Discovering Morphological Paradigms from Plain Text

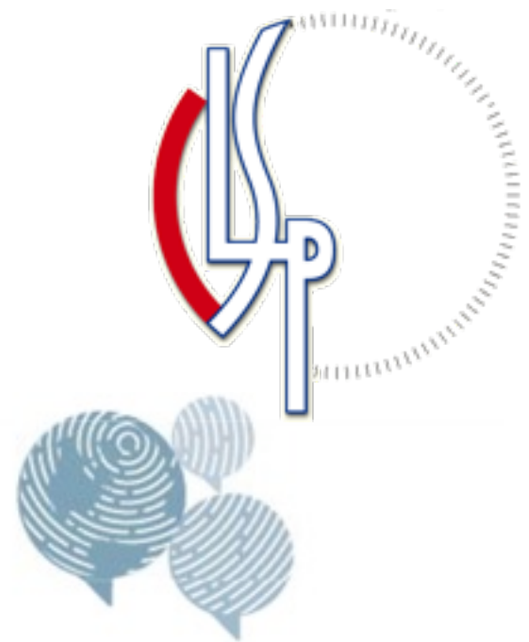
Using a Dirichlet Process Mixture Model

Markus Dreyer
SDL Language Weaver

Jason Eisner
Johns Hopkins University

This work was done at:

Center for Language and Speech Processing (CLSP)
Human Lang. Tech. Center of Excellence (HLTCOE)
Johns Hopkins University (JHU)



Motivation

Rich morphology

English text

break

break

jump break

jump

break

break

break

German text

brichst

brecht

springst brechen

springe

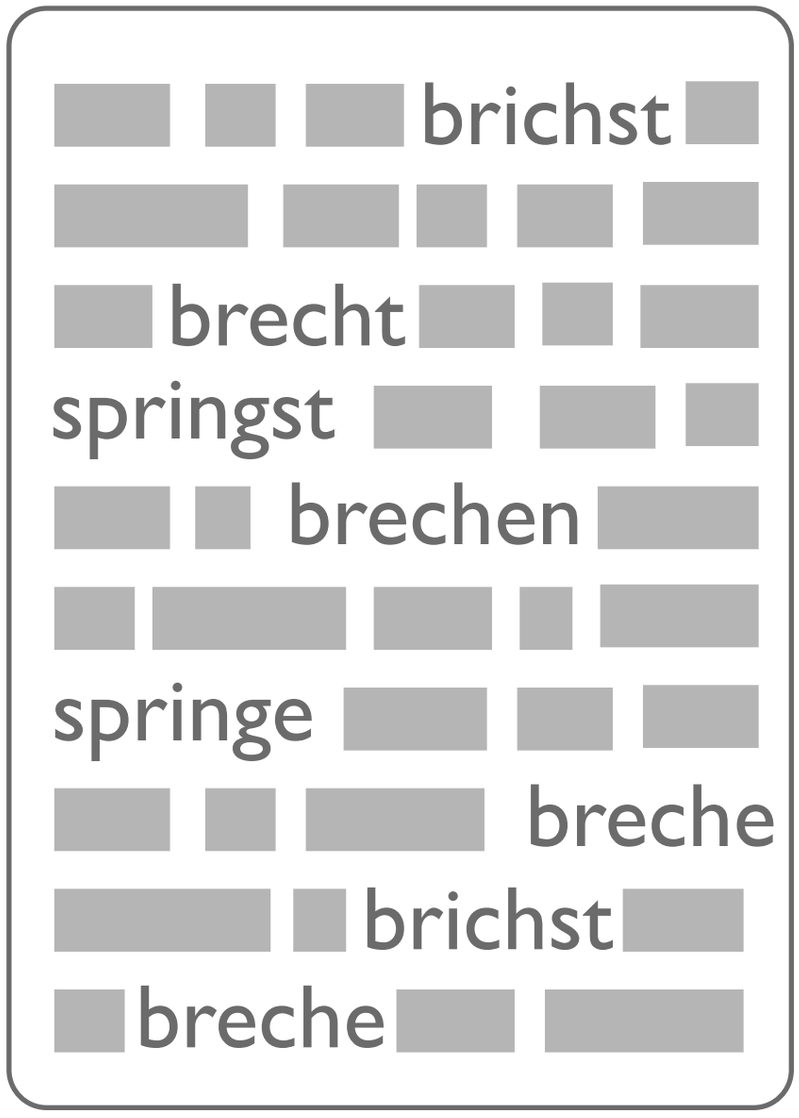
breche

brichst

breche

Motivation

- **Analyzing** text:
  - Lack of generalization
  - Data sparseness
- **Generating** text:
  - Generate correct forms
  - Produce correctly inflected text



There is a need for a general **morphology model** that knows **how to inflect words**.

Motivation

So how do you inflect a word?
You look it up in such a table, for example:

Inflectional Paradigm

infinitive	treffen			
	present		past	
	singular	plural	singular	plural
1st	treffe	treffen	traf	trafen
2nd	triffst	trefft	trafst	traft
3rd	trifft	treffen	traf	trafen

But creating such supervised data is **expensive**.

Let's use unannotated text to learn these paradigms!

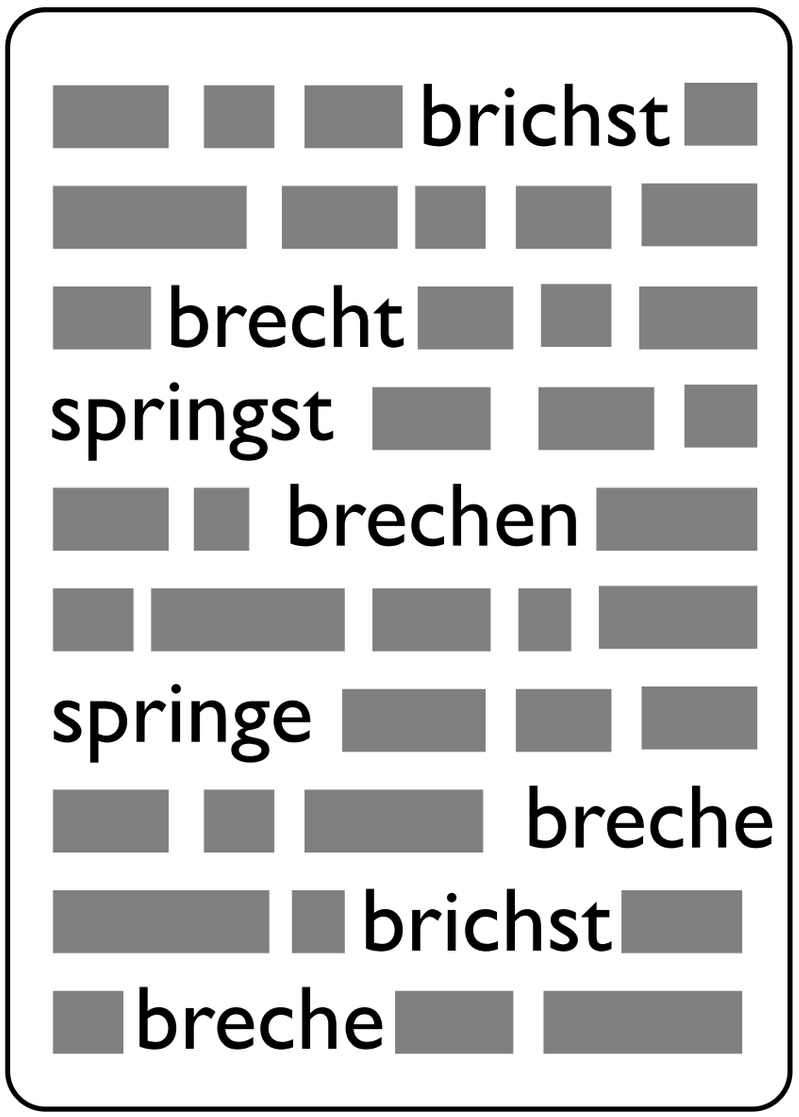
Motivation

- This talk is about a comprehensive **model for inflectional morphology**.
- **Main goal:**
  - Given some **unannotated text**, can we **learn how to inflect** the verbs of a language (incl. irregularities and exceptions)?
  - Discover the **inflectional paradigms** (tables) of a language, using minimal supervision

Motivation

1. Identify the different lexemes in text

German text



Paradigm

infinitive				
1st				
2nd				
3rd				
	singular	plural	singular	plural
	present		past	

Tokens

Types

Motivation

1. Identify the different lexemes in text

German text

brichst

brecht

springst

brechen

springe

breche

brichst

breche

Paradigm

infinitive				
1st				
2nd				
3rd				
	singular	plural	singular	plural
	present		past	

Tokens

Types

Motivation

1. Identify the different lexemes in text

German text

brichst
brecht
springst
brechen
springe
breche
brichst
breche

Paradigm

infinitive				
1st				
2nd				
3rd				
	singular	plural	singular	plural
	present		past	

Tokens

Types

Motivation

2. Place each form of a lexeme into its paradigm

German text

brichst

brecht

springst

brechen

springe

breche

brichst

breche

Paradigm

infinitive	brechen			
1st	breche			
2nd	brichst	brecht		
3rd				
	singular	plural	singular	plural
	present		past	

Tokens

Types

Motivation

2. Place each form of a lexeme into its paradigm

German text

brichst

brecht

springst

brechen

springe

breche

brichst

breche

Paradigm

infinitive	brechen			
1st	breche	brichen? brechen?	brichte? brach?	brichten? brachen?
2nd	brichst	brecht	brichtest? brachst?	brichtet? bracht?
3rd	bricht? brecht?	brichen? brechen?	brichte? brach?	brichten? brachen?
	singular	plural	singular	plural
	present		past	

Tokens

Types

Motivation

2. Place each form of a lexeme into its paradigm

German text

brichst

brecht

springst

brechen

springe

breche

brichst

breche

Paradigm

		present		past
		singular	plural	singular	plural
infinitive					
1st					
2nd					
3rd					

Tokens

Types

Motivation

2. Place each form of a lexeme into its paradigm

German text

brichst

brecht

springst

brechen

springe

breche

brichst

breche

Paradigm

infinitive				
1st	springe			
2nd	springst			
3rd				
	singular	plural	singular	plural
	present		past	

Tokens

Types

Motivation

2. Place each form of a lexeme into its paradigm

German text

brichst

brecht

springst

brechen

springe

breche

brichst

breche

Paradigm

infinitive	springen? sprengen?			
1st	springe	springen? sprengen?	springte? sprang?	springte? sprang?
2nd	springst	springt? sprengt?	springtest? sprangst?	springtet? sprangt?
3rd	springt? sprengt?	springen? sprengen?	springte? sprang?	springten? sprangen?
	singular	plural	singular	plural
	present		past	

Tokens

Types

Motivation

German text

brichst

brecht

springst

brechen

springe

breche

brichst

breche

Paradigm

infinitive	springen? sprengen?			
1st	springe	springen? sprengen?	springte? sprang?	springte? sprang?
2nd	springst	springt? sprengt?	springtest? sprangst?	springtet? sprangt?
3rd	springt? sprengt?	springen? sprengen?	springte? sprang?	springten? sprangen?
	singular	plural	singular	plural
	present		past	

