

# CS388: Natural Language Processing

## Lecture 24: Multilinguality + Morphology



Greg Durrett



# Administrivia

---

- ▶ Final project presentations next week
  - ▶ See Canvas announcement for who is presenting when
  - ▶ Can be “work in progress”, but there should be at least preliminary results
  - ▶ Final reports due on December 14; no slip days
- ▶ Project 2 graded; average = 19.0



# Dealing with other languages

---

- ▶ Other languages present some problems not seen in English at all!
- ▶ Many algorithms so far have been developed for English
  - ▶ Some structures like constituency parsing don't make sense for other languages
  - ▶ Neural methods are typically tuned to English-scale resources, may not be the best for other languages where less data is available
- ▶ Question:
  - 1) What other phenomena / challenges do we need to solve?
  - 2) How can we leverage existing resources to do better in other languages without just annotating massive data?



# This Lecture

---

- ▶ Morphology: effects and challenges
- ▶ Morphology tasks: analysis, inflection, word segmentation
- ▶ Cross-lingual tagging and parsing

# Morphology



# What is morphology?

---

- ▶ Study of how words form
- ▶ Derivational morphology: create a new *lexeme* from a base
  - estrangle (v) => estrangement (n)
  - become (v) => unbecoming (adj)
    - ▶ May not be totally regular: enflame => inflammable
- ▶ Inflectional morphology: word is inflected based on its context
  - I become / she becomes
    - ▶ Mostly applies to verbs and nouns





# Morphological Inflection

- In English: I arrive      you arrive      he/she/it arrives      [X] arrived  
we arrive      you arrive      they arrive

- In French:

		singular			plural		
		first	second	third	first	second	third
indicative		je (j')	tu	il, elle	nous	vous	ils, elles
(simple tenses)	present	arrive /a.ʁiv/	arrives /a.ʁiv/	arrive /a.ʁiv/	arrivons /a.ʁi.vɔ̃/	arrivez /a.ʁi.ve/	arrivent /a.ʁiv/
	imperfect	arrivais /a.ʁi.vɛ/	arrivais /a.ʁi.vɛ/	arrivait /a.ʁi.vɛ/	arrivions /a.ʁi.vjɔ̃/	arriviez /a.ʁi.vje/	arrivaient /a.ʁi.vɛ/
	past historic <sup>2</sup>	arrivai /a.ʁi.vɛ/	arrivas /a.ʁi.va/	arriva /a.ʁi.va/	arrivâmes /a.ʁi.vam/	arrivâtes /a.ʁi.vat/	arrivèrent /a.ʁi.vɛʁ/
	future	arriverai /a.ʁi.vʁɛ/	arriveras /a.ʁi.vʁa/	arrivera /a.ʁi.vʁa/	arriverons /a.ʁi.vʁɔ̃/	arriverez /a.ʁi.vʁe/	arriveront /a.ʁi.vʁɔ̃/
	conditional	arriverais /a.ʁi.vʁɛ/	arriverais /a.ʁi.vʁɛ/	arriverait /a.ʁi.vʁɛ/	arriverions /a.ʁi.və.ʁjɔ̃/	arriveriez /a.ʁi.və.ʁje/	arriveraient /a.ʁi.vʁɛ/



# Morphological Inflection

## ► In Spanish:

		singular			plural		
		1st person	2nd person	3rd person	1st person	2nd person	3rd person
indicative		yo	tú vos	él/ella/ello usted	nosotros nosotras	vosotros vosotras	ellos/ellas ustedes
	present	llego	llegas <sup>tú</sup> llegás <sup>vos</sup>	llega	llegamos	llegáis	llegan
	imperfect	llegaba	llegabas	llegaba	llegábamos	llegabais	llegaban
	preterite	llegué	llegaste	llegó	llegamos	llegasteis	llegaron
	future	llegaré	llegarás	llegará	llegaremos	llegaréis	llegarán
	conditional	llegaría	llegarías	llegaría	llegaríamos	llegaríais	llegarían





# Noun Inflection

- ▶ Not just verbs either; gender, number, case complicate things

Declension of Kind <span>[hide ▲]</span>					
	singular			plural	
	indef.	def.	noun	def.	noun
<b>nominative</b>	ein	das	Kind	die	Kinder
<b>genitive</b>	eines	des	Kindes, Kinds	der	Kinder
<b>dative</b>	einem	dem	Kind, Kinde <sup>1</sup>	den	Kindern
<b>accusative</b>	ein	das	Kind	die	Kinder

- ▶ Nominative: I/he/she, accusative: me/him/her, genitive: mine/his/hers
- ▶ Dative: merged with accusative in English, shows recipient of something  
I taught the children <=> Ich unterrichte die Kinder  
I give the children a book <=> Ich gebe den Kindern ein Buch



# Irregular Inflection

---

- ▶ Common words are often irregular
  - ▶ I am / you are / she is
  - ▶ Je suis / tu es / elle est
  - ▶ Yo soy / usted está / ella es
- ▶ However, less common words typically fall into some regular *paradigm* — these are somewhat predictable



# Agglutinating Languages

- Finnish/Turkish/Hungarian (Finno-Ugric): what a preposition would do in English is instead part of the verb

		active	passive
1st		halata	
long 1st <sup>2</sup>		halatakseen	
2nd	inessive <sup>1</sup>	halatessa	halattaessa
	instructive	halaten	—
3rd	inessive	halaamassa	—
	elative	halaamasta	—
	illative	halaamaan	—
	adessive	halaamalla	—
	abessive	halaamatta	—
	instructive	halaaman	halattaman
	nominative	halaaminen	
4th	partitive	halaamista	
5th <sup>2</sup>		halaamaisillaan	

