Syntax-Based
Statistical Machine Translation

(Or: "Can a Machine Translate Without Knowing What a Verb Is?")

Kevin Knight

USC/Information Sciences Institute
USC/Computer Science Department

Topics

- **Quick review of statistical MT essentials**
  - bilingual text
  - phrase substitution models
  - language models
  - decoding

- **Syntax-based statistical MT**
  - syntax-based translation models
  - learning syntactic transformation rules from data
 - decoding
  - tree automata

Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。
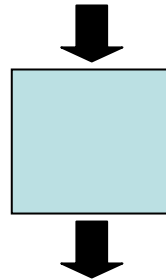
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Statistical Machine Translation
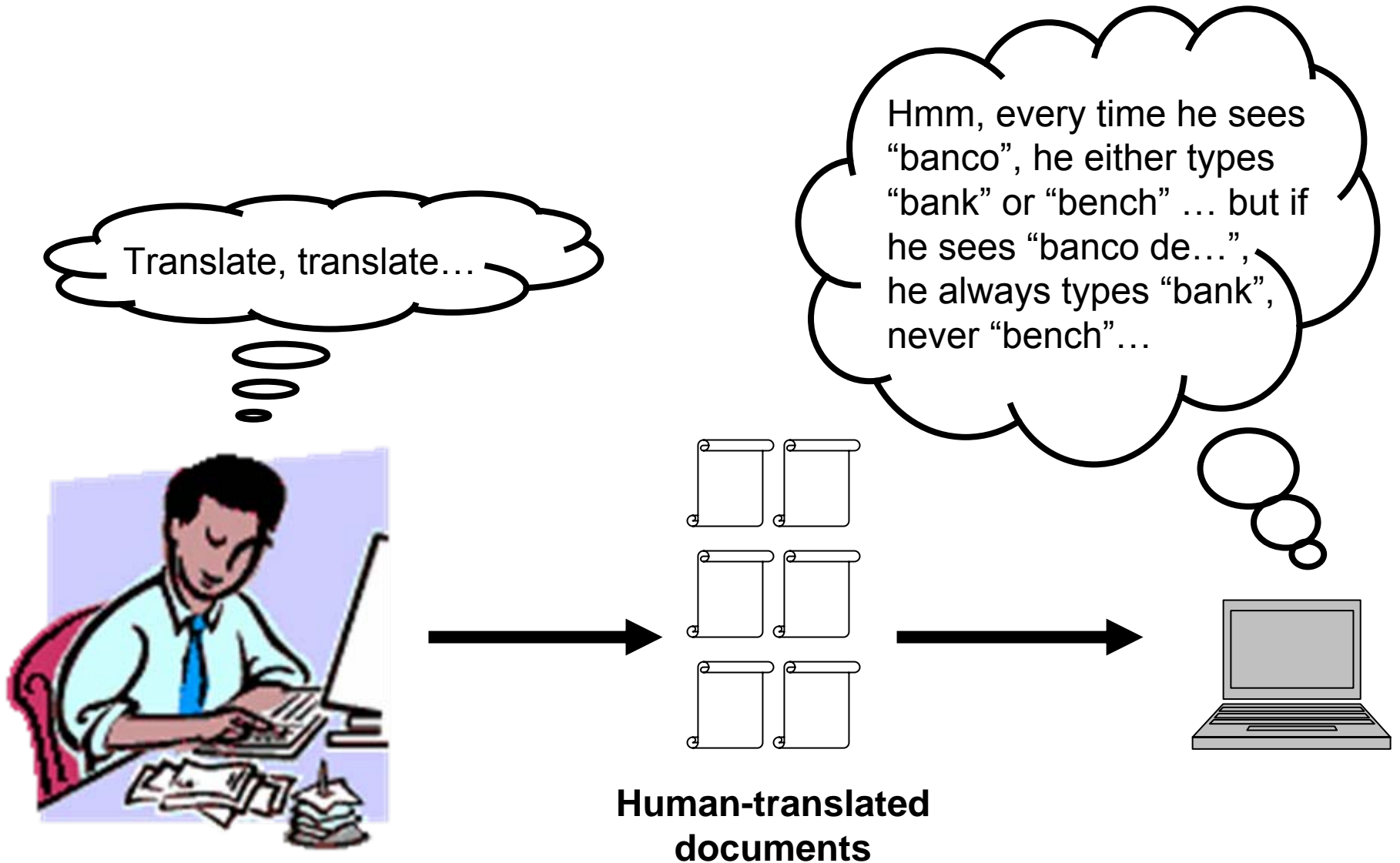
Spanish/English corpus