Tokens

Types

Motivation

German text

brichst

brecht

springst

brechen

springe

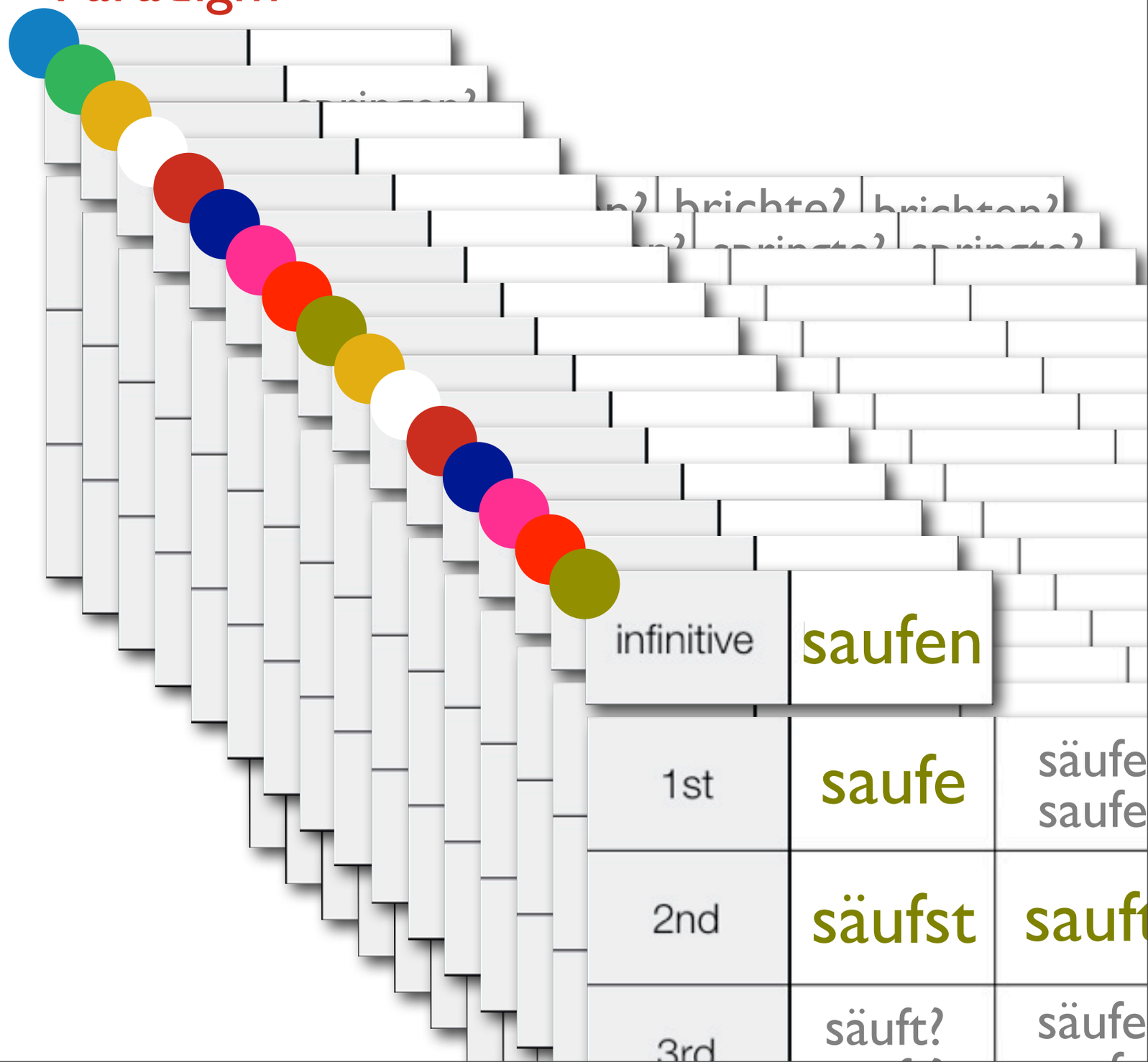
breche

brichst

breche

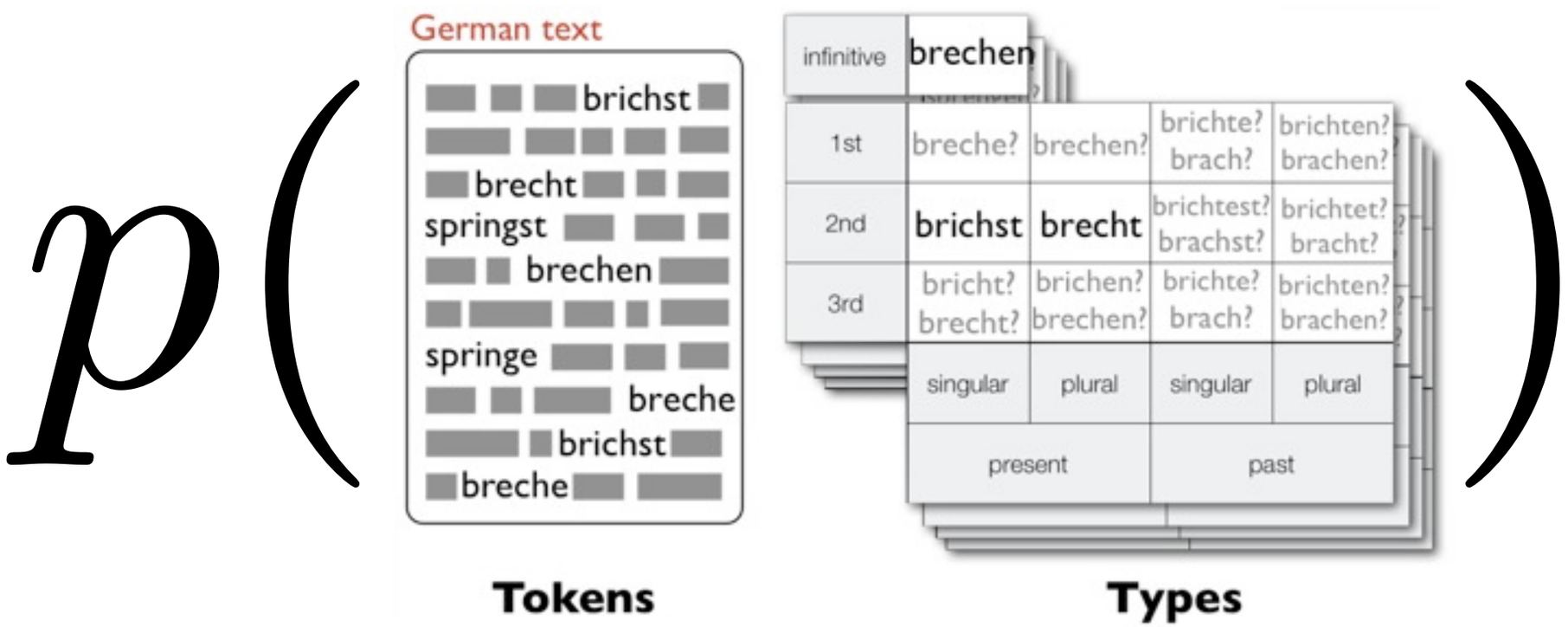
Tokens

Paradigm



Motivation

In order to perform this morphological knowledge discovery, we define a **probability distribution** over a **text corpus** and its (hidden) inflectional **paradigms**:



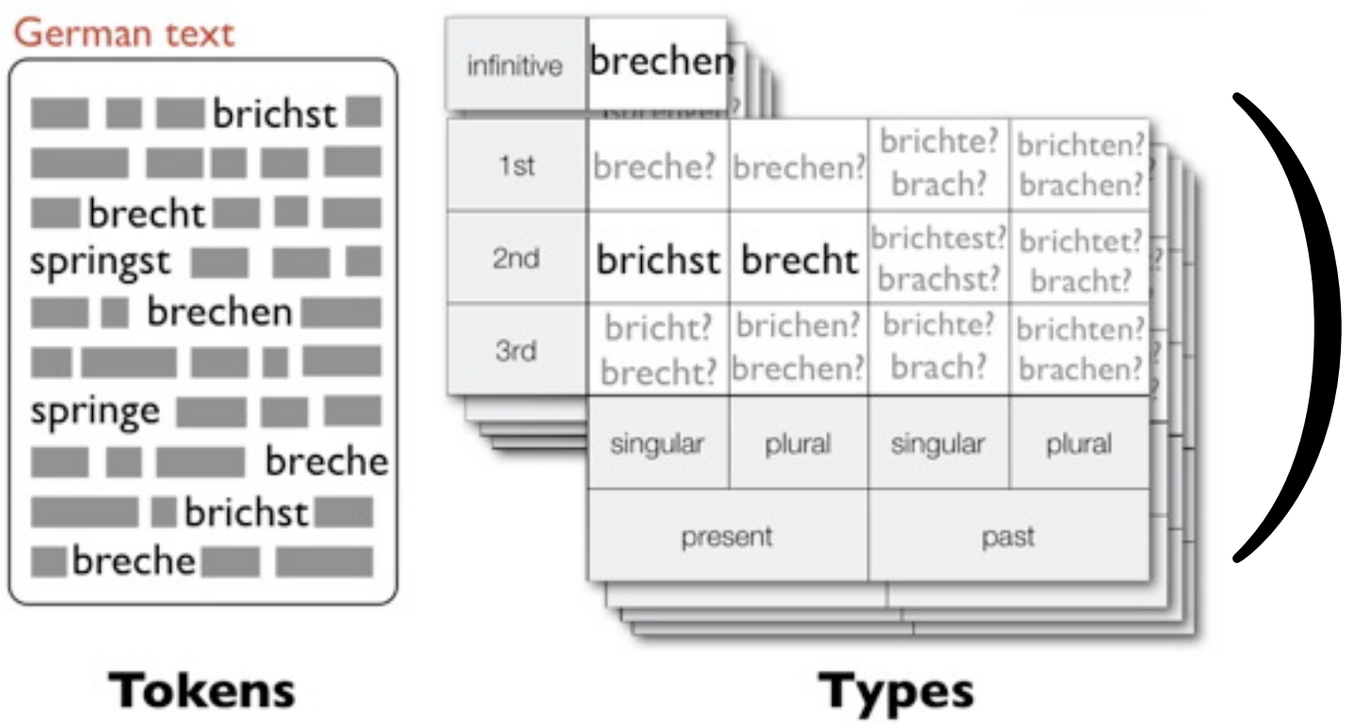
Overview

1 $p($

infinitive	brechen			
1st	breche?	brechen?	brichte? brach?	brichten? brachen?
2nd	brichst	brecht	brichtest? brachst?	brichtet? bracht?
3rd	bricht? brecht?	brichen? brechen?	brichte? brach?	brichten? brachen?
	singular	plural	singular	plural
	present		past	