indicative mood	positive	negative	perfect	positive	negative
present tense	halata	ei halata	halonut	ole halannut	ei ole halannut
past tense	halasin	ei halannut	haloin	olisin halannut	ei olisi halannut
2nd sing.	halat	ei halata	halait	olisit halannut	ei olisit halannut
2nd plur.	halatte	ei halata	halaitte	olitte halannut	ei olitte halannut
3rd sing.	halaa	ei halaa	haloo	on halannut	ei ole halannut
3rd plur.	halavat	ei halaa	halovat	ovat halannut	ei ole halannut
conditional mood	positive	negative	perfect	positive	negative
present	halaisin	ei halaisi	halonut	olisin halannut	ei olisi halannut
past	halaisin	ei halaisi	halonut	olisin halannut	ei olisi halannut
2nd sing.	halaisit	ei halaisi	halaisit	olisit halannut	ei olisit halannut
2nd plur.	halaisitte	ei halaisi	halaisitte	olitte halannut	ei olitte halannut
3rd sing.	halaisi	ei halaisi	halaisi	olisi halannut	ei olisi halannut
3rd plur.	halaisivat	ei halaisi	halaisivat	olisivat halannut	ei olisivat halannut
imperative mood	positive	negative	perfect	positive	negative
present	halaa	älä halaa	halonut	ole halannut	älä ole halannut
past	halasin	älä halasit	haloin	olisin halannut	älä olisi halannut
2nd sing.	halaa	älä halaa	halait	olisit halannut	älä olisit halannut
2nd plur.	halatkaa	älä halatkaa	halaitte	olitte halannut	älä olitte halannut
3rd sing.	halakoon	älä halakoon	haloo	on halannut	älä ole halannut
3rd plur.	halakoot	älä halakoot	halovat	ovat halannut	älä ole halannut
optative mood	positive	negative	perfect	positive	negative
present	halaisi	ei halaisi	halonut	olisin halannut	ei olisi halannut
past	halaisi	ei halaisi	halonut	olisin halannut	ei olisi halannut
2nd sing.	halaisit	ei halaisi	halaisit	olisit halannut	ei olisit halannut
2nd plur.	halaisitte	ei halaisi	halaisitte	olitte halannut	ei olitte halannut
3rd sing.	halaisi	ei halaisi	halaisi	olisi halannut	ei olisi halannut
3rd plur.	halaisivat	ei halaisi	halaisivat	olisivat halannut	ei olisivat halannut
participle forms	positive	negative	perfect	positive	negative
present	halaa	ei halaa	halonut	ole halannut	ei ole halannut
past	halasin	ei halasin	haloin	olisin halannut	ei olisi halannut
2nd sing.	halat	ei halat	halait	olisit halannut	ei olisit halannut
2nd plur.	halatte	ei halatte	halaitte	olitte halannut	ei olitte halannut
3rd sing.	halaa	ei halaa	haloo	on halannut	ei ole halannut
3rd plur.	halavat	ei halavat	halovat	ovat halannut	ei ole halannut
infinitive	halata	ei halata	halonut	ole halannut	ei ole halannut
2nd sing.	halat	ei halat	halait	olisit halannut	ei olisit halannut
2nd plur.	halatte	ei halatte	halaitte	olitte halannut	ei olitte halannut
3rd sing.	halaa	ei halaa	haloo	on halannut	ei ole halannut
3rd plur.	halavat	ei halavat	halovat	ovat halannut	ei ole halannut
supine	halattu	ei halattu	halonut	ole halannut	ei ole halannut
2nd sing.	halat	ei halat	halait	olisit halannut	ei olisit halannut
2nd plur.	halatte	ei halatte	halaitte	olitte halannut	ei olitte halannut
3rd sing.	halaa	ei halaa	haloo	on halannut	ei ole halannut
3rd plur.	halavat	ei halavat	halovat	ovat halannut	ei ole halannut
gerund	halattaessa	ei halattaessa	halonut	ole halannut	ei ole halannut
2nd sing.	halatessa	ei halatessa	halait	olisit halannut	ei olisit halannut
2nd plur.	halatessaa	ei halatessaa	halaitte	olitte halannut	ei olitte halannut
3rd sing.	halatessa	ei halatessa	haloo	on halannut	ei ole halannut
3rd plur.	halatessa	ei halatessa	halovat	ovat halannut	ei ole halannut
adpositional forms	positive	negative	perfect	positive	negative
present	halaa	ei halaa	halonut	ole halannut	ei ole halannut
past	halasin	ei halasin	haloin	olisin halannut	ei olisi halannut
2nd sing.	halat	ei halat	halait	olisit halannut	ei olisit halannut
2nd plur.	halatte	ei halatte	halaitte	olitte halannut	ei olitte halannut
3rd sing.	halaa	ei halaa	haloo	on halannut	ei ole halannut
3rd plur.	halavat	ei halavat	halovat	ovat halannut	ei ole halannut
infinitive	halata	ei halata	halonut	ole halannut	ei ole halannut
2nd sing.	halat	ei halat	halait	olisit halannut	ei olisit halannut
2nd plur.	halatte	ei halatte	halaitte	olitte halannut	ei olitte halannut
3rd sing.	halaa	ei halaa	haloo	on halannut	ei ole halannut
3rd plur.	halavat	ei halavat	halovat	ovat halannut	ei ole halannut
supine	halattu	ei halattu	halonut	ole halannut	ei ole halannut
2nd sing.	halat	ei halat	halait	olisit halannut	ei olisit halannut
2nd plur.	halatte	ei halatte	halaitte	olitte halannut	ei olitte halannut
3rd sing.	halaa	ei halaa	haloo	on halannut	ei ole halannut
3rd plur.	halavat	ei halavat	halovat	ovat halannut	ei ole halannut
gerund	halattaessa	ei halattaessa	halonut	ole halannut	ei ole halannut
2nd sing.	halatessa	ei halatessa	halait	olisit halannut	ei olisit halannut
2nd plur.	halatessaa	ei halatessaa	halaitte	olitte halannut	ei olitte halannut
3rd sing.	halatessa	ei halatessa	haloo	on halannut	ei ole halannut
3rd plur.	halatessa	ei halatessa	halovat	ovat halannut	ei ole halannut
adpositional forms	positive	negative	perfect	positive	negative
present	halaa	ei halaa	halonut	ole halannut	ei ole halannut
past	halasin	ei halasin	haloin	olisin halannut	ei olisi halannut
2nd sing.	halat	ei halat	halait	olisit halannut	ei olisit halannut
2nd plur.	halatte	ei halatte	halaitte	olitte halannut	ei olitte halannut
3rd sing.	halaa	ei halaa	haloo	on halannut	ei ole halannut
3rd plur.	halavat	ei halavat	halovat	ovat halannut	ei ole halannut
infinitive	halata	ei halata	halonut	ole halannut	ei ole halannut
2nd sing.	halat	ei halat	halait	olisit halannut	ei olisit halannut
2nd plur.	halatte	ei halatte	halaitte	olitte halannut	ei olitte halannut
3rd sing.	halaa	ei halaa	haloo	on halannut	ei ole halannut
3rd plur.	halavat	ei halavat	halovat	ovat halannut	ei ole halannut
supine	halattu	ei halattu	halonut	ole halannut	ei ole halannut
2nd sing.	halat	ei halat	halait	olisit halannut	ei olisit halannut
2nd plur.	halatte	ei halatte	halaitte	olitte halannut	ei olitte halannut
3rd sing.	halaa	ei halaa	haloo	on halannut	ei ole halannut
3rd plur.	halavat	ei halavat	halovat	ovat halannut	ei ole halannut
gerund	halattaessa	ei halattaessa	halonut	ole halannut	ei ole halannut
2nd sing.	halatessa	ei halatessa	halait	olisit halannut	ei olisit halannut
2nd plur.	halatessaa	ei halatessaa	halaitte	olitte halannut	ei olitte halannut
3rd sing.	halatessa	ei halatessa	haloo	on halannut	ei ole halannut
3rd plur.	halatessa	ei halatessa	halovat	ovat halannut	ei ole halannut