Translate: Clients do not sell pharmaceuticals in Europe.

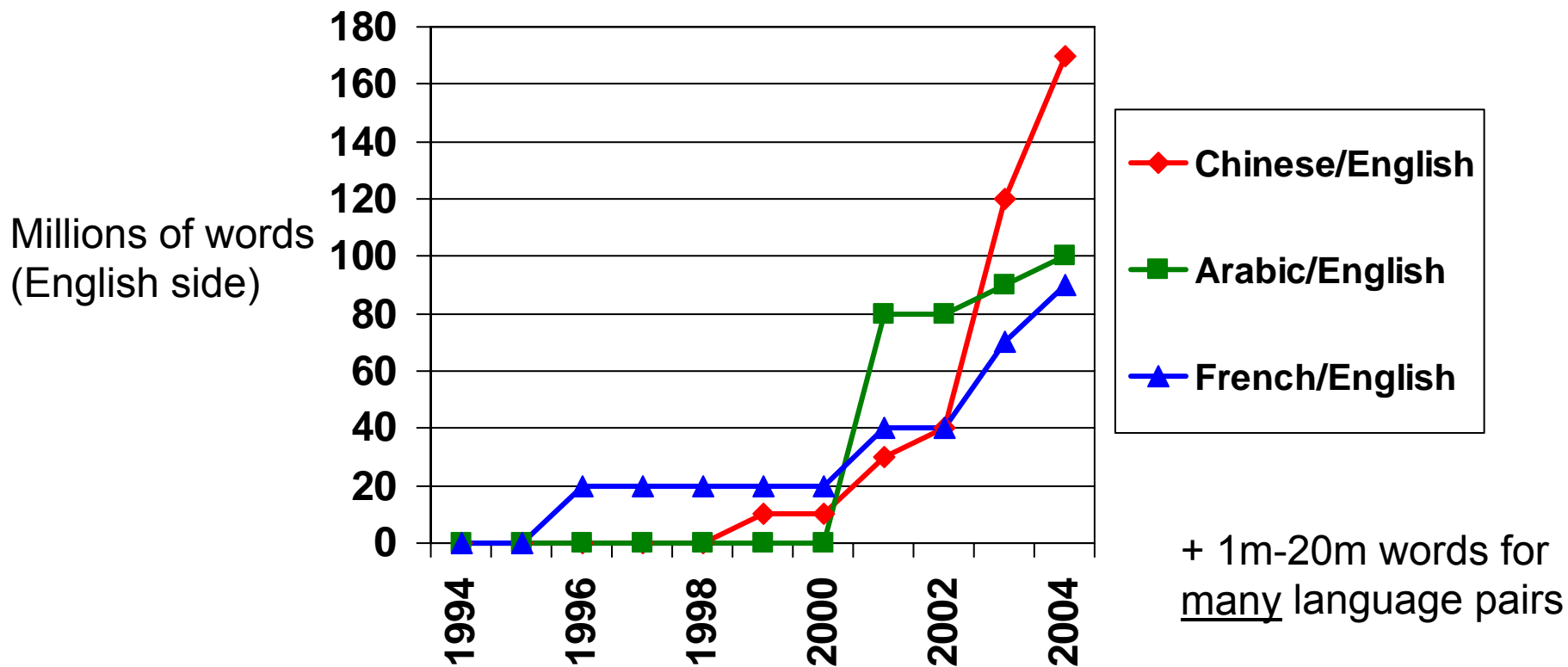|	
1a. Garcia and associates . 1b. Garcia y asociados .	7a. the clients and the associates are enemies . 7b. los clients y los asociados son enemigos .
2a. Carlos Garcia has three associates . 2b. Carlos Garcia tiene tres asociados .	8a. the company has three groups . 8b. la empresa tiene tres grupos .
3a. his associates are not strong . 3b. sus asociados no son fuertes .	9a. its groups are in Europe . 9b. sus grupos estan en Europa .
4a. Garcia has a company also . 4b. Garcia tambien tiene una empresa .	10a. the modern groups sell strong pharmaceuticals . 10b. los grupos modernos venden medicinas fuertes .
5a. its clients are angry . 5b. sus clientes estan enfadados .	11a. the groups do not sell zenzanine . 11b. los grupos no venden zanzanina .
6a. the associates are also angry . 6b. los asociados tambien estan enfadados .	12a. the small groups are not modern . 12b. los grupos pequenos no son modernos .