$)$

2 $p($



$)$

Overview

1 $p($

infinitive	brechen			
1st	breche?	brechen?	brichte? brach?	brichten? brachen?
2nd	brichst	brecht	brichtest? brachst?	brichtet? bracht?
3rd	bricht? brecht?	brichen? brechen?	brichte? brach?	brichten? brachen?
	singular	plural	singular	plural
	present		past	

$)$

2 $p($

Tokens

Types

$)$

1 Paradigms

Why build probability model over paradigms?

infinitive	brichen? brechen?	
1st	breche	brichen? brechen?
2nd	brichst	brecht
3rd	bricht? brecht?	brichen? brechen?
	singular	plural
	present	

- Jointly predict missing string values
- Compute marginals
- Know what spellings are likely in the different cells

1 Paradigms

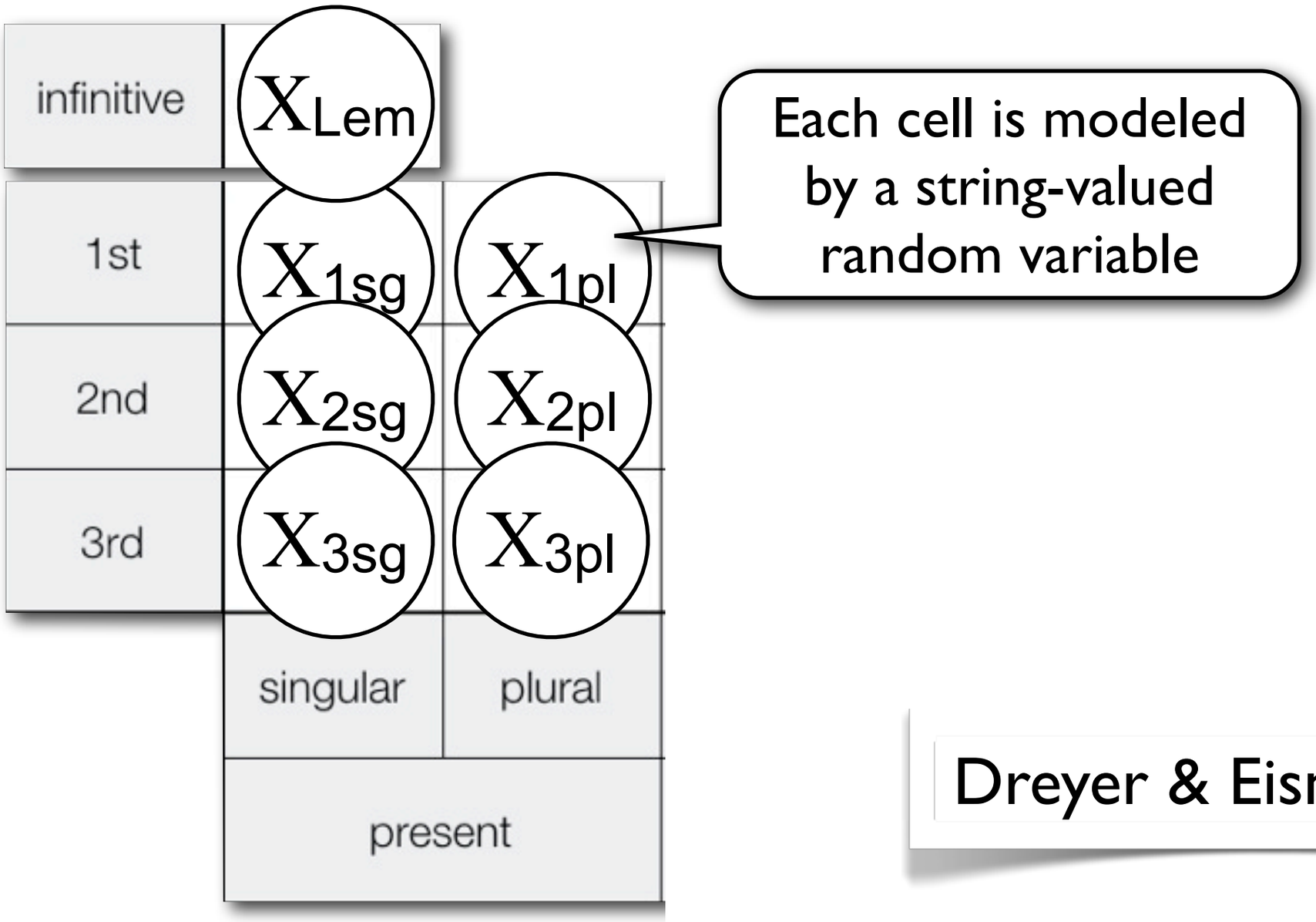
How to build probability model over paradigms?

infinitive	brichen? brechen?	
1st	breche	brichen? brechen?
2nd	brichst	brecht
3rd	bricht? brecht?	brichen? brechen?
	singular	plural
	present	

Dreyer & Eisner (2009)

1 Paradigms

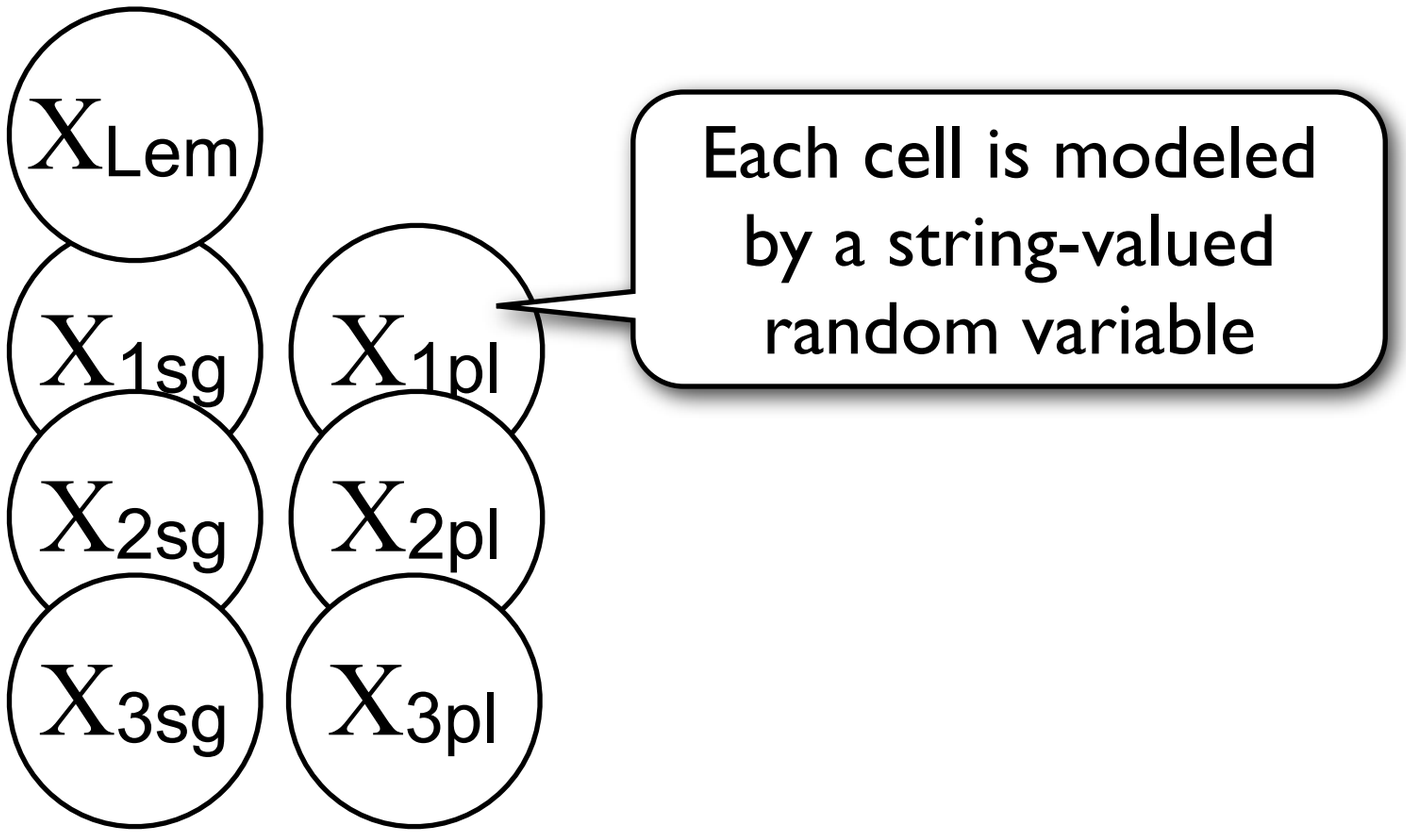
How to build probability model over paradigms?



Dreyer & Eisner (2009)

1 Paradigms

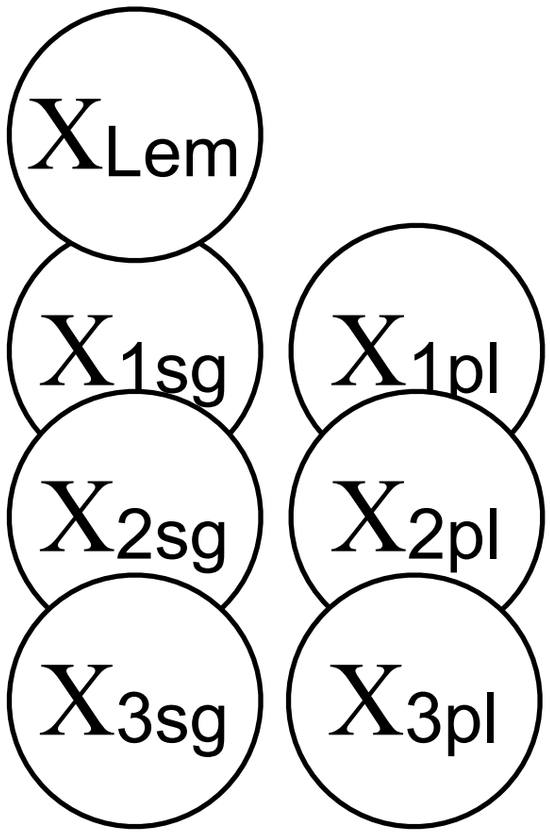
How to build probability model over paradigms?