illative: “into”

adessive: “on”

- Many possible forms — and in newswire data, only a few are observed



# Morphologically-Rich Languages

---

- ▶ Many languages spoken all over the world have much richer morphology than English (Chinese is the main exception)
- ▶ CoNLL 2006 / 2007: dependency parsing + morphological analyses for ~15 mostly Indo-European languages
- ▶ SPMRL shared tasks (2013-2014): Syntactic Parsing of Morphologically-Rich Languages
- ▶ Word piece / byte-pair encoding models for MT are pretty good at handling these if there's enough data





# Morphologically-Rich Languages



MORGAN & CLAYPOOL PUBLISHERS

## Linguistic Fundamentals for Natural Language Processing

*100 Essentials from  
Morphology and Syntax*

Emily M. Bender

**SYNTHESIS LECTURES ON  
HUMAN LANGUAGE TECHNOLOGIES**

Graeme Hirst, *Series Editor*

- ▶ Great resources for challenging your assumptions about language and for understanding multilingual models!

# Morphological Analysis/Inflection





# Morphological Analysis

---

- ▶ In English, not that many word forms, lexical features on words and word vectors are pretty effective
- ▶ In other languages, \*lots\* more unseen words! Affects parsing, translation, ...
- ▶ When we're building systems, we probably want to know base form + morphological features explicitly
- ▶ How to do this kind of *morphological analysis*?



# Morphological Analysis

But the government does not recommend reducing taxes.

Ám a kormány egyetlen adó csökkentését sem javasolja .

n=singular / case=nominative / proper=no  
deg=positive / n=singular / case=nominative  
n=singular / case=nominative / proper=no  
n=singular / case=accusative / proper=no / pperson=3rd / pnumber=singular  
mood=indicative / t=present / p=3rd / n=singular / def=yes

► Why is this useful?



# Morphological Analysis

---

- ▶ Given a word, need to recognize what its morphological features are
- ▶ Basic approach:
  - ▶ Lexicon: tells you what possibilities are
  - ▶ Analyzer: statistical model that disambiguates
- ▶ Models are largely CRF-like: score morphological features in context
- ▶ Lots of work on Arabic inflection (high amounts of ambiguity)



# Predicting Inflection

- ▶ Other direction: given base form + features, inflect the word
  - ▶ Hard for unknown words — need models that generalize

*w i n d e n* →

conjugation of winden						[hide ▲]
infinitive			winden			
present participle			windend			
past participle			gewunden			
auxiliary			haben			
	indicative			subjunctive		
present	ich winde	wir winden	i	ich winde	wir winden	
	du windest	ihr windet		du windest	ihr windet	
	er windet	sie winden		er winde	sie winden	
preterite	ich wand	wir wanden	ii	ich wände	wir wänden	
	du wandest	ihr wandet		du wändest	ihr wändet	
	er wand	sie wanden		er wände	sie wänden	
imperative	winde (du)	windet (ihr)				
composed forms of winden						[show ▼]