Centauri/Arcturan [Knight 97]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .	7a. lalok farok ororok lalok sprok izok enemok .
1b. at-voon bichat dat .	7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok .	8a. lalok brok anok plok nok .
2b. at-drubel at-voon pippat rrat dat .	8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok .	9a. wiwok nok izok kantok ok-yurp .
3b. totat dat arrat vat hilat .	9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok .	10a. lalok mok nok yorok ghirok clok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok hihok yorok zanzanok .
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok .	12a. lalok rarok nok izok hihok mok .
6b. wat dat krat quat cat .	12b. wat nnat forat arrat vat gat .

Ready-to-Use Online Bilingual Data

Millions of words
(English side)

Chinese/English

Arabic/English

French/English

+ 1m-20m words for
<u>many</u> language pairs

(Data stripped of formatting, in sentence-pair format, available
from the Linguistic Data Consortium at UPenn).

Bilingual Text (200m words)

English
strings

Word
alignments

Chinese
strings

Word-Aligned bilingual text

...

...
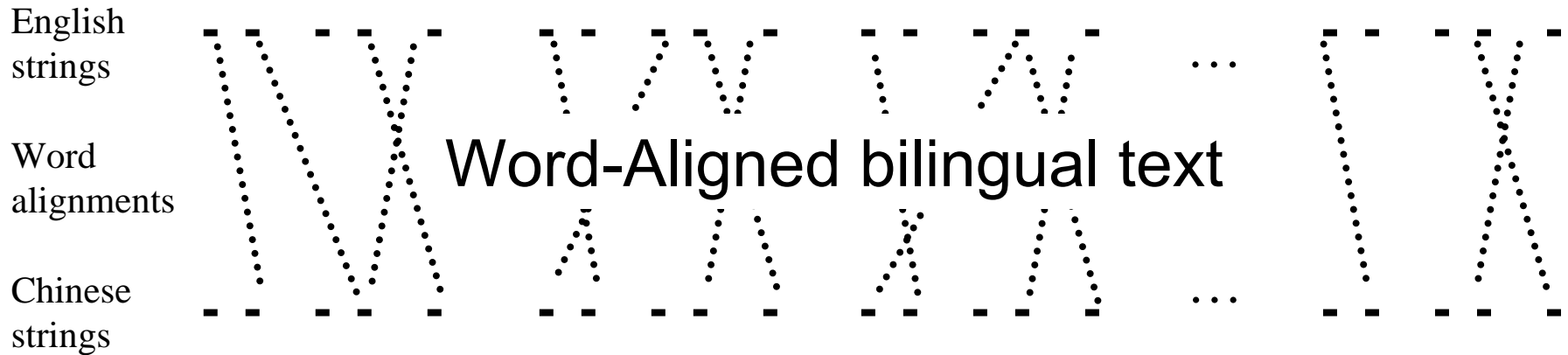
Phrase Pair Extraction [Och & Ney, 2004]

Vast Database of Phrase Pairs

California
ISI NLP
PROTECT OUR COAST AND OCEAN

Phrase-Based Translation

这　7人　中包括　来自　法国　和　俄罗斯　的　宇航　员　.