Dreyer & Eisner (2009)

1 Paradigms

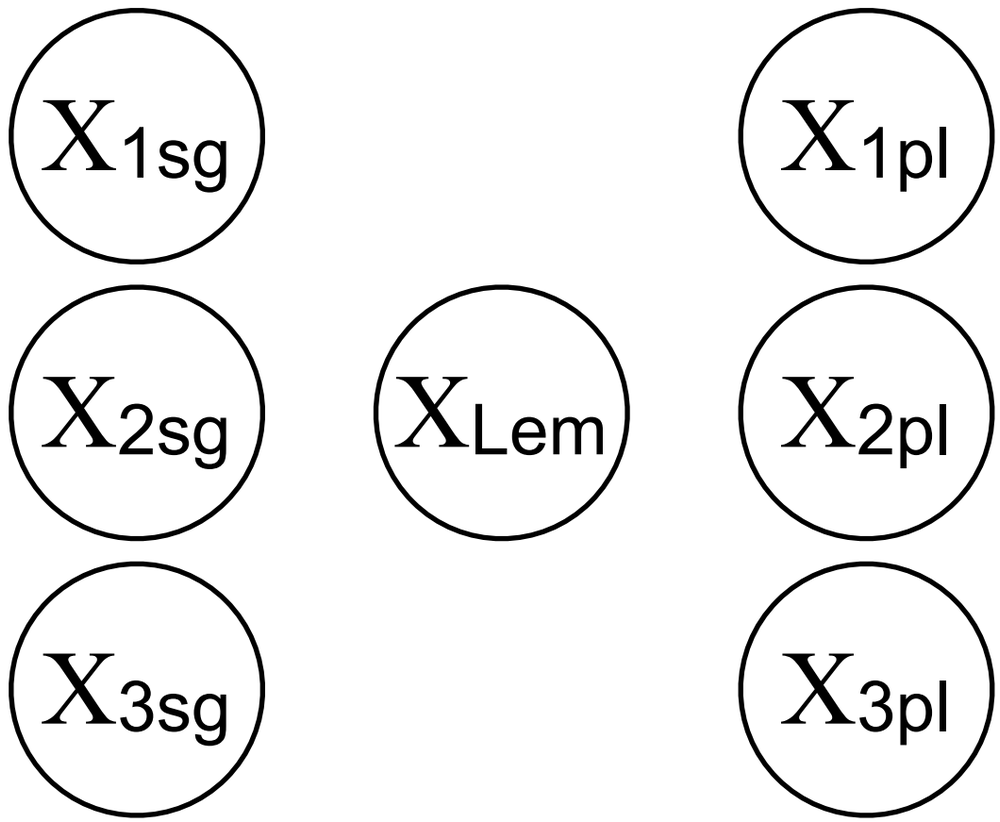
How to build probability model over paradigms?



Dreyer & Eisner (2009)

1 Paradigms

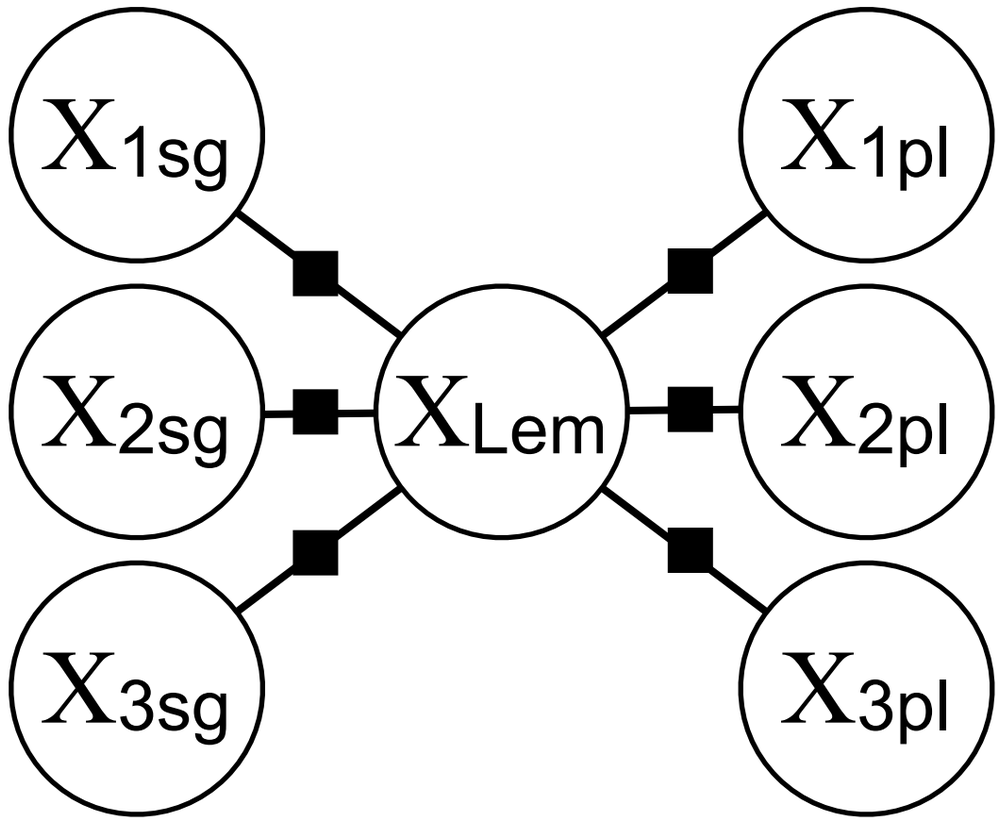
How to build probability model over paradigms?



Dreyer & Eisner (2009)

1 Paradigms

How to build probability model over paradigms?

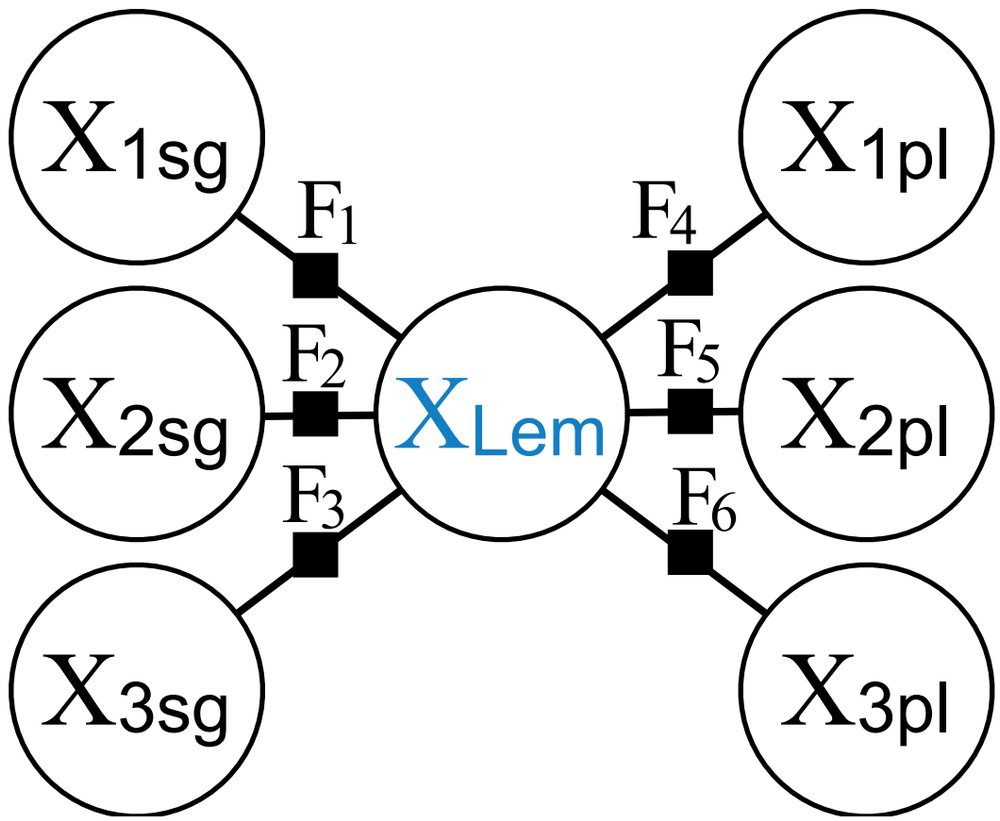


Dreyer & Eisner (2009)

1 Paradigms

How to build probability model over paradigms?

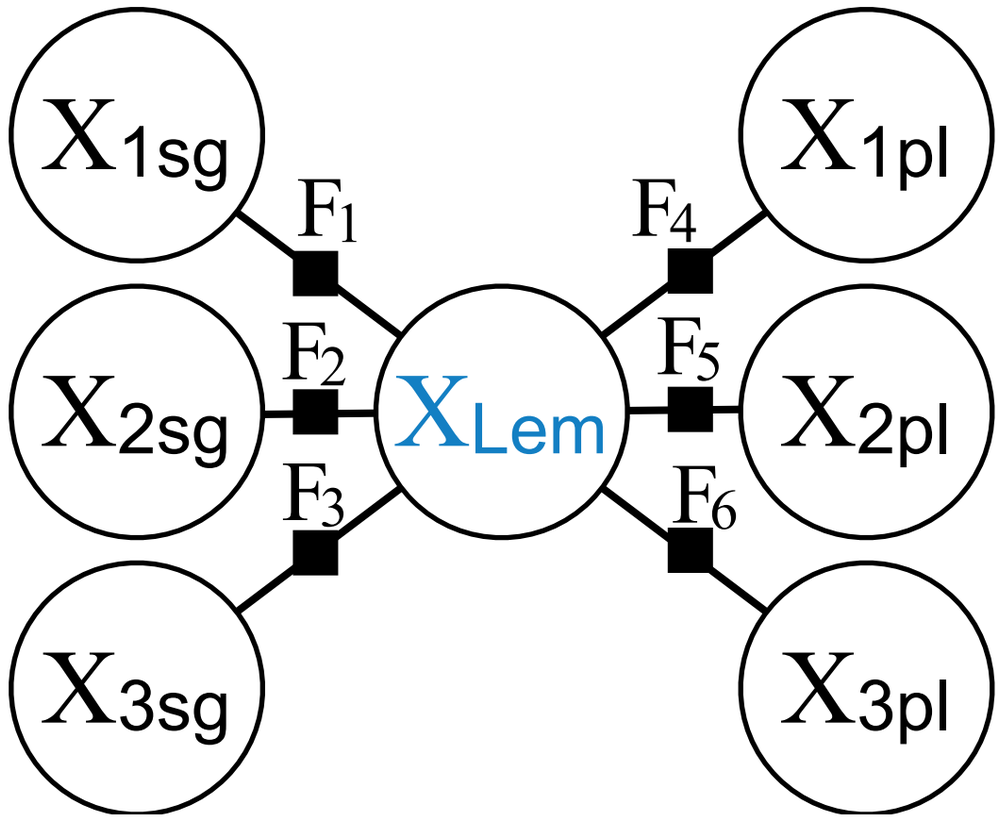
$$p(X_{Lem}, X_{1sg}, X_{2sg}, X_{3sg}, X_{1pl}, X_{2pl}, X_{3pl}) =$$



Markov Random Field over string-valued variables

$$\frac{1}{Z}$$
$$\times F_1(X_{Lem}, X_{1sg})$$
$$\times F_2(X_{Lem}, X_{2sg})$$
$$\times F_3(X_{Lem}, X_{3sg})$$
$$\times F_4(X_{Lem}, X_{1pl})$$
$$\times F_5(X_{Lem}, X_{2pl})$$
$$\times F_6(X_{Lem}, X_{3pl})$$

