

# Statistical NLP

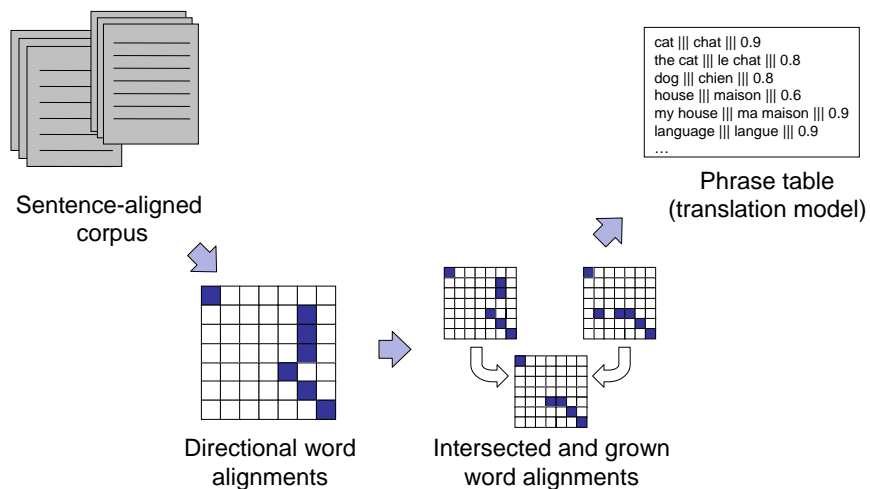
## Spring 2007



### Lecture 12: Phrase Decoding

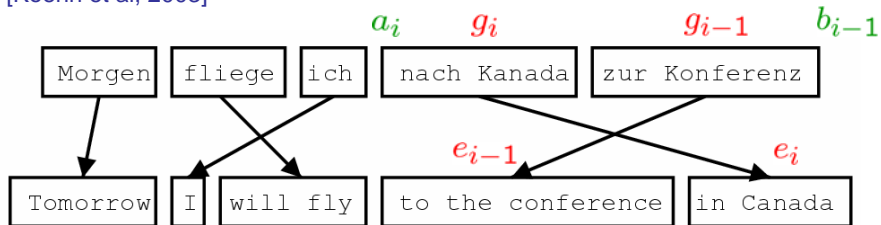
Dan Klein – UC Berkeley

## Overview: Extracting Phrases



# Pharaoh's Model

[Koehn et al, 2003]



$$P(e|g) = P(\{\bar{g}_i\}|g) \prod_i \phi(\bar{e}_i|\bar{g}_i) d(a_i - b_{i-1})$$

↙
↓
↘  
Segmentation
Translation
Distortion

# Pharaoh's Model

$$P(f|e) = P(\{\bar{e}_i\}|e) \prod_i \phi(\bar{f}_i|\bar{e}_i) d(a_i - b_{i-1})$$

↙
↓
↘  
 $\frac{1}{K}$ 
 $\frac{\text{count}(\bar{f}_i, \bar{e}_i)}{\text{count}(\bar{e}_i)}$ 
 $\alpha^{|a_i - b_{i-1}|}$

Where do we get these counts?

# Phrase-Based Decoding

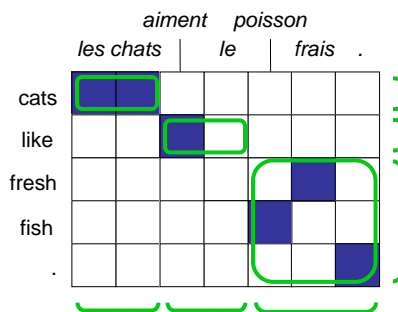
这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

the	7 people	including	by some		and	the russian	the	the astronauts	
it	7 people included		by france		and the	the russian		international astronautical	of rapporteur .
this	7 out	including the	from	the french			the fifth		.
these	7 among	including from		the french and	and to	of the russian	of	space	members .
that	7 persons	including from the		of france		russian	of the	aerospace	members .
	7 include		from the	of france and		russian		astronauts	. the
	7 numbers include		from france		and russian			of astronauts who	.
	7 populations include		those from france		and russian			astronauts .	
	7 deportees include		come from	france	and russia		in	astronautical	personnel ;
	7 philtrum	including those from		france and	and russia		a space		member
		including representatives from		france and the	and russia			astronaut	
		include	came from	france and russia			by cosmonauts		
		include representatives from		french	and russia			cosmonauts	
		include	came from france		and russia 's			cosmonauts .	
		includes	coming from	french and	russia 's			cosmonaut	
				french and	russian		's	astronavigation	member .
				french	and russia			astronauts	
					and russia 's				special rapporteur
					, and	russia			rapporteur
					, and russia				rapporteur .
					, and russia				
					or	russia 's			

