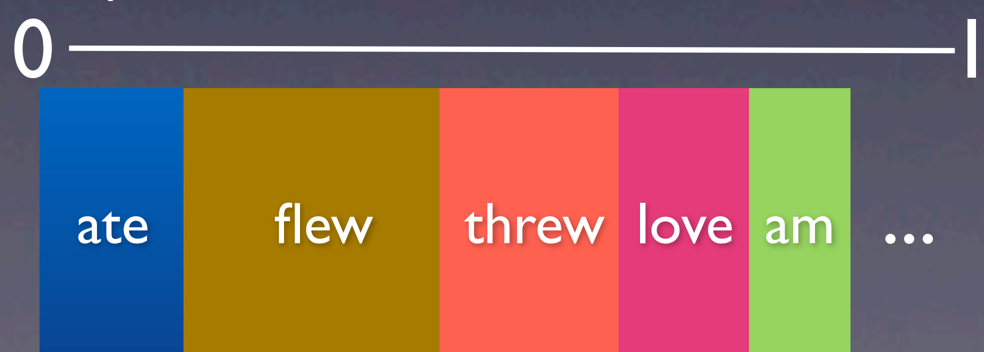
Current situation:

What's the word distribution for ditransitives?

p(word | S\N)



p(word | (S\N)/N)

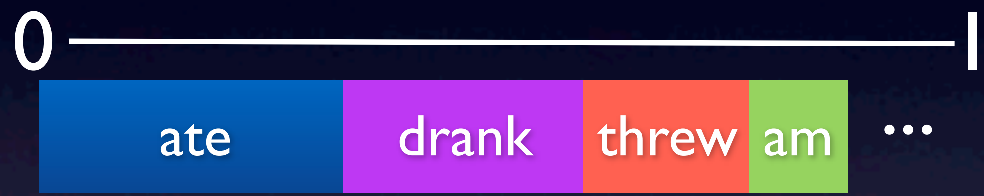


Outline

- Grammar Induction
- Combinatory Categorial Grammar
- Dirichlet Processes
- HDP-CCG

Hierarchical Dirichlet Processes

p(word | S\N)



p(word | (S\N)/N)

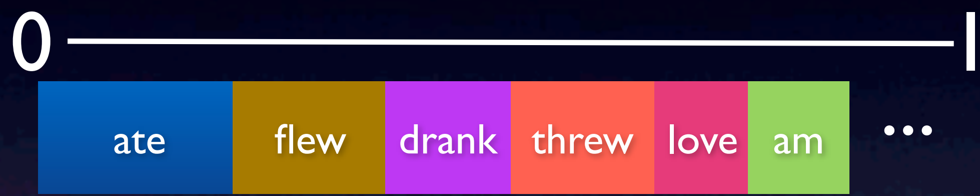
Average?

p(word | ((S\N)/N)/N)

Base DP

p(word | verb)

$W_0 \sim DP(\alpha_0, \{words\})$



p(word | S\N)

p(word | (S\N)/N)

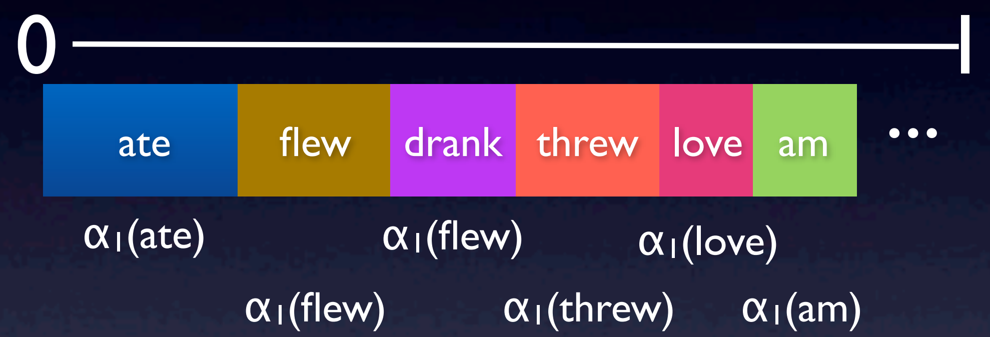
p(word | ((S\N)/N)/N)