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		and the russian	the fifth			.
these	7 among	including from		the french and		of the russian	of	space	members	.
that	7 persons	including from the		of france	and to	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 included	come from	france	and russia	in	astronautical		personnel		;
	7 philtrum	including those from	france and	russia	a space		member			
		including representatives from	france and the	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				

Table 1: #11# the seven - member crew includes astronauts from france and russia .

Scoring: Try to use phrase pairs that have been frequently observed.
Try to output a sentence with frequent English word sequences.

Phrase-Based Translation

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	and the russian		the fifth		.	
these	7 among	including from		the french and		of the russian	of	space	members	.
that	7 persons	including from the		of france	and to	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 included	come from	france	and russia		in	astronautical	personnel	;
	7 philtrum	including those from	france and	russia		a space		member	
		including representatives from	france and the	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			

Table 1: #11# the seven - member crew includes astronauts from france and russia .

Scoring: Try to use phrase pairs that have been frequently observed.
          Try to output a sentence with frequent English word sequences.

Phrase-Based Translation

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	and the russian		the fifth		.	
these	7 among	including from		the french and		of the russian	of	space	members	.
that	7 persons	including from the		of france	and to	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 included		come from	france	and russia		in	astronautical	personnel	;
	7 philtrum	including those from		france and	russia		a space		member	
		including representatives from		france and the	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				

Table 1: #11# the seven - member crew includes astronauts from france and russia .

Scoring: Try to use phrase pairs that have been frequently observed.
        Try to output a sentence with frequent English word sequences.

Phrase-Based Translation

| 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	and the russian		the fifth			.
these	7 among	including from		the french and	of the russian		of	space		members	.
that	7 persons	including from the		of france	and to	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 included	come from		france	and russia		in	astronautical		personnel	;
	7 philtrum	including those from	france and		russia	a space		member	
		including representatives from	france and the		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		

Table 1: #11# the seven - member crew includes astronauts from france and
russia .

Scoring: Try to use phrase pairs that have been frequently observed.
          Try to output a sentence with frequent English word sequences.

Components

- Training algorithms
  - Word alignment, phrase pair extraction…
    - P(chinese | english) = product of conditional phrase pair probabilities
  - English n-gram models…
    - P(english) = product of trigram probabilities
    - P(w3 | w1 w2)
- Decoding algorithm
  - argmax e P(english | chinese)   =
    argmax e P(english) * P(chinese | english)

Features and Tuning

- English n-gram language model

- Phrase pairs

    - Corpus probability of phrase pair

    - Bad-phrase spotter

    - Word-drop spotter

    - "Move Me" preference

- English output length
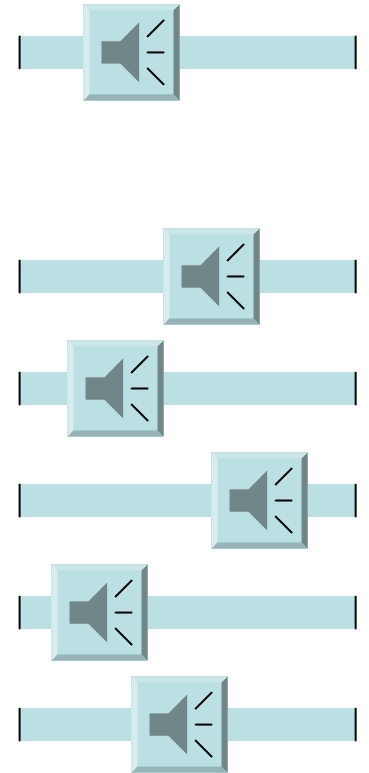
We compute a total score for each possible translation -- a linear weighted combination of these six values. This generalizes the formula from the previous slide, if we switch to log probs.