Decoder design is important: [Koehn et al. 03]

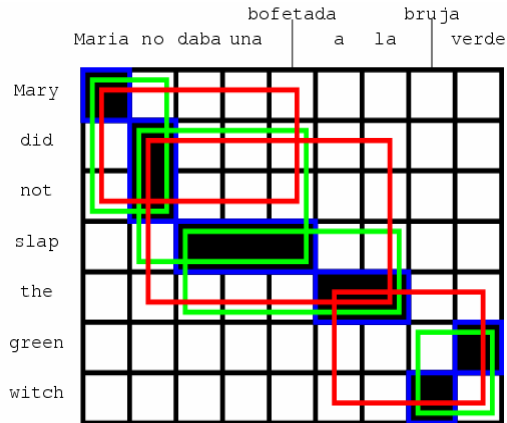
# Phrase Scoring

$$\phi_{new}(\bar{e}_j | \bar{f}_i) = \frac{c(\bar{f}_i, \bar{e}_j)}{c(\bar{f}_i)}$$



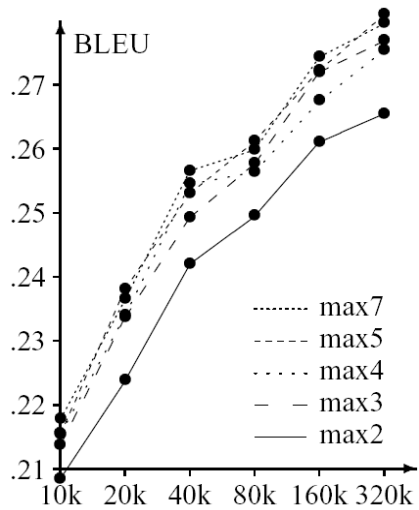
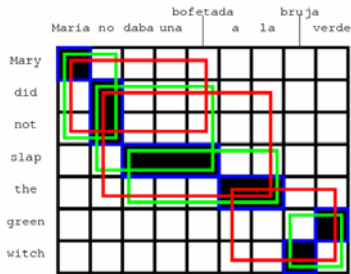
- Learning weights has been tried, several times:
  - [Marcu and Wong, 02]
  - [DeNero et al, 06]
  - ... and others
- Seems not to work, for a variety of only partially understood reasons
- Main issue: big chunks get all the weight, obvious priors don't help