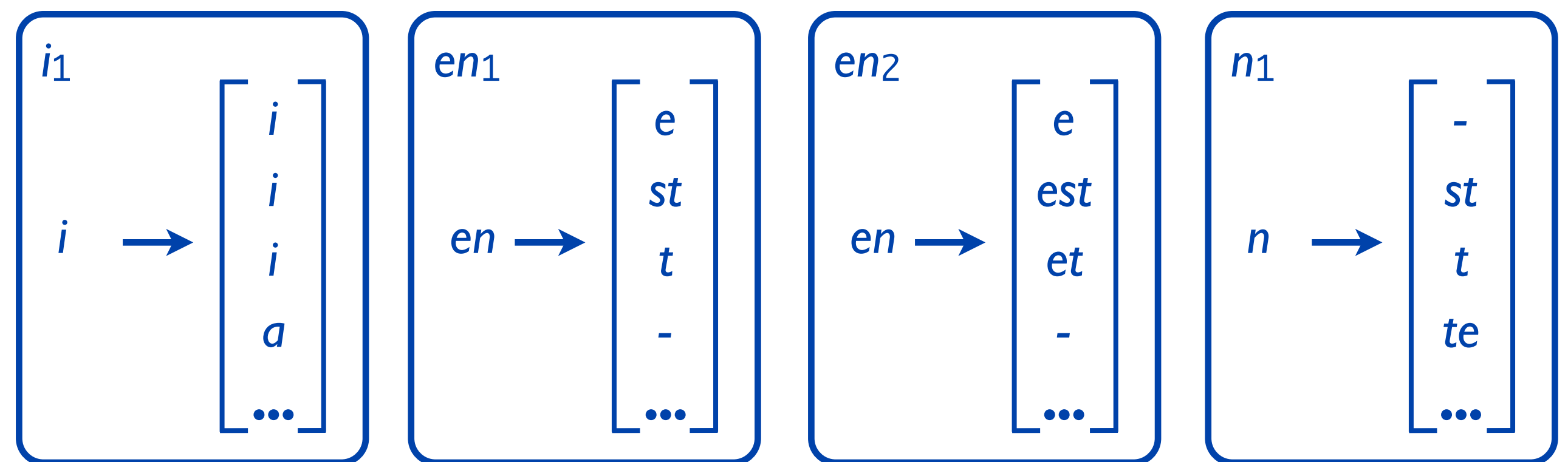
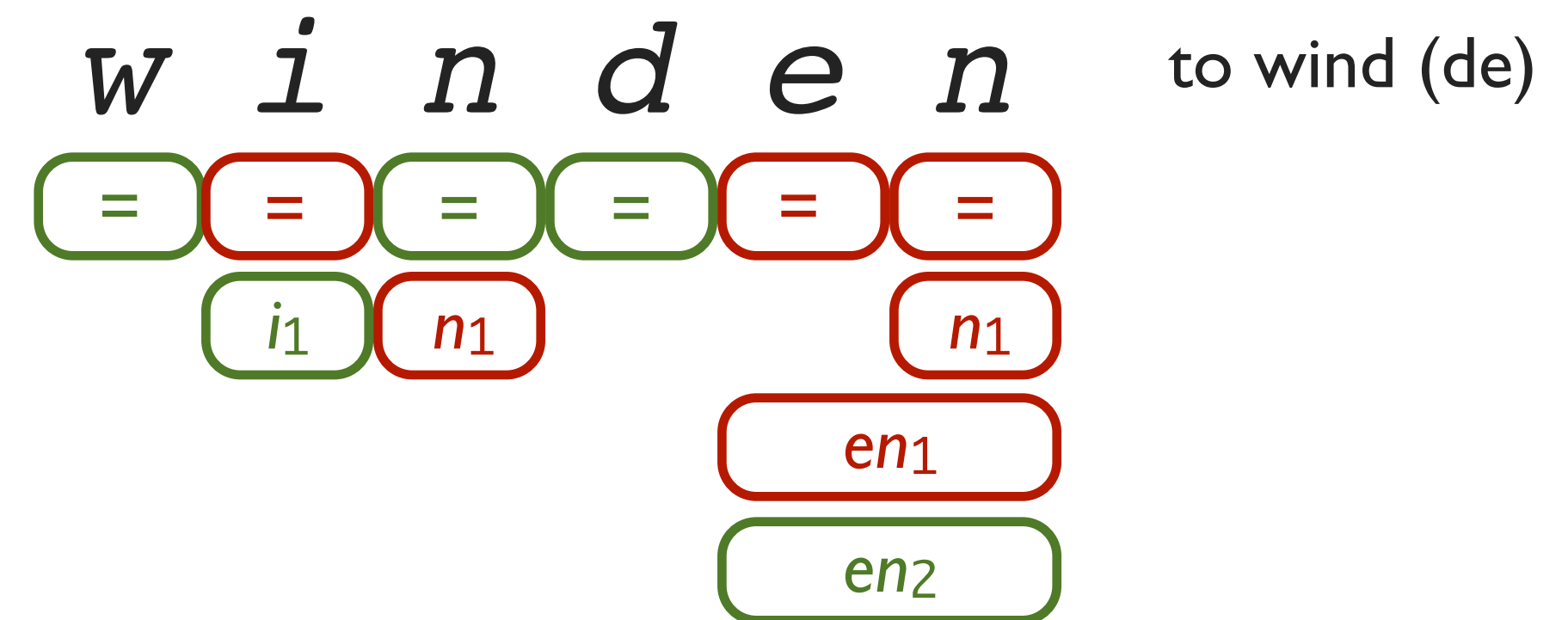


# Predicting Inflection

- ▶ Other direction: given base form + features, inflect the word
  - ▶ Hard for unknown words — need models that generalize

- ▶ Take a bunch of existing verbs from Wiktionary, extract these change rules using character alignments

- ▶ Train a CRF with character n-gram context features to learn where to apply them







она **пыталась** пересечь пути на ее велосипеде  
 she had attempted to cross the road on her bike  
 C50 C473 C28 C8 C275 C37 C43 C82 C94 C331  
 PRP VBD VBN TO VB DT NN IN PRP\$ NN  
 aux  
 nsubj root xcomp

- ▶ Machine translation where phrase table is defined in terms of lemmas
- ▶ “Translate-and-inflect”: translate into uninflected words and predict inflection based on source side

# Chahuneau et al. (2013)



# Word Segmentation



# Morpheme Segmentation

---

- ▶ Can we do something unsupervised rather than these complicated analyses?
- ▶ unbecoming => un+becom+ing — we should be able to recognize these common pieces and split them off
- ▶ How do we do this?



# Morpheme Segmentation

- ▶ Simple probabilistic model  $\text{Cost}(\text{Source text}) = \sum_{\text{morph tokens}} -\log p(m_i)$
- ▶  $p(m_i) = \text{count}(\text{token}) / \text{count}(\text{all tokens})$

- ▶ Train with EM: E-step involves estimating best segmentation with Viterbi, M-step: collect token counts

*allowed expected need needed all+owe+d expe+cted n+e+ed ne+ed+ed E0*

M0: ed has count 3 *all+ow+ed expect+ed ne+ed ne+ed+ed E1*

- ▶ Some heuristics: reject rare morphemes, one-letter morphemes
- ▶ Doesn't handle stem changes: becoming => becom + ing

Creutz and Lagus (2002)



# Chinese Word Segmentation

- ▶ Some languages including Chinese are totally untokenized
- ▶ LSTMs over character embeddings / character bigram embeddings to predict word boundaries
- ▶ Having the right segmentation can help machine translation

冬天 (winter), 能 (can) 穿 (wear) 多少 (amount) 穿 (wear) 多少 (amount); 夏天 (summer), 能 (can) 穿 (wear) 多 (more) 少 (little) 穿 (wear) 多 (more) 少 (little)。

Without the word “夏天 (summer)” or “冬天 (winter)”, it is difficult to segment the phrase “能穿多少穿多少”.

- separating nouns and pre-modifying adjectives:  
高血压 (*high blood pressure*)  
→ 高(*high*) 血压(*blood pressure*)
- separating compound nouns:  
民政部 (*Department of Internal Affairs*)  
→ 民政(*Internal Affairs*) 部(*Department*).