Weight settings really affect translation quality!

Translation accuracy

W_{length}

W_{NGLM}
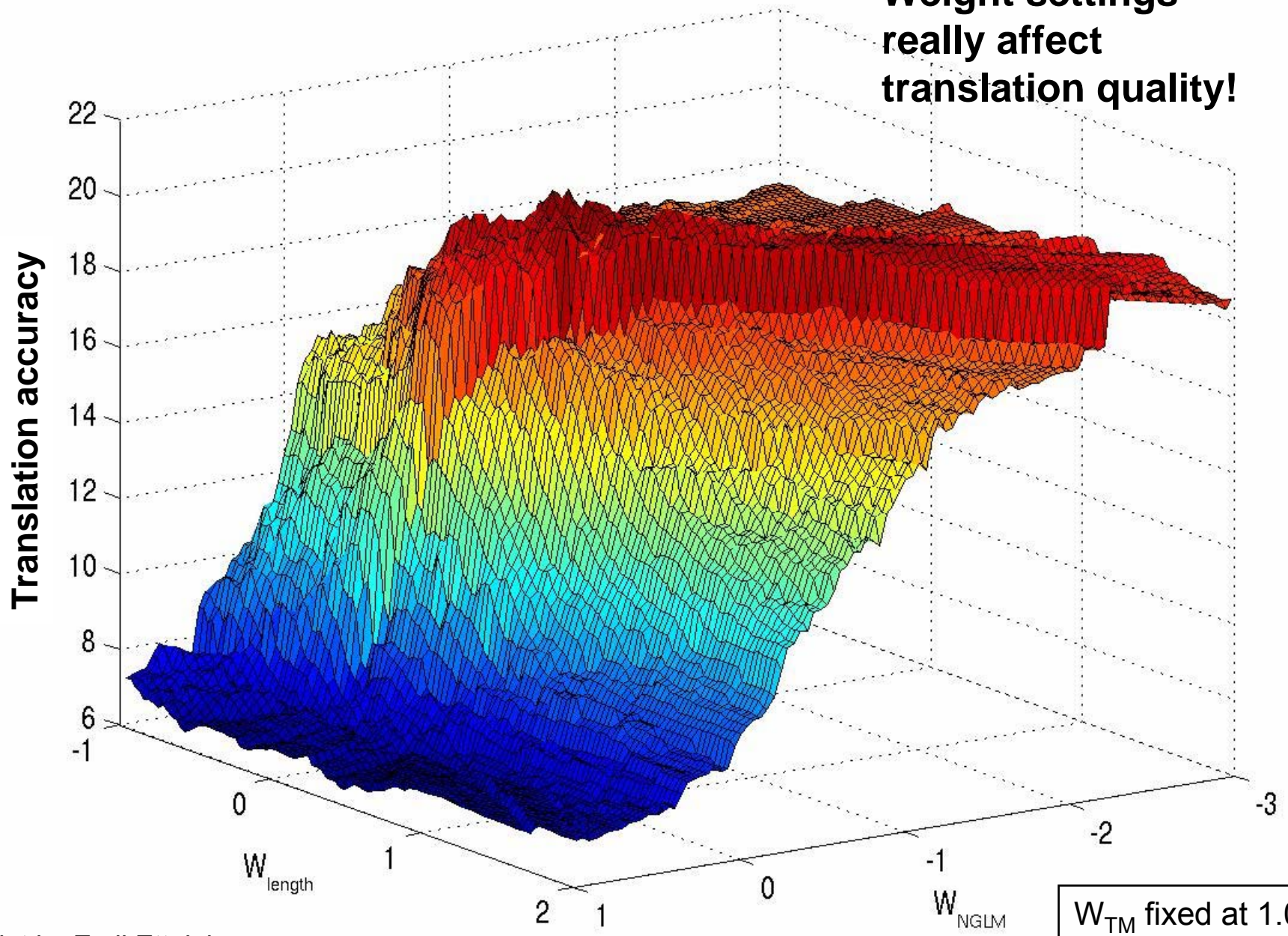