# Extracting Phrases

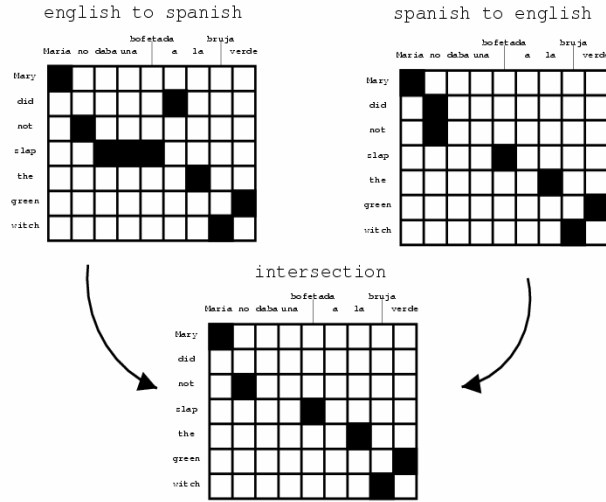


# Phrase Size

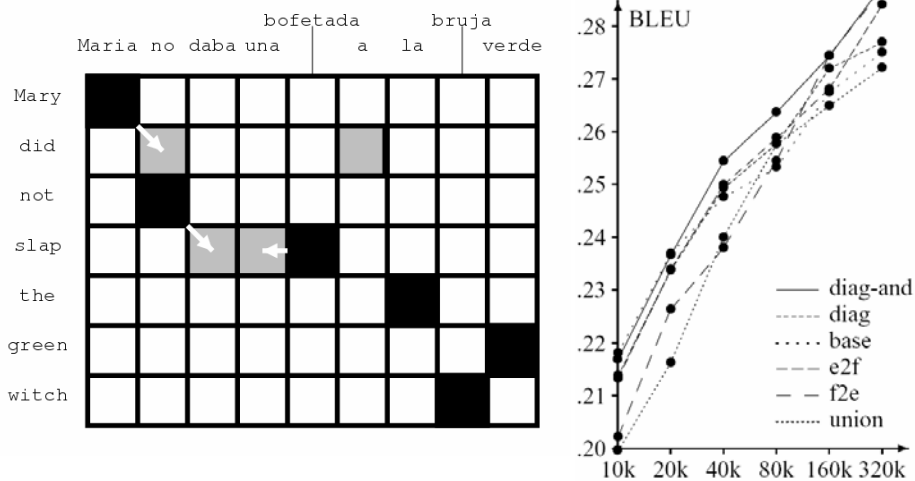
- Phrases do help
  - But they don't need to be long
  - Why should this be?



# Bidirectional Alignment

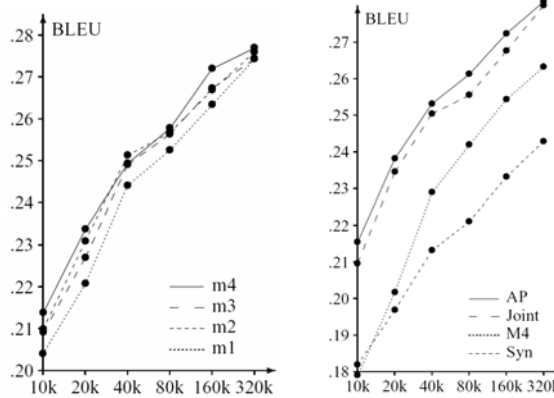


# Alignment Heuristics



# Sources of Alignments

Method	Training corpus size					
	10k	20k	40k	80k	160k	320k
AP	84k	176k	370k	736k	1536k	3152k
Joint	125k	220k	400k	707k	1254k	2214k
Syn	19k	24k	67k	105k	217k	373k

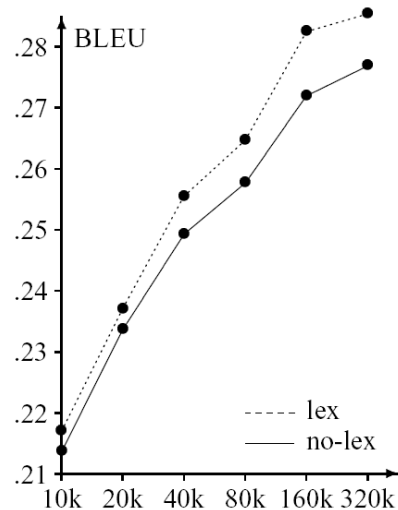


# Lexical Weighting

$$\phi(\bar{f}_i|\bar{e}_i) = \frac{\text{count}(\bar{f}_i, \bar{e}_i)}{\text{count}(\bar{e}_i)} p_w(\bar{f}_i|\bar{e}_i)$$

	f1	f2	f3
NULL	--	--	##
e1	##	--	--
e2	--	##	--
e3	--	##	--

$$\begin{aligned}
 p_w(\bar{f}|\bar{e}, a) &= p_w(f_1 f_2 f_3 | e_1 e_2 e_3, a) \\
 &= w(f_1|e_1) \\
 &\quad \times \frac{1}{2}(w(f_2|e_2) + w(f_2|e_3)) \\
 &\quad \times w(f_3|\text{NULL})
 \end{aligned}$$