Assume distributions are variants of the Base Dirichlet Process

Base DP

p(word | verb)

$W_0 \sim DP(\alpha_0, \{words\})$



p(word | S\N)

$W_{S\backslash N} \sim DP(\alpha_1, W_0)$

0

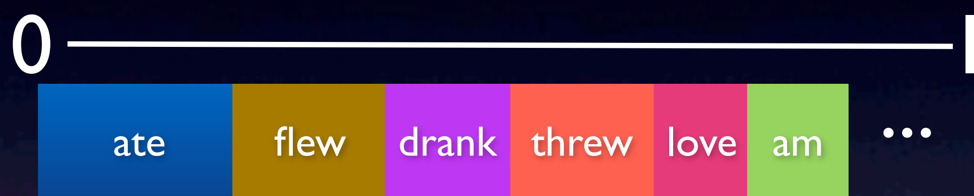
...

$$V_i \sim \mathrm{Beta}(1, \alpha_1)$$

$$p_i = V_i \prod_{j<i} (1 - V_j)$$

Base DP

p(word | verb) $W_0 \sim DP(\alpha_0, \{words\})$



Variance from base DP

p(word | S\N) $W_{S\backslash N} \sim DP(\boxed{\alpha_1}, W_0)$

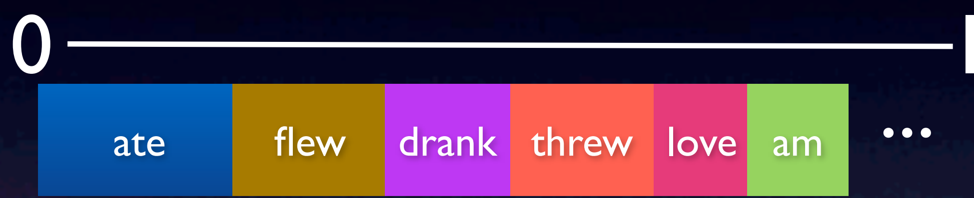
p(word | (S\N)/N) $W_{(S\backslash N)/N} \sim DP(\boxed{\alpha_2}, W_0)$

p(word | ((S\N)/N)/N) $W_{((S\backslash N)/N)/N} \sim DP(\boxed{\alpha_3}, W_0)$

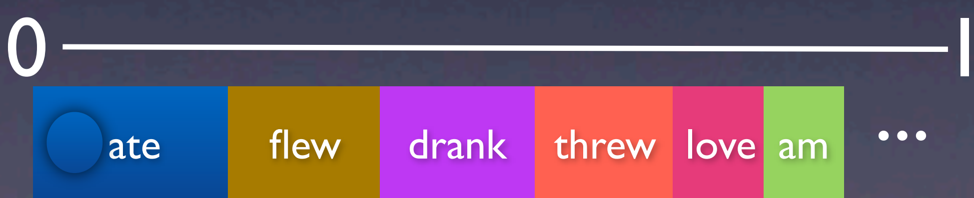
What's the effect?

$W_0 \sim DP(\alpha_0, \{words\})$

p(word | verb)

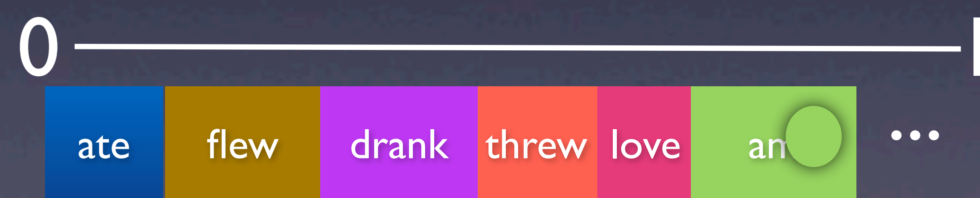


Shared Parameters



p(word | (S\N)/N)

$W_{(S\backslash N)/N} \sim DP(\alpha_2, W_0)$



p(word | S\N)

$W_{S\backslash N} \sim DP(\alpha_1, W_0)$

Effect on performance

	Arabic	Swedish	Basque	English
No Sharing	41.6	70.1	29.6	59.5
HDP-CCG	**66.4**	**74.5**	**50.6**	**70.7**

Length 10 sentences