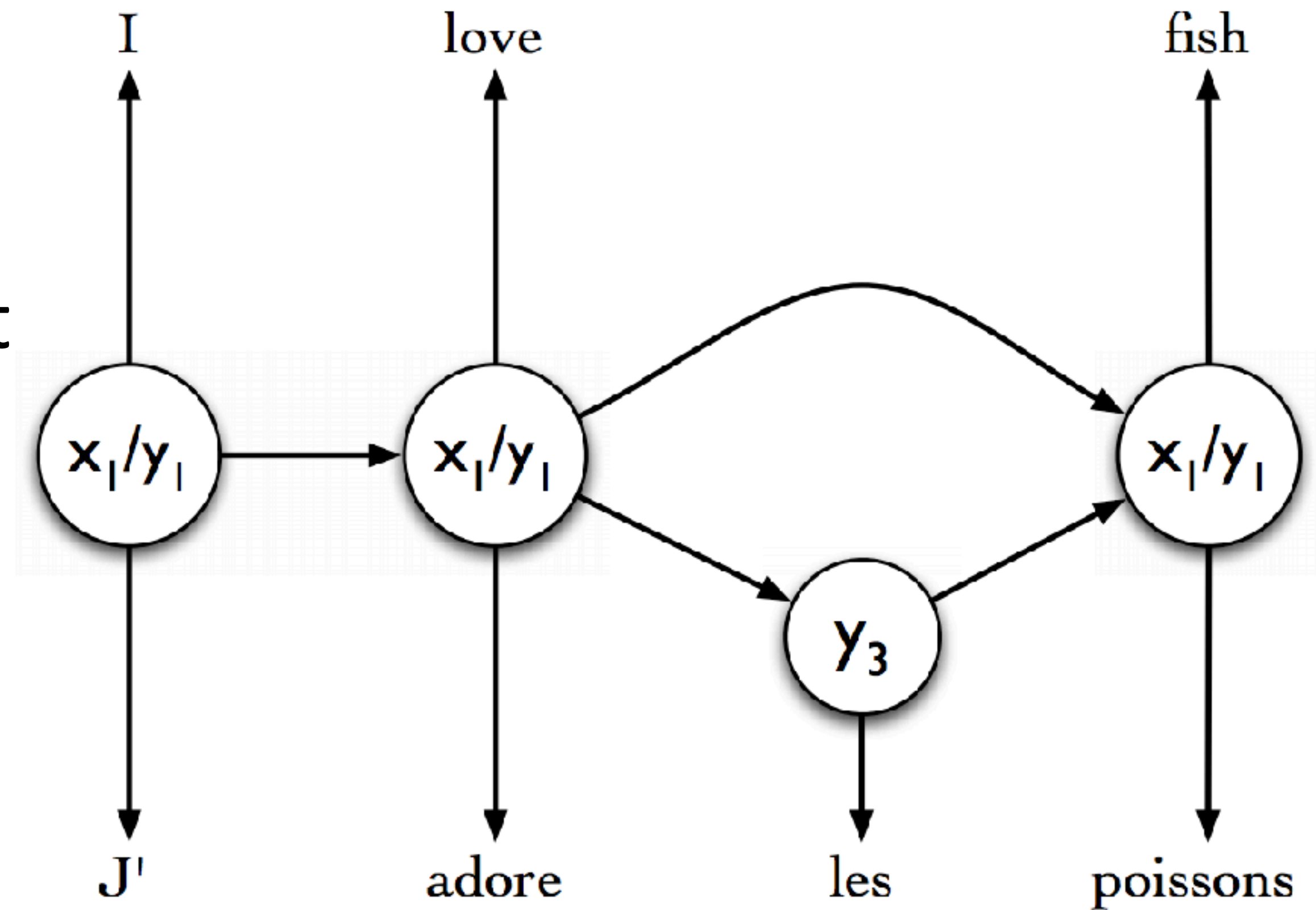
# Cross-Lingual Tagging and Parsing





# Cross-Lingual Tagging

- ▶ Multilingual POS induction
- ▶ Generative model of two languages simultaneously, joint alignment + tag learning
- ▶ Complex generative model, requires Gibbs sampling for inference



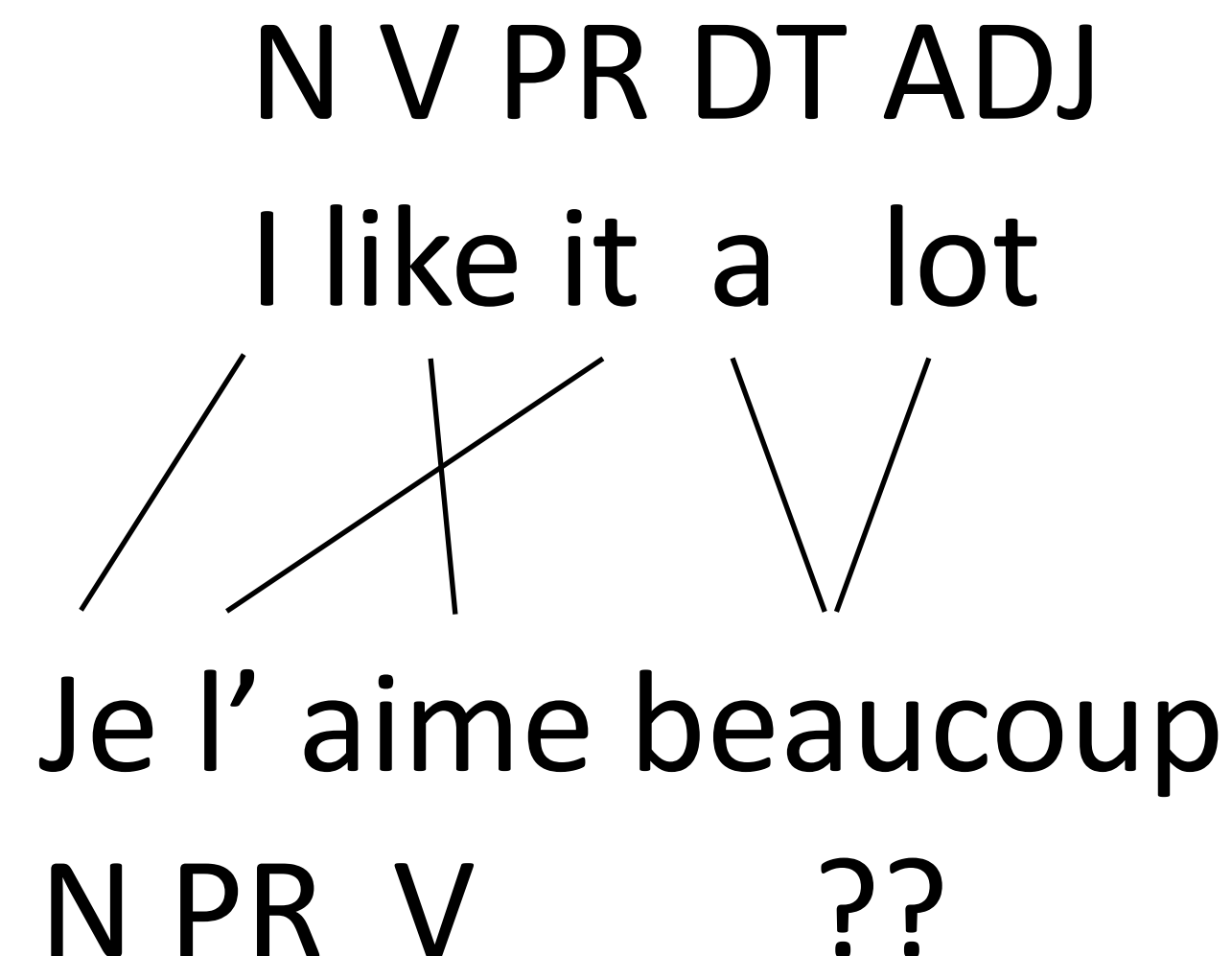
Snyder et al. (2008)