# The Pharaoh Decoder

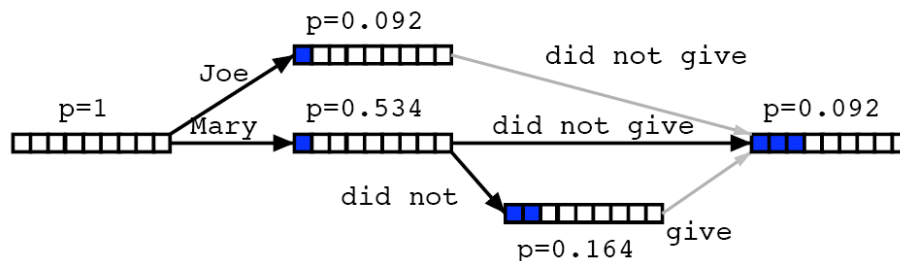
Maria	no	dio	una	bofetada	a	la	bruja	verde
Mary	not	give	a	slap	to	the	witch	green
	did not		a slap		by		green witch	
	no		slap		to the			
	did not give				to			
					the			
			slap			the witch		



- Probabilities at each step include LM and TM

# Hypothesis Lattices

Maria	no	dio	una	bofetada	a	la	bruja	verde
Mary	not	give	a	slap	to	the	witch	green
	did not		a slap		by		green witch	
	no		slap		to the			
	did not give				to			
					the			
			slap			the witch		



# Pruning

---

Maria no dio una bofetada a la bruja verde

\_\_\_\_\_

e: Mary did not  
f: \*\*-----  
p: 0.154

**better  
partial  
translation**

\_\_\_\_\_

e: the  
f: -----\*\*--  
p: 0.354

**covers  
easier part  
--> lower cost**

- **Problem: easy partial analyses are cheaper**
  - Solution 1: use beams per foreign subset
  - Solution 2: estimate forward costs (A\*-like)

# WSD?

---

- **Remember when we discussed WSD?**
  - Word-based MT systems rarely have a WSD step
  - Why not?

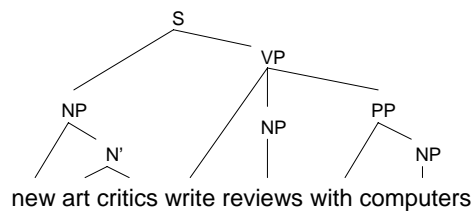
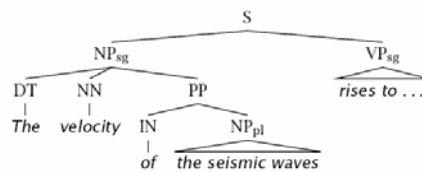


## What's Next?

- Modeling syntax
  - PCFGs and phrase structure
  - Syntactic parsing
  - Grammar induction
  - Syntactic language and translation models

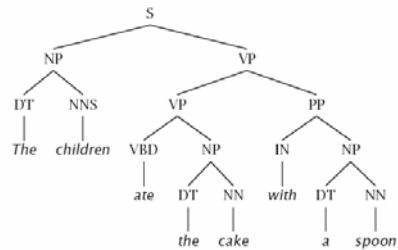
## Phrase Structure Parsing

- Phrase structure parsing organizes syntax into *constituents* or *brackets*
- In general, this involves nested trees
- Linguists can, and do, argue about details
- Lots of ambiguity
- Not the only kind of syntax...



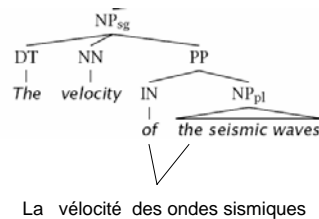
# Constituency Tests

- How do we know what nodes go in the tree?
- Classic constituency tests:
  - Substitution by *proform*
  - Question answers
  - Semantic reference
  - Dislocation
- Cross-linguistic arguments, too



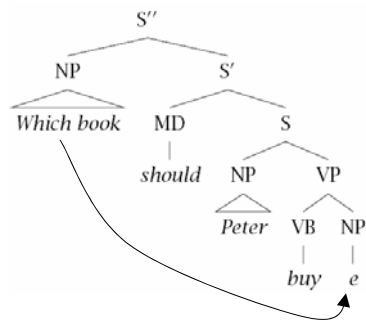
# Conflicting Tests

- Constituency isn't always clear
  - Units of transfer:
    - think about ~ penser à
    - talk about ~ hablar de
  - Phonological reduction:
    - I will go → I'll go
    - I want to go → I wanna go
    - a le centre → au centre