1 Paradigms



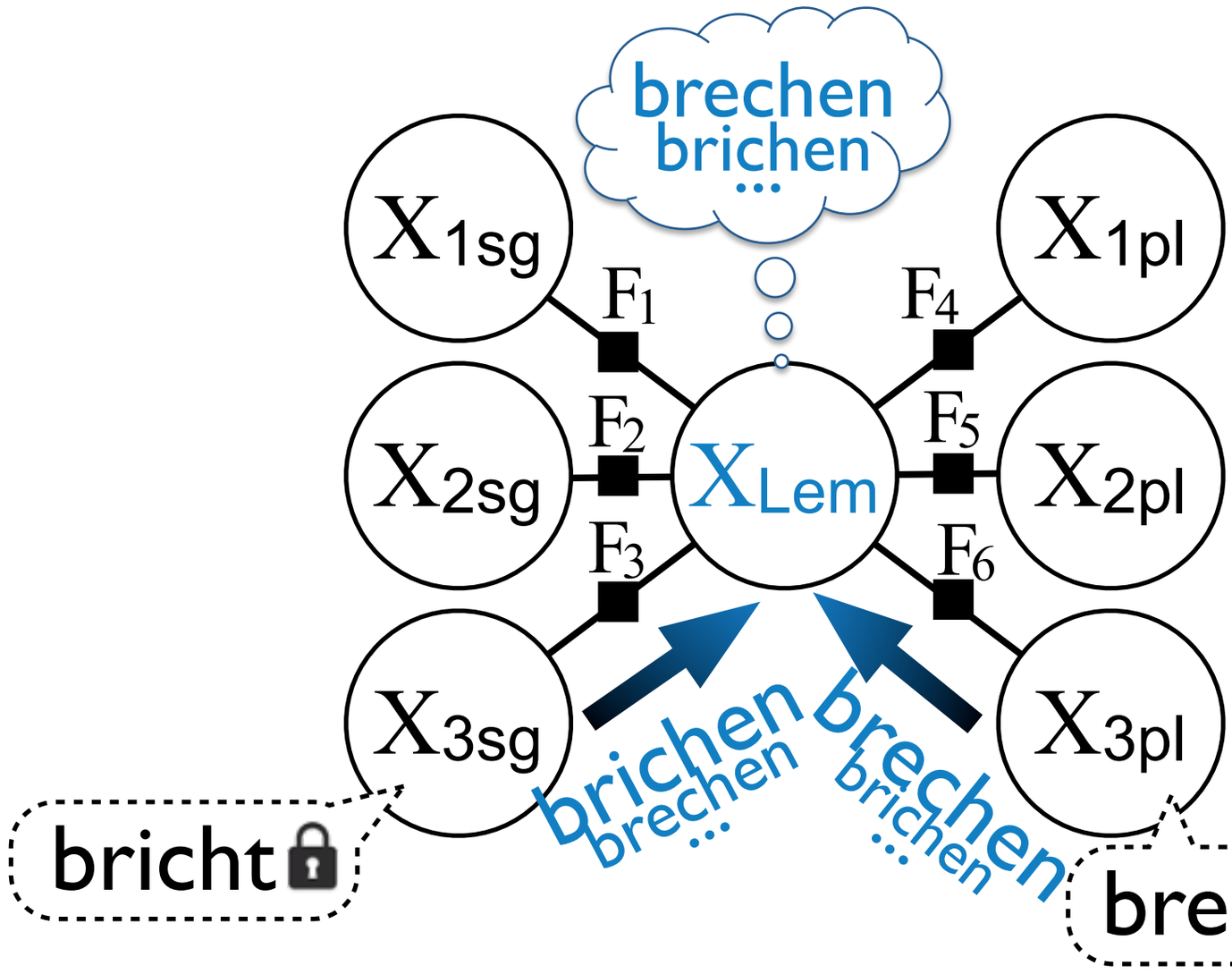
Markov Random Field over string-valued variables

Belief Propagation:

- Standard inference algorithm
- Computes Marginals through message passing

- We use finite-state variant of this algorithm, Dreyer & Eisner (2009)

1 Paradigms



Belief Propagation:

- Standard inference algorithm
- Computes Marginals through message passing

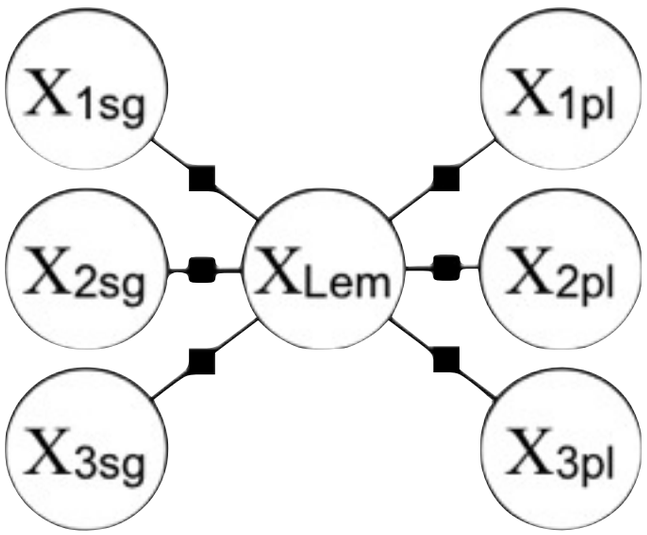
- We use finite-state variant of this algorithm, Dreyer & Eisner (2009)

1 Paradigms

Summary

- Paradigms modeled as Markov Random Fields (**MRF**)
- Weighted finite-state transducers (**FST**) relate the various spellings to one another
- They encode morphological knowledge ("**grammar**")
- Use finite-state-based belief propagation (**BP**) to compute string marginals

infinitive	brichen? brechen?	
1st	**breche**	brichen? brechen?
2nd	**brichst**	**brecht**
3rd	bricht? brecht?	brichen? brechen?
	singular	plural
	present	



1 Paradigms

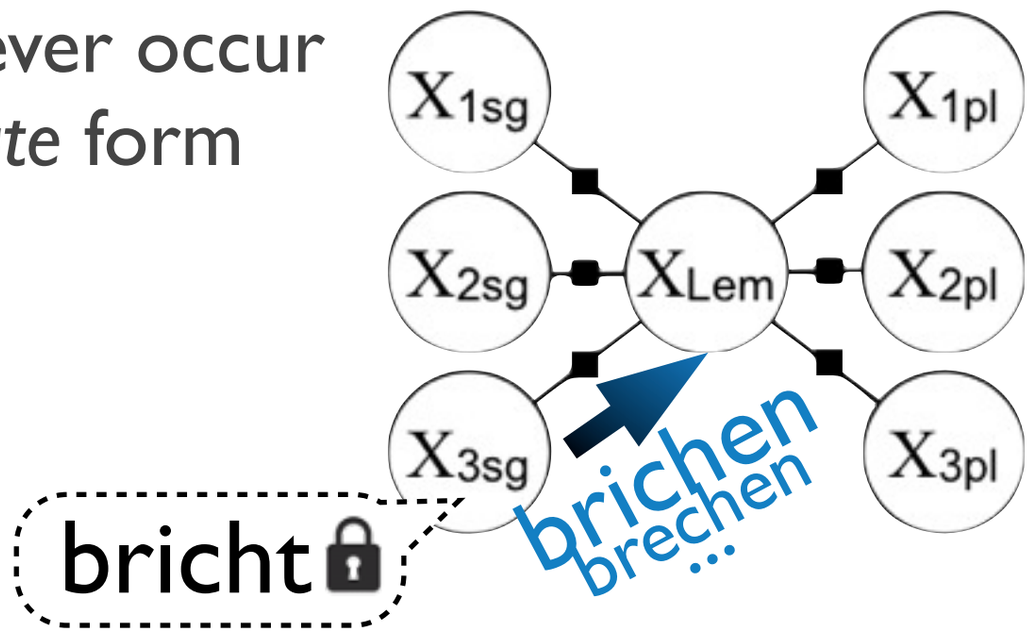
Summary

- Dreyer & Eisner (2009):
  - Learn purely from example paradigms (training data)
  - Then use model to predict unseen forms

infinitive	brichen? brechen?	
1st	**breche**	brichen? brechen?
2nd	**brichst**	**brecht**
3rd	bricht? brecht?	brichen? brechen?
singular	plural	
present		

- **Disadvantages:**
  - Training data is expensive
  - Predicts forms that would never occur in real text (where an *alternate* form may be preferred)

We will now address these problems.



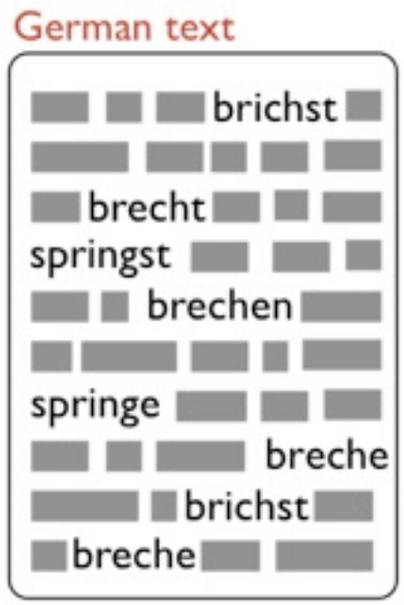
Overview

1 $p($

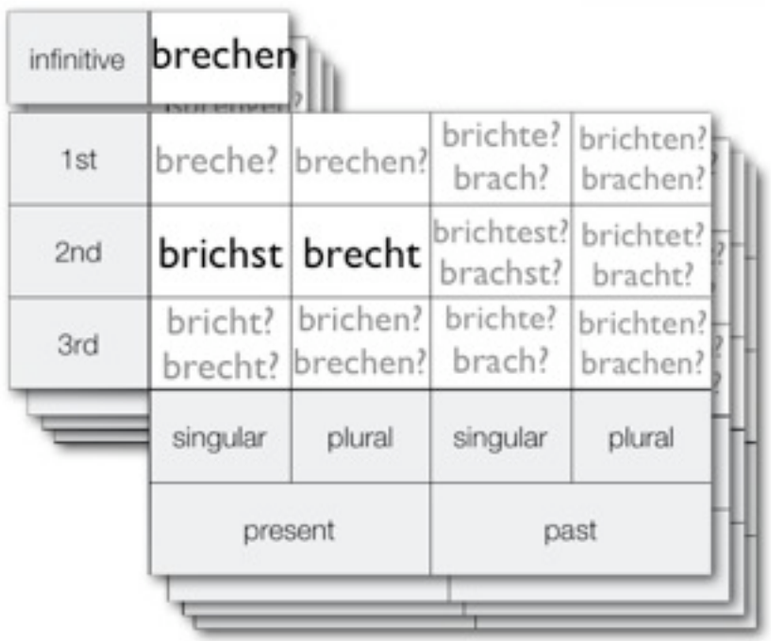
infinitive	brechen			
1st	breche?	brechen?	brichte? brach?	brichten? brachen?
2nd	brichst	brecht	brichtest? brachst?	brichtet? bracht?
3rd	bricht? brecht?	brichen? brechen?	brichte? brach?	brichten? brachen?
	singular	plural	singular	plural
	present		past	

$)$

2 $p($



Tokens



Types

$)$

2 Lexicon & Corpus

- We use the paradigms to construct a **probabilistic lexicon** that specifies which inflections of which lexemes are *common* and how they are *spelled*.