# Cross-Lingual Tagging

- ▶ We have resources for languages like English — can we use these more directly?

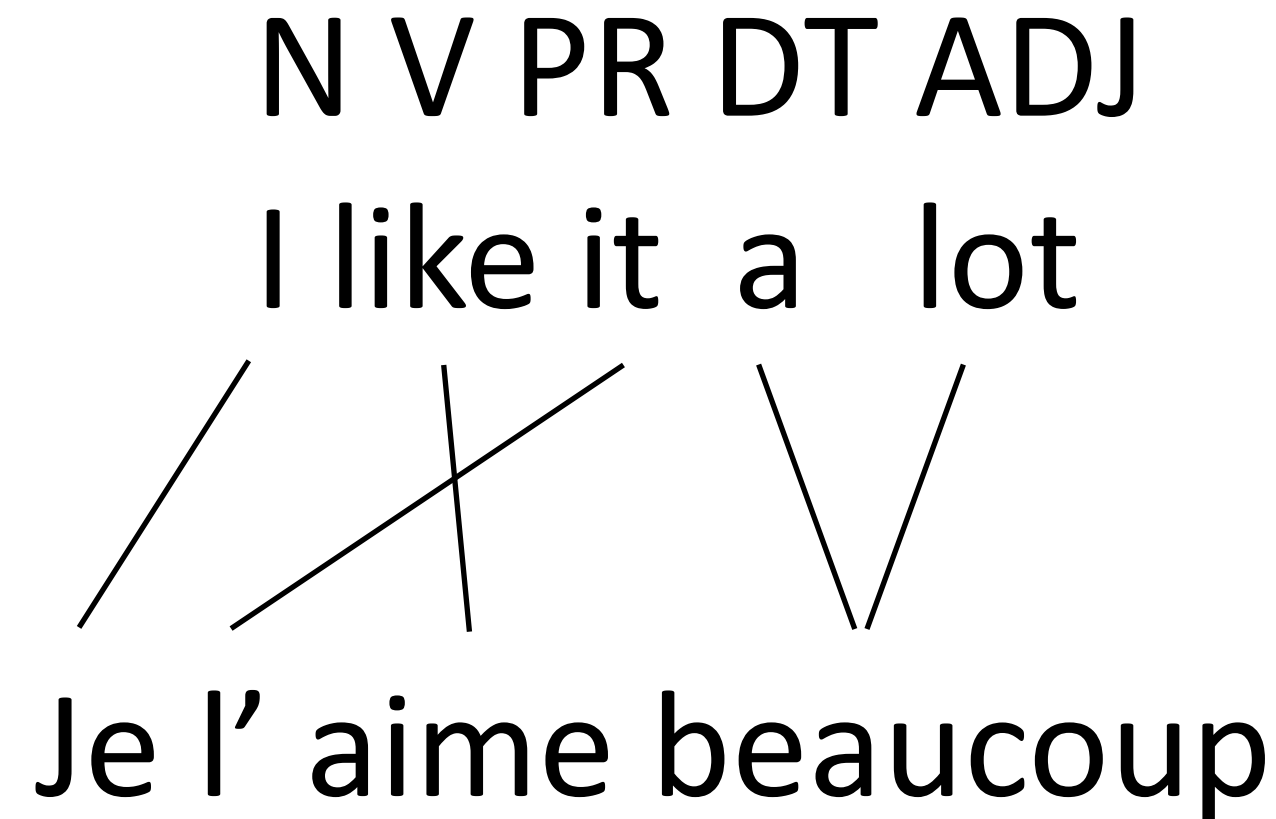


- ▶ Tag with English tagger, project across bitext, train French tagger?
- ▶ Can do something smarter

Das and Petrov (2011)



# Cross-Lingual Tagging



- Form a graph of trigrams, use these to propagate knowledge about tags

---

## Algorithm 1 Bilingual POS Induction

---

**Require:** Parallel English and foreign language data  $\mathcal{D}^e$  and  $\mathcal{D}^f$ , unlabeled foreign training data  $\Gamma^f$ ; English tagger.