W_{TM} fixed at 1.0

plot by Emil Ettelaie

(A View from the Back)
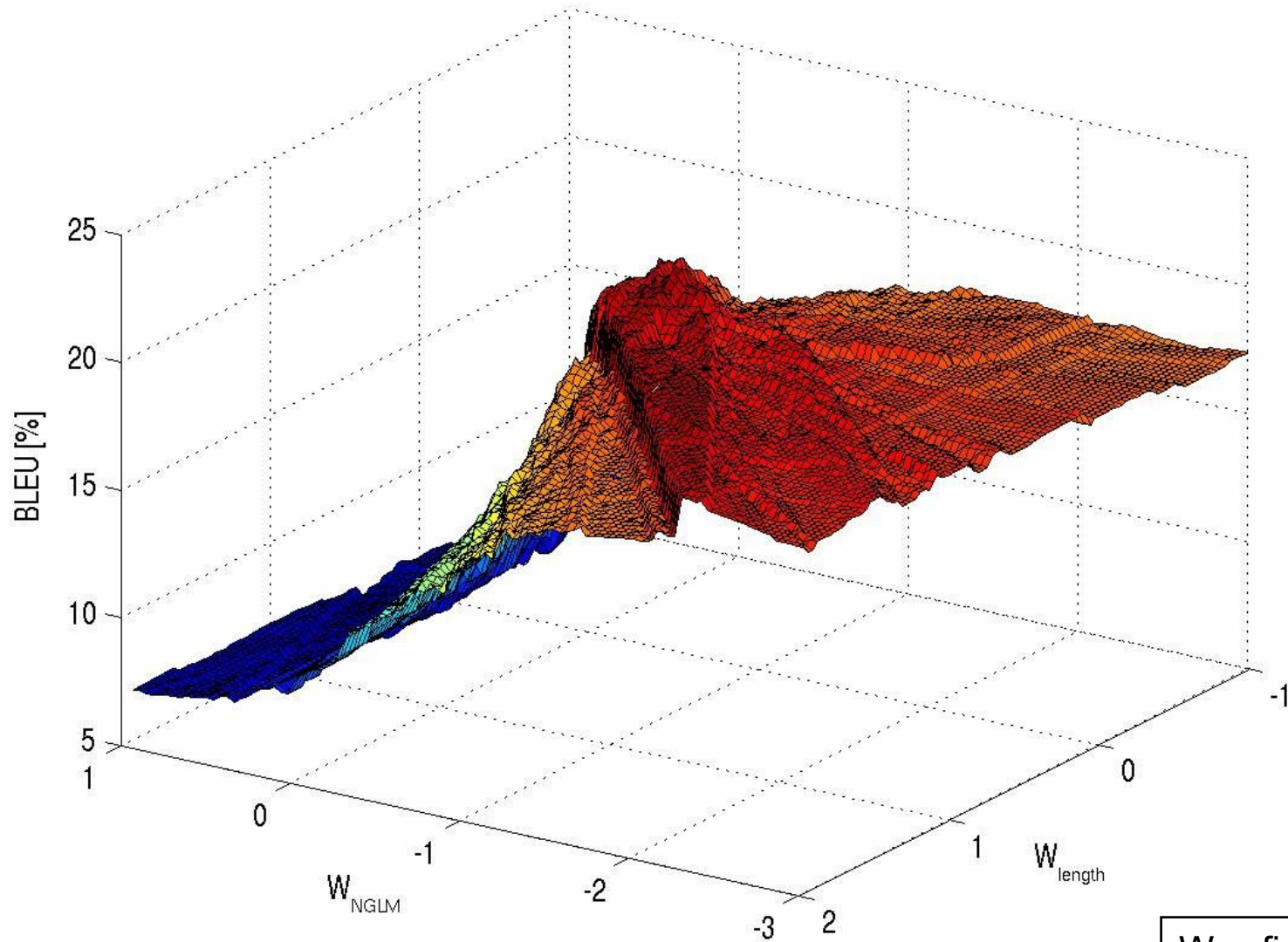
W_{TM} fixed at 1.0