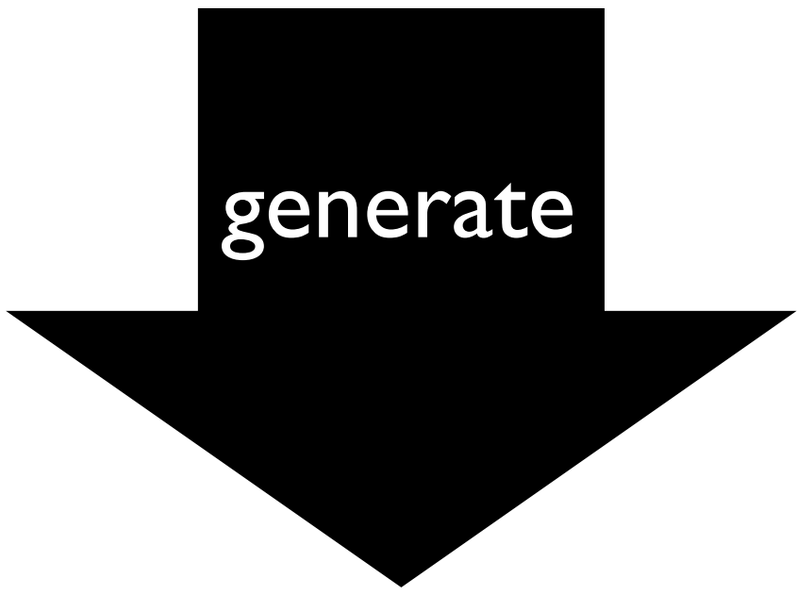
- We define a generative **probability model** of the lexicon and a text corpus.

- This allows for **clean inference procedure** to learn morphology from text and discover its inflectional paradigms

2 Lexicon & Corpus

Generative story

Model



Data

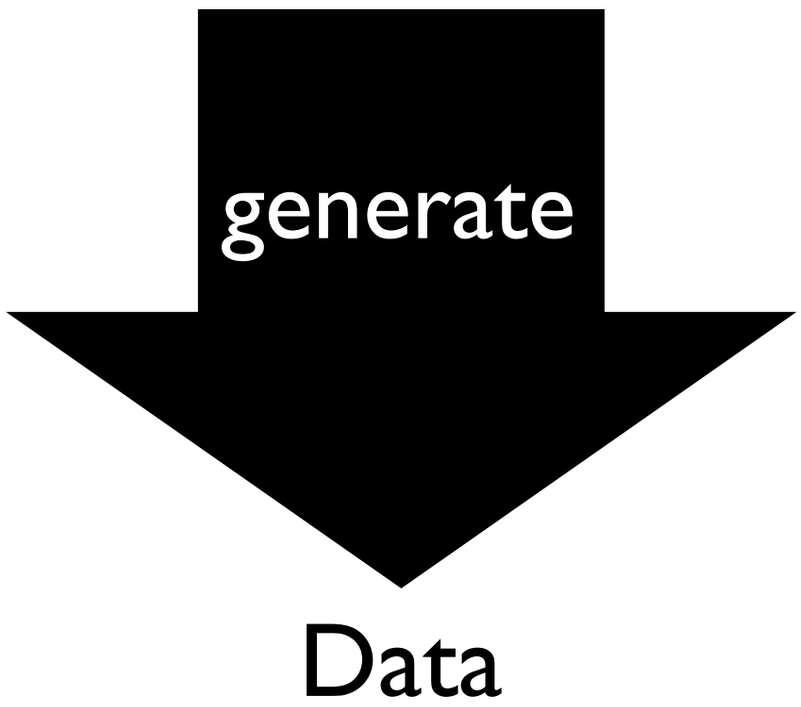
Inference (Sampling)

Model

Data

Lexicon & Corpus

Generative story



To generate from our model:

- First, generate the **lexicon** (types).
- Then, use it to generate the **corpus** (tokens).

2 Lexicon & Corpus

Generating...

(1) Choose a distribution over lexemes

Stick-breaking process

2 Lexicon & Corpus

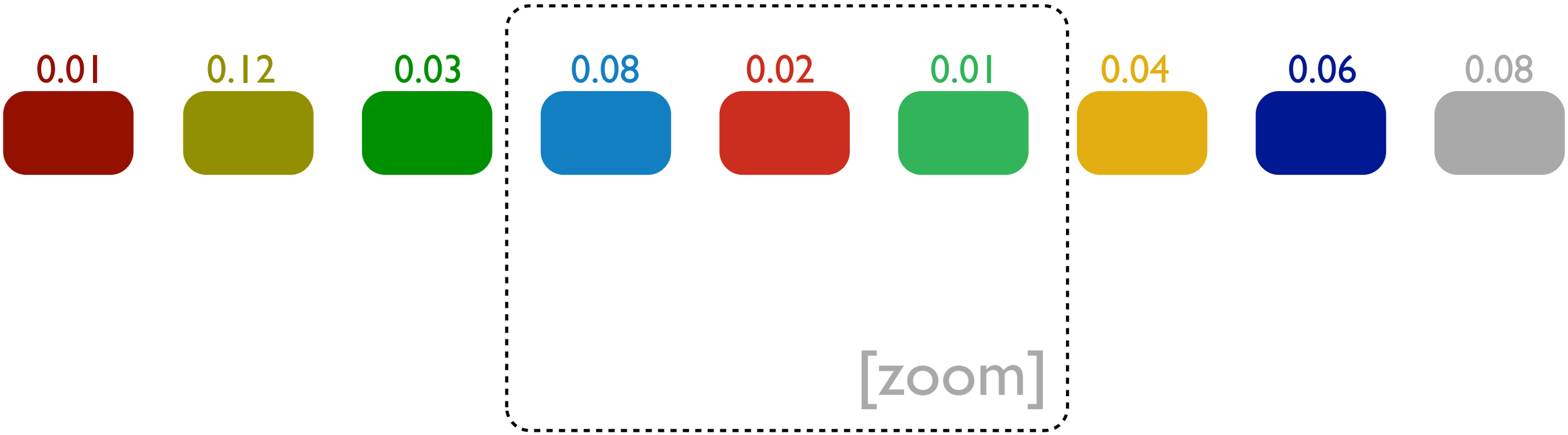
Generating...

(1) Choose a distribution over lexemes

Stick-breaking process

2 Lexicon & Corpus

Generating...



(1) Choose a distribution over lexemes

Stick-breaking process

2 Lexicon & Corpus

Generating...

(1) Choose a distribution over lexemes

2 Lexicon & Corpus

Generating...

(2) For each lexeme, choose a distribution over its inflections

2 Lexicon & Corpus

Generating...

(2) For each lexeme, choose a distribution over its inflections

2 Lexicon & Corpus

Generating...

0.08

.3	1st sg	1st pl	.12
.07	2nd sg	2nd pl	.05
.26	3rd sg	3rd pl	.2

0.02

.06	1st sg	1st pl	.08
.3	2nd sg	2nd pl	.03
.4	3rd sg	3rd pl	.13

0.01

.4	1st sg	1st pl	.11
.08	2nd sg	2nd pl	.25
.06	3rd sg	3rd pl	.1

(3) For each lexeme, choose a paradigm that expresses the lexeme orthographically

2 Lexicon & Corpus

Generating...

0.08

.3	breche	brechen	.12
.07	brichst	brecht	.05
.26	bricht	brechen	.2

0.02

.06	treffe	treffen	.08
.3	triffst	trefft	.03
.4	trifft	treffen	.13

0.01

.4	springe	springen	.11
.08	springst	springt	.25
.06	springt	springen	.1

(3) For each lexeme, choose a paradigm that expresses the lexeme orthographically

2 Lexicon & Corpus

Generating... Done!

0.08

.3	breche	brechen	.12
.07	brichst	brecht	.05
.26	bricht	brechen	.2

0.02

.06	treffe	treffen	.08
.3	triffst	trefft	.03
.4	trifft	treffen	.13

0.01

.4	springe	springen	.11
.08	springst	springt	.25
.06	springt	springen	.1

(3) For each lexeme, choose a paradigm that expresses the lexeme orthographically

2 Lexicon & Corpus

Generating... Done!

0.08

.3	breche	brechen	.12
.07	brichst	brecht	.05
.26	bricht	brechen	.2

0.02

.06	treffe	treffen	.08
.3	triffst	trefft	.03
.4	trifft	treffen	.13

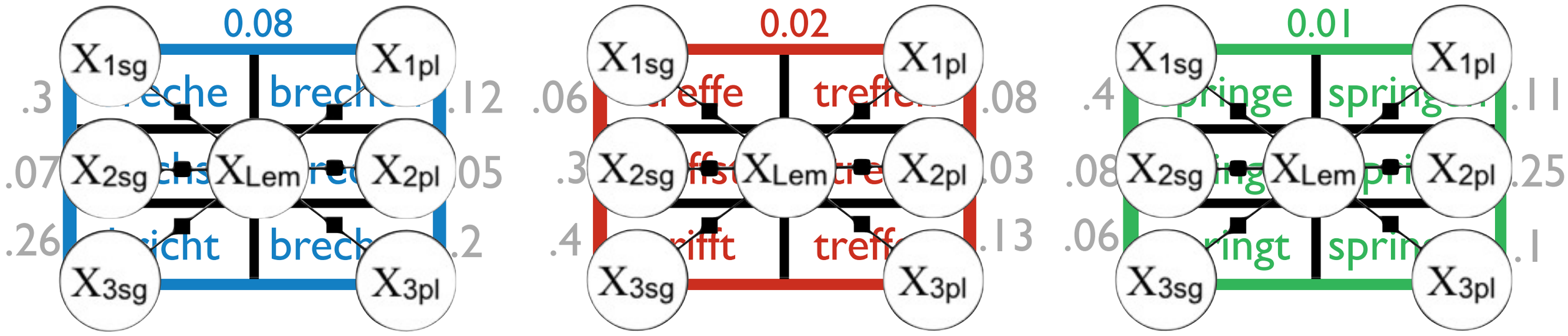
0.01

.4	springe	springen	.11
.08	springst	springt	.25
.06	springt	springen	.1

(3) For each lexeme, choose a paradigm that expresses the lexeme orthographically

2 Lexicon & Corpus

Generating... Done!



(3) For each lexeme, choose a paradigm that expresses the lexeme orthographically

2 Lexicon & Corpus

Generating... Done!