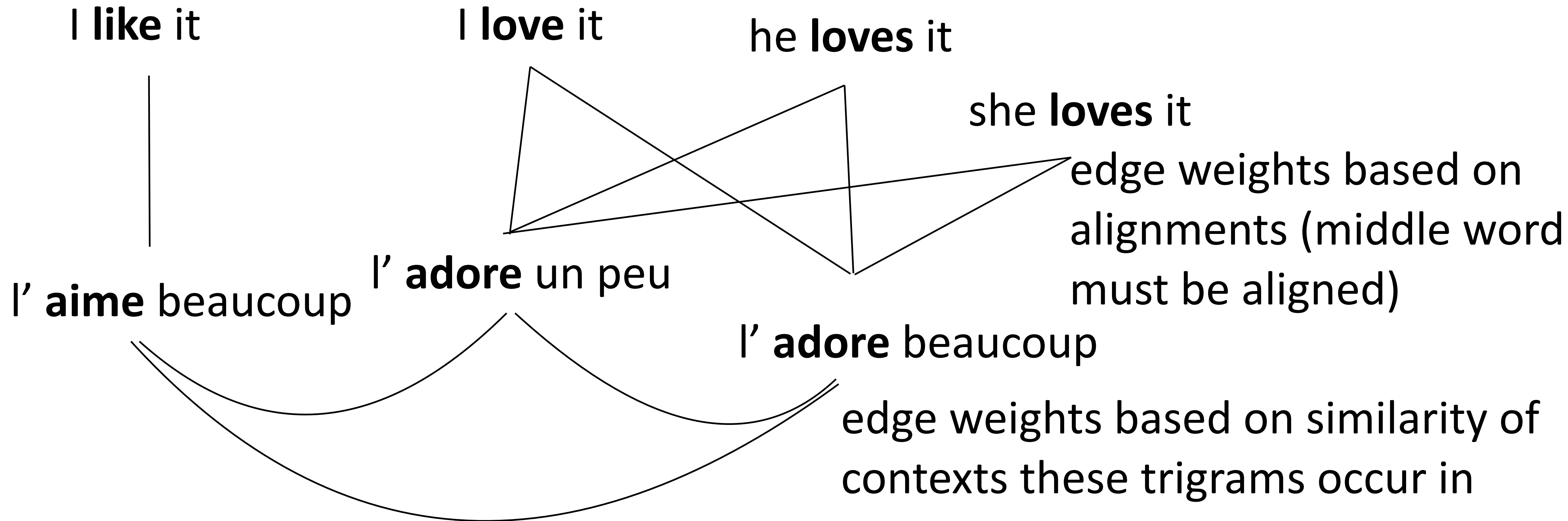
**Ensure:**  $\Theta^f$ , a set of parameters learned using a constrained unsupervised model (§5).

- 1:  $\mathcal{D}^{e \leftrightarrow f} \leftarrow \text{word-align-bitext}(\mathcal{D}^e, \mathcal{D}^f)$
  - 2:  $\widehat{\mathcal{D}}^e \leftarrow \text{pos-tag-supervised}(\mathcal{D}^e)$
  - 3:  $\mathcal{A} \leftarrow \text{extract-alignments}(\mathcal{D}^{e \leftrightarrow f}, \widehat{\mathcal{D}}^e)$
  - 4:  $G \leftarrow \text{construct-graph}(\Gamma^f, \mathcal{D}^f, \mathcal{A})$
  - 5:  $\tilde{G} \leftarrow \text{graph-propagate}(G)$
  - 6:  $\Delta \leftarrow \text{extract-word-constraints}(\tilde{G})$
  - 7:  $\Theta^f \leftarrow \text{pos-induce-constrained}(\Gamma^f, \Delta)$
  - 8: Return  $\Theta^f$
-



# Cross-Lingual Tagging

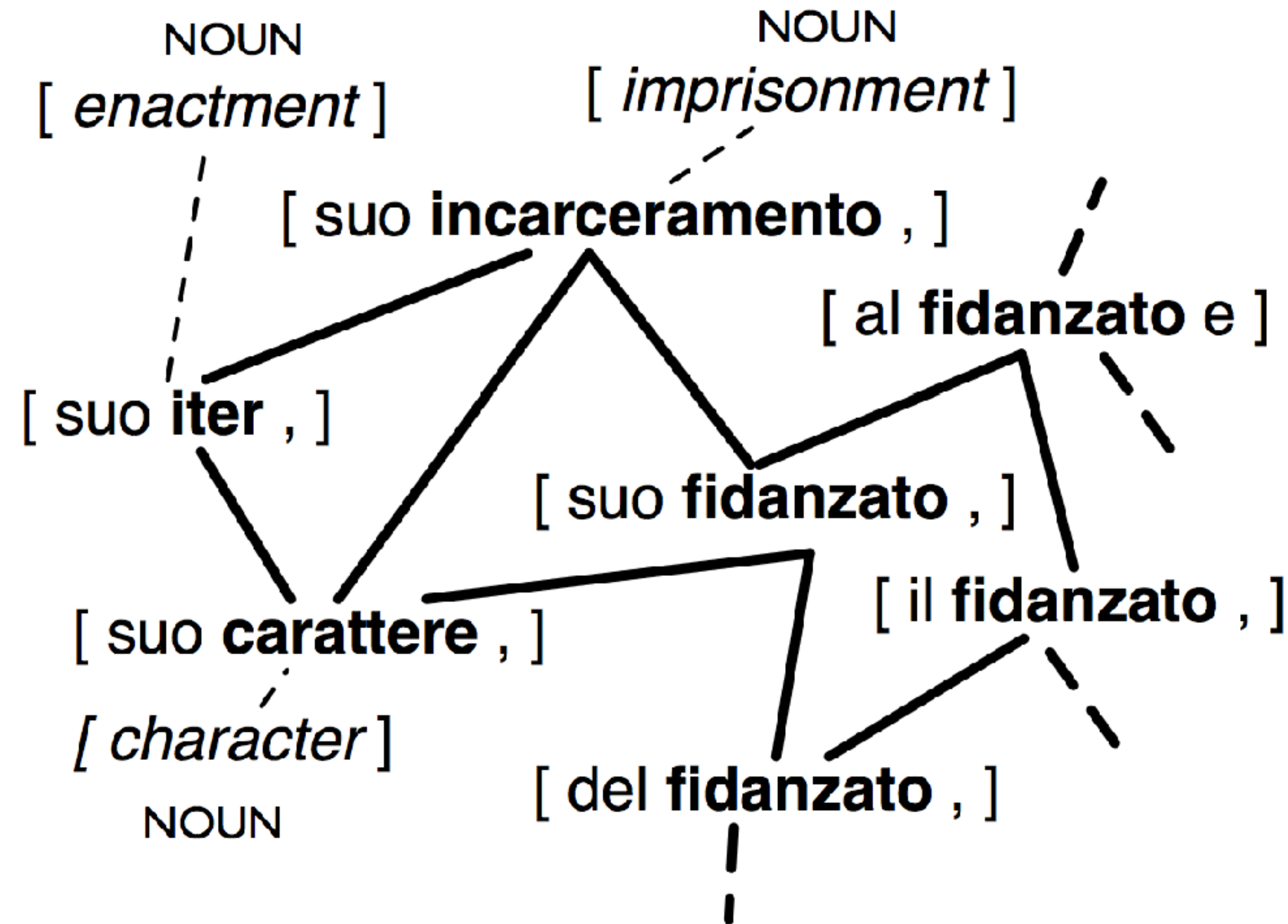
Das and Petrov (2011)