Hill climbing with Minimum Error-Rate Training (MERT) aka MaxBleu Training (Och, 2003)

W_{TM} fixed at 1.0
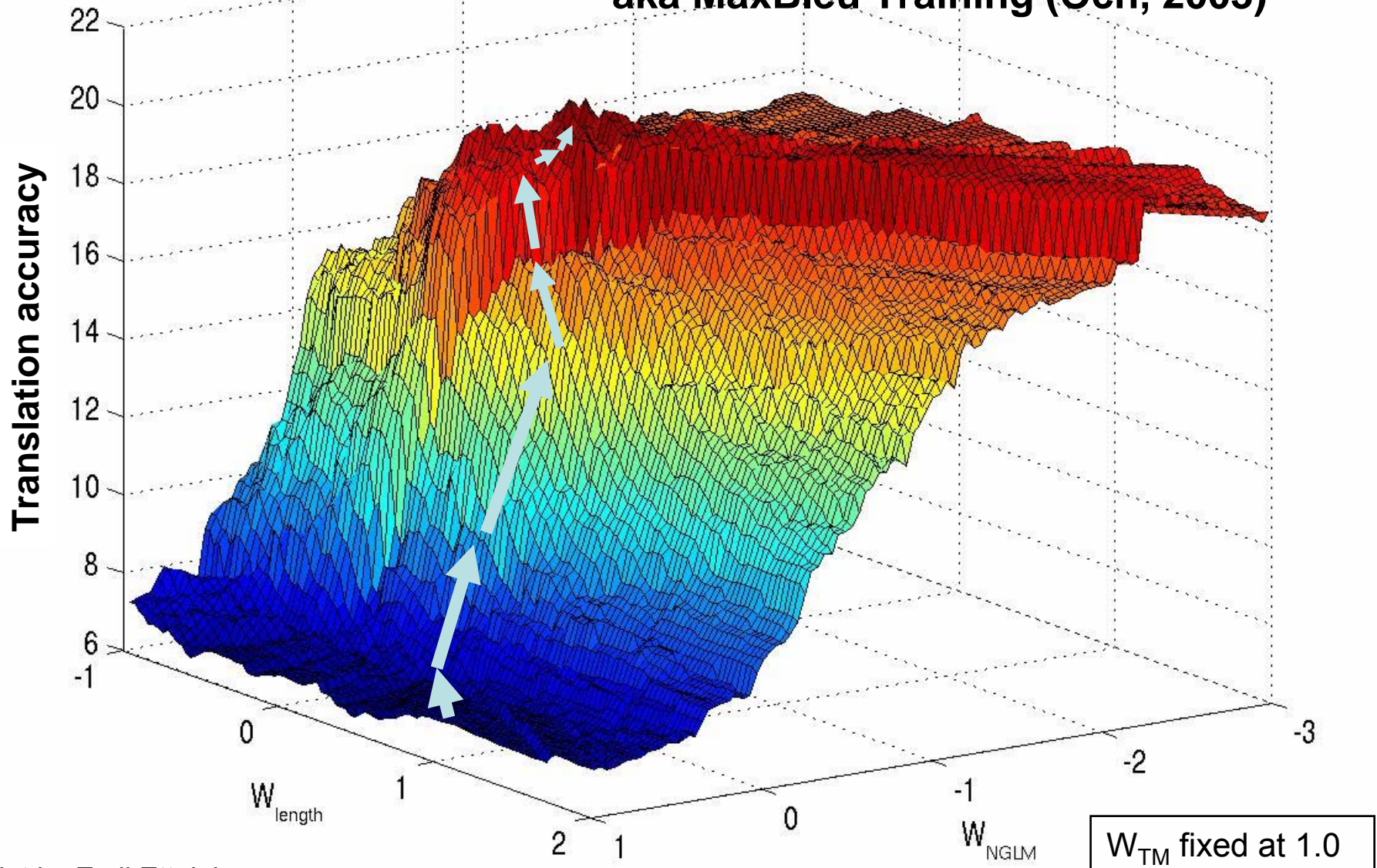
plot by Emil Ettelaie

These Ideas Work!