0.08

.3	breche	brechen	.12
.07	brichst	brecht	.05
.26	bricht	brechen	.2

0.02

.06	treffe	treffen	.08
.3	triffst	trefft	.03
.4	trifft	treffen	.13

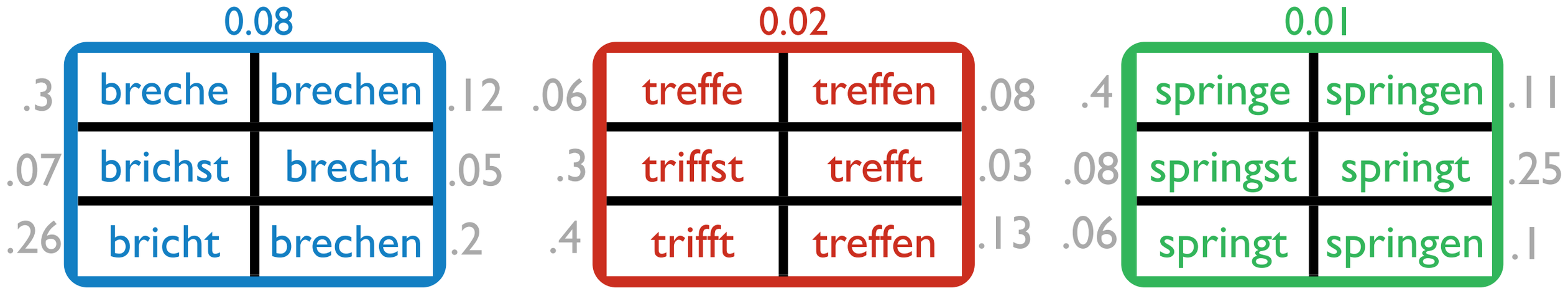
0.01

.4	springe	springen	.11
.08	springst	springt	.25
.06	springt	springen	.1

(3) For each lexeme, choose a paradigm that expresses the lexeme orthographically

2 Lexicon & Corpus

Generating...



The lexicon has been generated
(i.e., the types of the language).
Now generate the corpus
(i.e., the tokens).

2 Lexicon & Corpus

Generating... Done!

0.08			
.3	breche	brechen	.12
.07	brichst	brecht	.05
.26	bricht	brechen	.2

0.02			
.06	treffe	treffen	.08
.3	triffst	trefft	.03
.4	trifft	treffen	.13

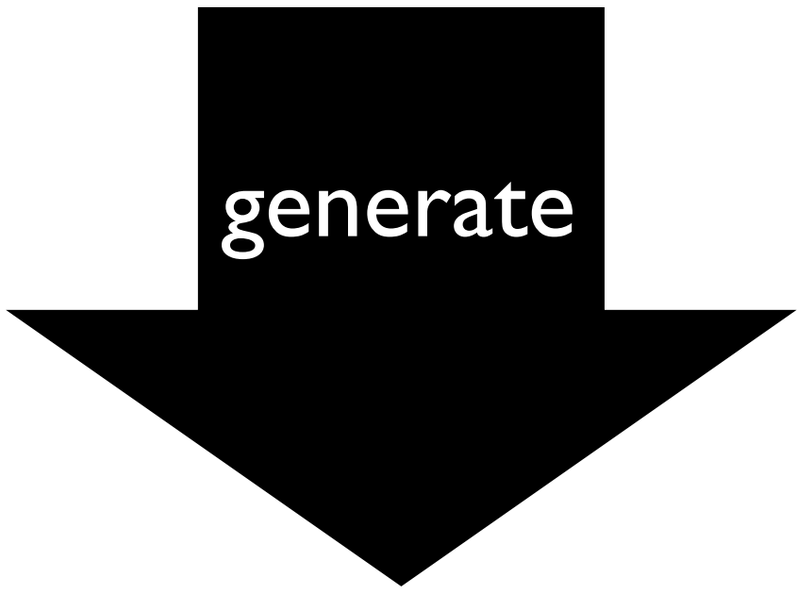
0.01			
.4	springe	springen	.11
.08	springst	springt	.25
.06	springt	springen	.1

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	PREP	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		1st pl		2nd pl		1st pl		2nd sg
Spelling	bricht		brechen		springt		brechen		triffst

2 Lexicon & Corpus

Generative story

Model

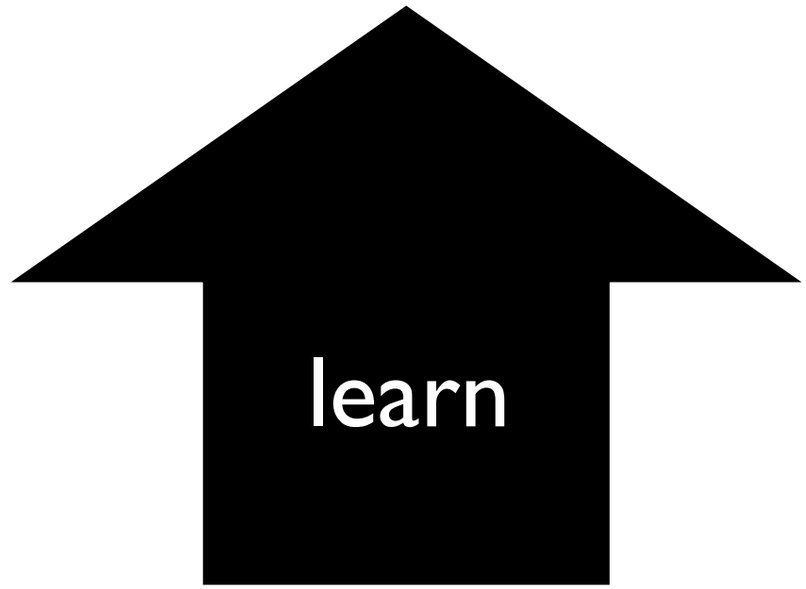


Data

2 Lexicon & Corpus

Inference (Sampling)

Model



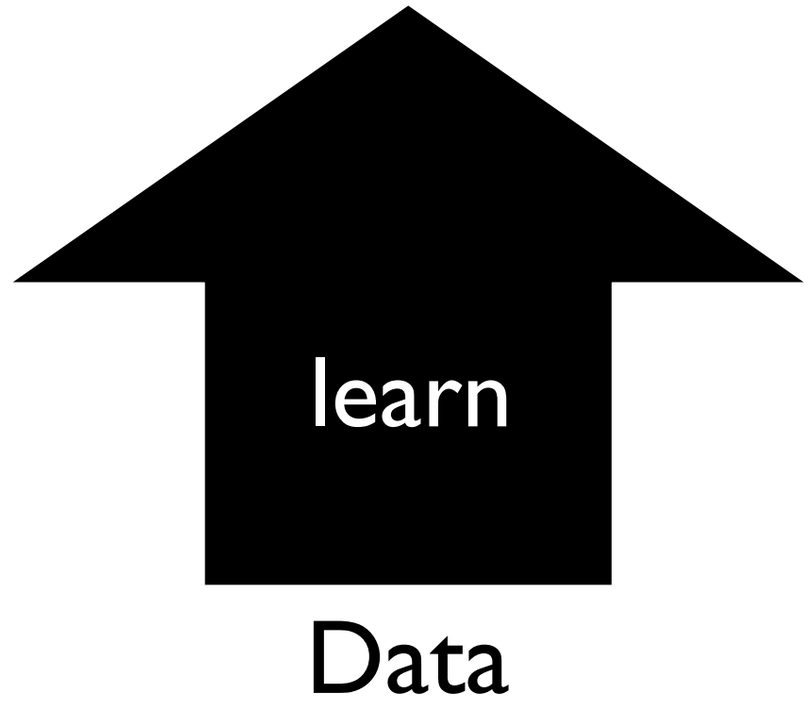
Data

2 Lexicon & Corpus

To do inference:

- Start with observed corpus
- Construct lexicon and **estimate** all distributions

Inference (Sampling)



2 Lexicon & Corpus

0.08			
.3	breche	brechen	.12
.07	brichst	brecht	.05
.26	bricht	brechen	.2

0.02			
.06	treffe	treffen	.08
.3	triffst	trefft	.03
.4	trifft	treffen	.13

0.01			
.4	springe	springen	.11
.08	springst	springt	.25
.06	springt	springen	.1

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	PREP	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		1st pl		2nd pl		1st pl		2nd sg
Spelling	bricht		brechen		springt		brechen		triffst

2 Lexicon & Corpus

Index	1	2	3	4	5	6	7	8	9
POS									
Lexeme									
Inflection									
Spelling	bricht		brechen		springt		brechen		triffst

2 Lexicon & Corpus

Index	1	2	3	4	5	6	7	8	9
POS									
Lexeme									
Inflection									
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection									
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

Seed paradigm

treffe	treffen
triffst	trefft
trifft	treffen

Minimal supervision:
We observe **some paradigms**, from which we can estimate an initial θ (which parameterizes the finite-state MRFs)

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection									
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

Seed paradigm

treffe	treffen
triffst	trefft
trifft	treffen

Train initial θ values

"morphological grammar"

"e" is likely to change into "i"

3rd sg ends in "t"

from 3rd sg to 1st pl, change vowel

...

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection									
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection									
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

1

The red lexeme is completely specified and “bricht” does **not** fit in.

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection									
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

1

The red lexeme is completely specified and “bricht” does **not** fit in.

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg								
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

The red lexeme is completely specified and “bricht” does **not** fit in.

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection									
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

1st sg	1st pl
2nd sg	2nd pl
3rd sg	3rd pl

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection									
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

1st sg	1st pl
2nd sg	2nd pl
bricht ❶	3rd pl

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection									
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

1st sg	1st pl
2nd sg	2nd pl
bricht	3rd pl

❶

Index	❶	❷	❸	❹	❺	❻	❼	❽	❾
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg								
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brichen? brechen?

1

We immediately run finite-state-based **belief propagation** in this new paradigm.

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg								
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brichen? brechen?

1

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg								
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 3

We have just removed the nonsense form "brichen" (and others)

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg								
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 3

We have just removed the nonsense form "brichen" (and others)

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl						
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 3

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl						
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 3

1st sg	1st pl
2nd sg	2nd pl
3rd sg	3rd pl

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl						
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 3

1st sg	1st pl
2nd sg	2nd pl
springt	3rd pl

5

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl						
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 3

1st sg	1st pl
2nd sg	2nd pl
springt	3rd pl

5

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl		3rd sg				
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 3

springe? sprenge?	springen? sprengen?
springst? sprengst?	springt? sprengt?
springt	springen? sprengen?

5

Run belief propagation!

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl		3rd sg				
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

briche? breche?	brichen? **brechen**?
brichst? brechst?	bricht? brecht?
bricht	**brechen**

1 3

springe? sprenge?	springen? sprengen?
springst? sprengst?	springt? sprengt?
springt	springen? sprengen?

5

It would fit well in two of the cells

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl		3rd sg				
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 3

springe? sprenge?	springen? sprengen?
springst? sprengst?	springt? sprengt?
springt	springen? sprengen?

5

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl		3rd sg				
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 7 3

springe? sprenge?	springen? sprengen?
springst? sprengst?	springt? sprengt?
springt	springen? sprengen?

5

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl		3rd sg		3rd pl		
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 7 3

springe? sprenge?	springen? sprengen?
springst? sprengst?	springt? sprengt?
springt	springen? sprengen?

5

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl		3rd sg		3rd pl		
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

9

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 7 3

springe? sprenge?	springen? sprengen?
springst? sprengst?	springt? sprengt?
springt	springen? sprengen?

5

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl		3rd sg		3rd pl		
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Lexicon & Corpus

treffe	treffen
triffst	trefft
trifft	treffen

9

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 7 3

springe? sprenge?	springen? sprengen?
springst? sprengst?	springt? sprengt?
springt	springen? sprengen?