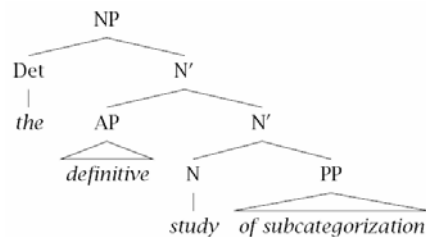
# Non-Local Phenomena

- Dislocation / gapping
  - Why did the postman think that the neighbors were home?
  - A debate arose which continued until the election.
- Binding
  - Reference
    - The IRS audits itself
  - Control
    - I want to go
    - I want you to go

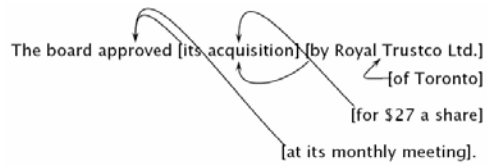
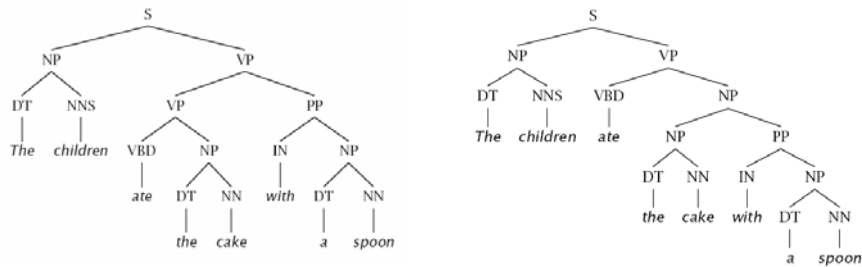


# Regularity of Rules

- Argumentation
- Adjunction
- Coordination
- X' Theory



# PP Attachment



# PP Attachment

V	N1	P	N2	Attachment
join	board	as	director	V
is	chairman	of	N.V.	N
using	crocidolite	in	filters	V
bring	attention	to	problem	V
is	asbestos	in	products	N
making	paper	for	filters	N
including	three	with	cancer	N

Method	Accuracy
Always noun attachment	59.0
Most likely for each preposition	72.2
Average Human (4 head words only)	88.2
Average Human (whole sentence)	93.2

## Attachment is a Simplification

- I cleaned the dishes from dinner
- I cleaned the dishes with detergent
- I cleaned the dishes in the sink

## Syntactic Ambiguities I

- **Prepositional phrases:**  
*They cooked the beans in the pot on the stove with handles.*
- **Particle vs. preposition:**  
*A good pharmacist dispenses with accuracy.  
The puppy tore up the staircase.*
- **Complement structures**  
*The tourists objected to the guide that they couldn't hear.  
She knows you like the back of her hand.*
- **Gerund vs. participial adjective**  
*Visiting relatives can be boring.  
Changing schedules frequently confused passengers.*

## Syntactic Ambiguities II

---

- **Modifier scope within NPs**  
*impractical design requirements*  
*plastic cup holder*
- **Multiple gap constructions**  
*The chicken is ready to eat.*  
*The contractors are rich enough to sue.*
- **Coordination scope:**  
*Small rats and mice can squeeze into holes or cracks in the wall.*

## Treebank Sentences

---

```
( (S (NP-SBJ The move)
  (VP followed
    (NP (NP a round)
      (PP of
        (NP (NP similar increases)
          (PP by
            (NP other lenders))
          (PP against
            (NP Arizona real estate loans))))))
    (S-ADV (NP-SBJ *)
      (VP reflecting
        (NP (NP a continuing decline)
          (PP-LOC in
            (NP that market))))))
  .))
```

# Human Processing

---

- Garden pathing:

the man who hunts ducks out on weekends

the cotton shirts are made from grows in Mississippi

the old train the young

the daughter of the king's son loves himself

- **Ambiguity maintenance**

Have the police . . . eaten their supper?

come in and look around.

taken out and shot.