- ▶ Each node is associated with a distribution over tags, label propagation updates these using the graph



# Cross-Lingual Tagging







# Cross-Lingual Tagging

---

- ▶ Label propagation: encourages nodes with higher-weight edges between them to have similar tags
- ▶ Take these trigrams and treat them as “soft training examples” and learn an HMM tagger
- ▶ Prune to only keep tags above some probability to get the lexicon (valid tag-word pairs)



# Cross-Lingual Tagging

	Model	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Avg
<i>baselines</i>	EM-HMM	68.7	57.0	75.9	65.8	63.7	62.9	71.5	68.4	66.7
	Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
	Projection	73.6	77.0	83.2	79.3	79.7	82.6	80.1	74.7	78.8
<i>our approach</i>	No LP	79.0	78.8	82.4	76.3	84.8	87.0	82.8	79.4	81.3
	With LP	<b>83.2</b>	<b>79.5</b>	82.8	<b>82.5</b>	<b>86.8</b>	<b>87.9</b>	<b>84.2</b>	<b>80.5</b>	83.4
<i>oracles</i>	TB Dictionary	93.1	94.7	93.5	96.6	96.4	94.0	95.8	85.5	93.7
	Supervised	96.9	94.9	98.2	97.8	95.8	97.2	96.8	94.8	96.6

- ▶ EM-HMM/feature HMM: unsupervised methods with a greedy mapping from learned tags to gold tags
- ▶ Projection: project tags across bitext to make pseudogold corpus, train on that

Das and Petrov (2011)





# Cross-Lingual Parsing

- ▶ Now that we can POS tag other languages, can we parse them too?
- ▶ Direct transfer: train a parser over POS sequences in one language, then apply it to another language

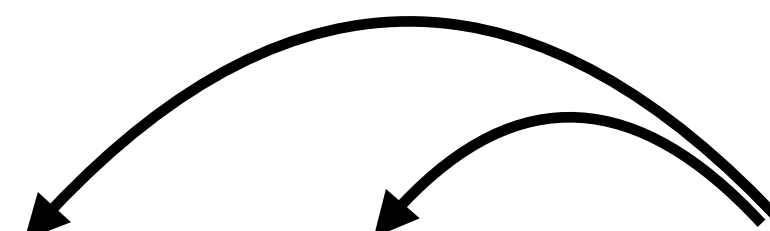
PRON VERB NOUN  
I like tomatoes



PRON VERB PRON  
I like them

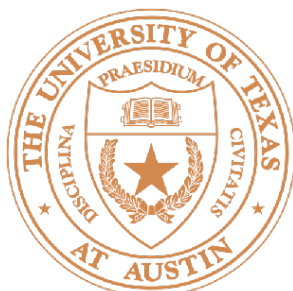


PRON PRON VERB  
Je les aime



- ▶ Even though we've never seen this sequence in English **and** don't know the words, we can still figure it out

McDonald et al. (2011)



# Cross-Lingual Parsing

	best-source		avg-source gold-POS	gold-POS		pred-POS	
	source	gold-POS		multi-dir.	multi-proj.	multi-dir.	multi-proj.
da	it	48.6	46.3	48.9	49.5	46.2	47.5
de	nl	55.8	48.9	56.7	56.6	51.7	52.0
el	en	63.9	51.7	60.1	65.1	58.5	63.0
es	it	68.4	53.2	64.2	64.5	55.6	56.5
it	pt	69.1	58.5	64.1	65.0	56.8	58.9
nl	el	62.1	49.9	55.8	65.7	54.3	64.4
pt	it	74.8	61.6	74.0	75.6	67.7	70.3
sv	pt	66.8	54.8	65.3	68.0	58.3	62.1
avg		63.7	51.6	61.1	63.8	56.1	59.3

► Multi-dir: transfer a parser trained on several source treebanks to the target language

► Multi-proj: more complex annotation projection approach

McDonald et al. (2011)



# Cross-Lingual Embeddings

---

- ▶ Learn a shared multilingual embedding space so *any* neural system can transfer over
- ▶ multiCluster: use bilingual dictionaries to form clusters of words that are translations of one another, replace corpora with cluster IDs, train “monolingual” embeddings over all these corpora
- ▶ multiCCA: “project” all other languages into English
  - ▶ CCA: learn a projection of aligned data points into a shared space



# Cross-Lingual Embeddings

Task	multiCluster	multiCCA
dependency parsing	48.4 [72.1]	<b>48.8</b> [69.3]
doc. classification	90.3 [52.3]	<b>91.6</b> [52.6]
mono. wordsim	14.9 [71.0]	<b>43.0</b> [71.0]
cross. wordsim	12.8 [78.2]	<b>66.8</b> [78.2]
word translation	30.0 [38.9]	<b>83.6</b> [31.8]

- ▶ Word vectors work pretty well at “intrinsic” tasks, some improvement on things like document classification and dependency parsing as well





# Where are we now?

---

- ▶ Universal dependencies: treebanks (+ tags) for 70+ languages
- ▶ Many languages are still small, so projection techniques may still help
- ▶ More corpora in other languages, less and less reliance on structured tools like parsers, and pretraining on unlabeled data means that performance on other languages is better than ever
- ▶ BERT has pretrained multilingual models that seem to work pretty well (trained on a whole bunch of languages)





# Takeaways

---

- ▶ Many languages have richer morphology than English and pose distinct challenges
- ▶ Problems: how to analyze rich morphology, how to generate with it
- ▶ Can leverage resources for English using bitexts
- ▶ Next time: wrapup + ethics of NLP