5

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl		3rd sg		3rd pl		2nd sg
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2

We will now re-estimate θ, given our new "observations" (samples).
This training method is called MCEM.

treffe	treffen
triffst	trefft
trifft	treffen

9

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 7 3

springe? sprenge?	springen? sprengen?
springst? sprengst?	springt? sprengt?
springt	springen? sprengen?

5

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl		3rd sg		3rd pl		
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2

We will now re-estimate θ, given our new "observations" (samples).
This training method is called MCEM.

treffe	treffen
triffst	trefft
trifft	treffen

9

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 7 3

springe? sprenge?	springen? sprengen?
springst? sprengst?	springt? sprengt?
springt	springen? sprengen?

5

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl		3rd sg		3rd pl		2nd sg
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 We go over the corpus over and over again, **re-analyzing words** in the light of **newly acquired knowledge** about table frequencies, inflection frequencies and the updated "morphological grammar" θ.

treffe	treffen
triffst	trefft
trifft	treffen

9

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 7 3

springe? sprenge?	springen? sprengen?
springst? sprengst?	springt? sprengt?
springt	springen? sprengen?

5

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl		3rd sg		3rd pl		
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 We go over the corpus over and over again, **re-analyzing words** in the light of **newly acquired knowledge** about table frequencies, inflection frequencies and the updated "morphological grammar" θ.

treffe	treffen
triffst	trefft
trifft	treffen

9

briche? breche?	brichen? brechen?
brichst? brechst?	bricht? brecht?
bricht	brechen

1 7 3

springe? sprenge?	springen? sprengen?
springst? sprengst?	springt? sprengt?
springt	springen? sprengen?

5

Index	1	2	3	4	5	6	7	8	9
POS	VERB	PRON	VERB	NOUN	VERB	CONJ	VERB	NOUN	VERB
Lexeme									
Inflection	3rd sg		3rd pl		3rd sg		3rd pl		2nd sg
Spelling	bricht	...	brechen	...	springt	...	brechen	...	triffst

2 Text & Paradigms

Summary of the sampling process:

- Constantly update **frequency estimates** for lexemes and inflections
- Often update the "morphological grammar" θ
- Keep **re-analyzing words** accordingly
- Run **finite-state BP** to fill in missing paradigm cells
- **Important:** Often, BP will produce a regular and some more irregular candidates, one of them is found in the corpus and placed in the cell, so we "learn" it!

2 Text & Paradigms

Obtaining results for evaluation

- We add many paradigms, in which **only the lemma form** is given, but the other slots are empty.
- Just **keep track** of what corpus tokens the sampler places in those empty cells, or what candidates will be suggested from belief propagation.
- To get an answer for particular cell, get its **marginal** probability distribution at end of each iteration. At the end, get **average prob. per spelling** and report highest-scoring one

3 Results

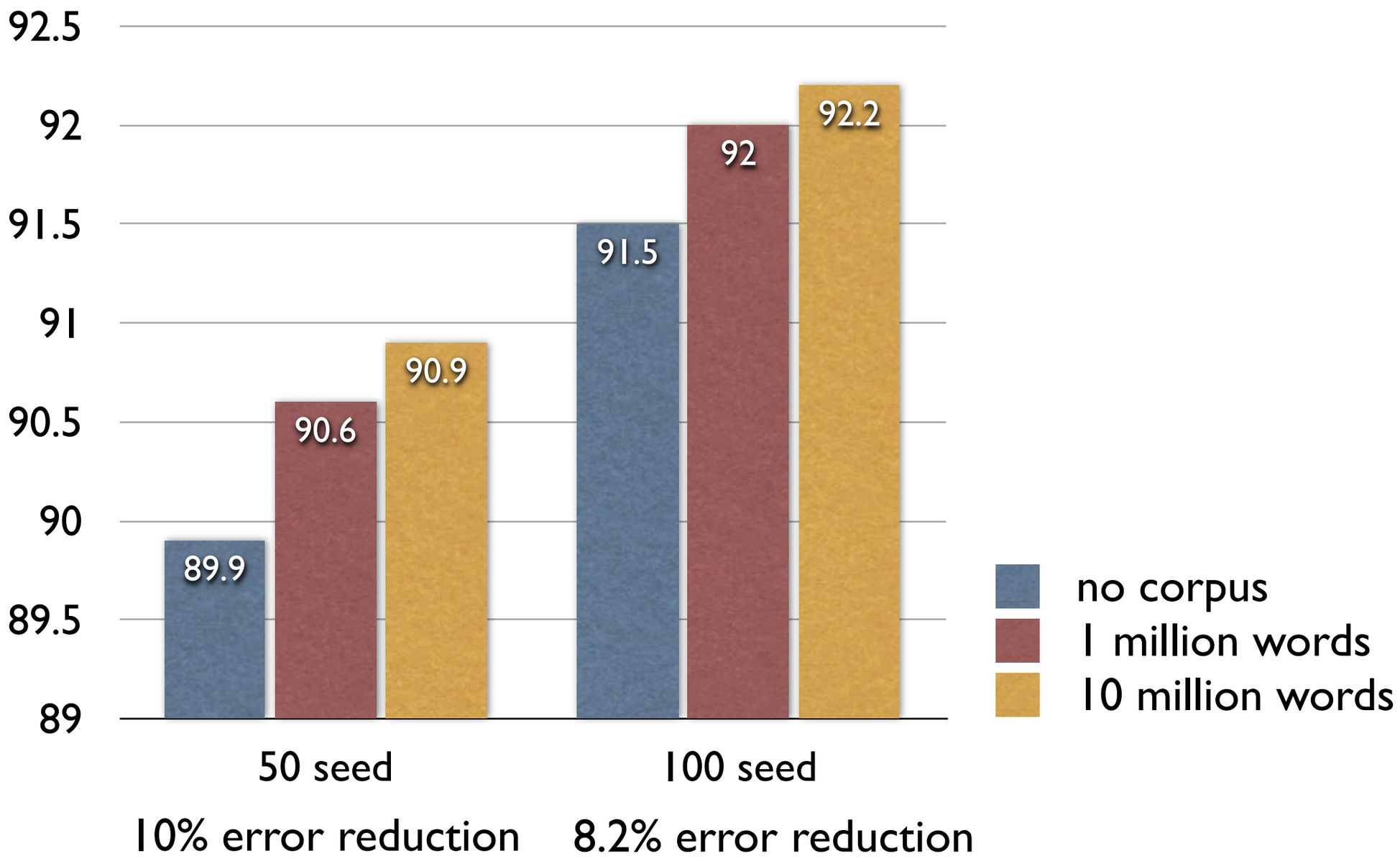
Experiment:
Learn German inflectional morphology

- Given:
  - 50 seed paradigms (from CELEX)
  - German corpus of 10 million words (from "WaCKy" corpus)
- Test:
  For 5,415 German verbs, predict paradigms with 21 inflections each

3

Results

Adding a large text corpus significantly improves prediction accuracy.

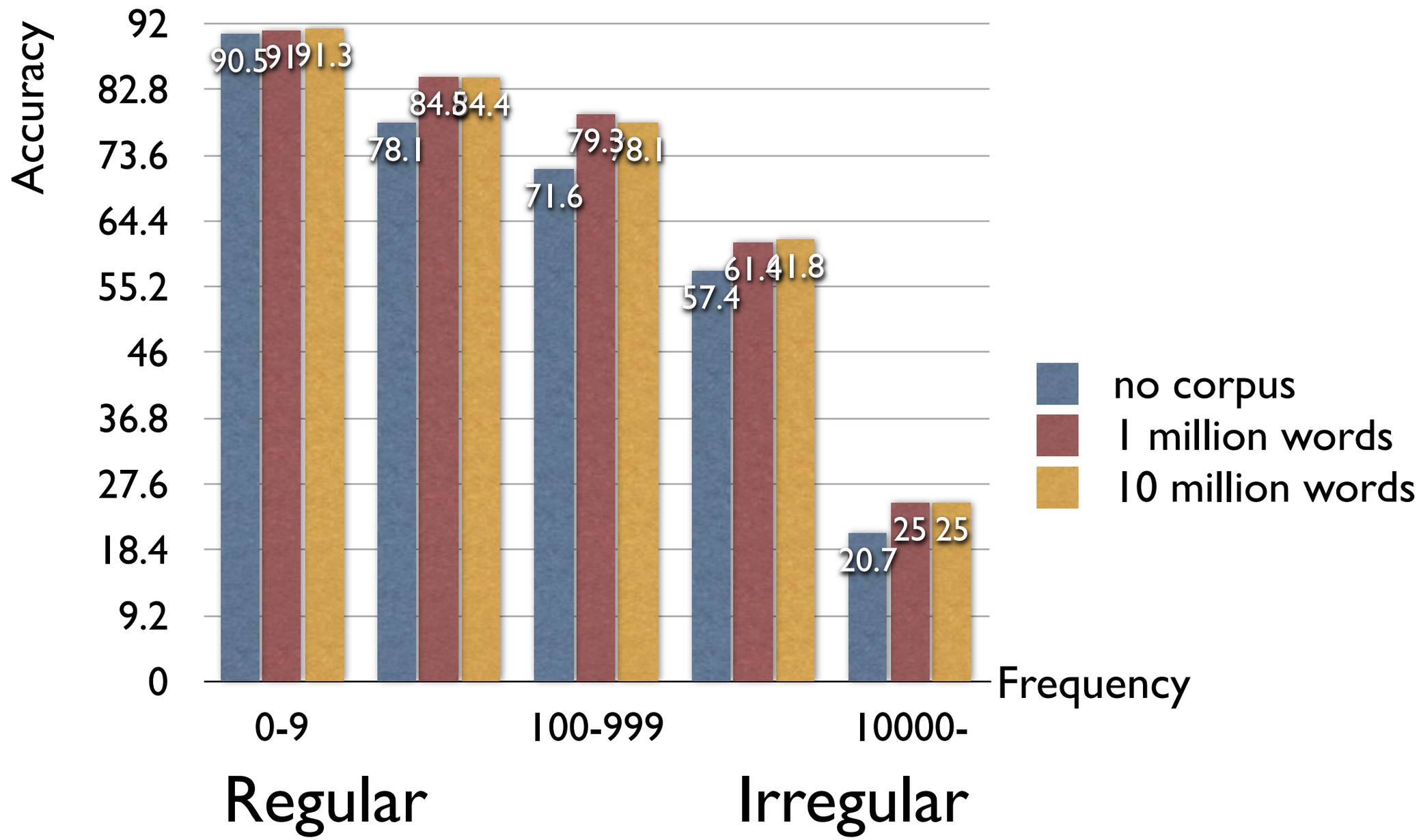


Regular: 85.4

3

Results

Large gains for irregular forms



Conclusions

- Formulated a **principled framework** for semi-supervised learning of structured morphological paradigms
- Jointly **tagged** corpus tokens and learned **non-concatenative spelling changes** between morphological types in the lexicon
- Filled complete structured paradigms with **observed and predicted** word forms
- Ran sampler on **large corpora** (up to 10 million words), reducing prediction error by up to 10%