Translation Quality
(BLEU)

Phrase-based MT Progress

70
60
50
40
30
20
10

2002   2003   2004   2005

NIST Common Evaluations
(Arabic/English)
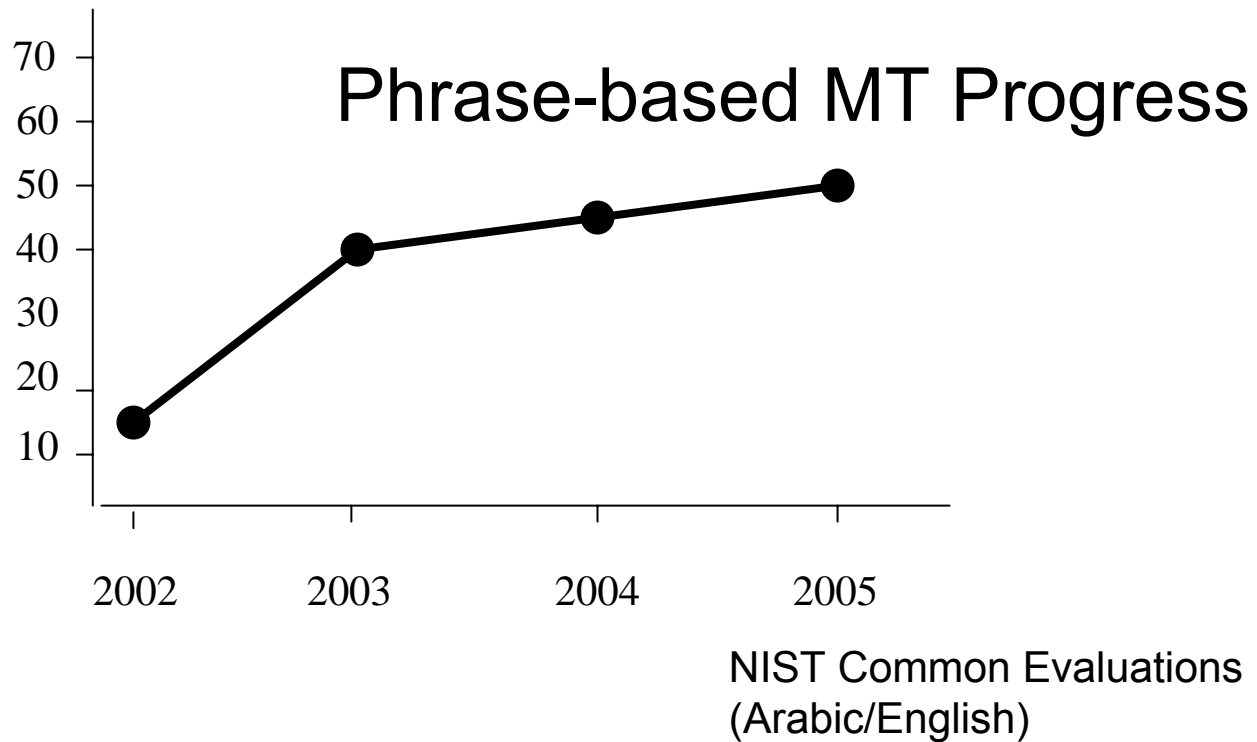
Some Lessons

- The simpler, the better
- It takes a long time just to get the bugs out!
- Every change has to be carefully checked
- Good ideas often don't help
- Have to try lots of things
- It's highly experimental

Statistical MT Research is Highly Experimental

Translation Accuracy (BLEU)

Actual progress at undisclosed laboratory!

Chinese/English NIST 2002 Test Set

35

30

25

20

15

Mar 1

Apr 1

May 1

2005
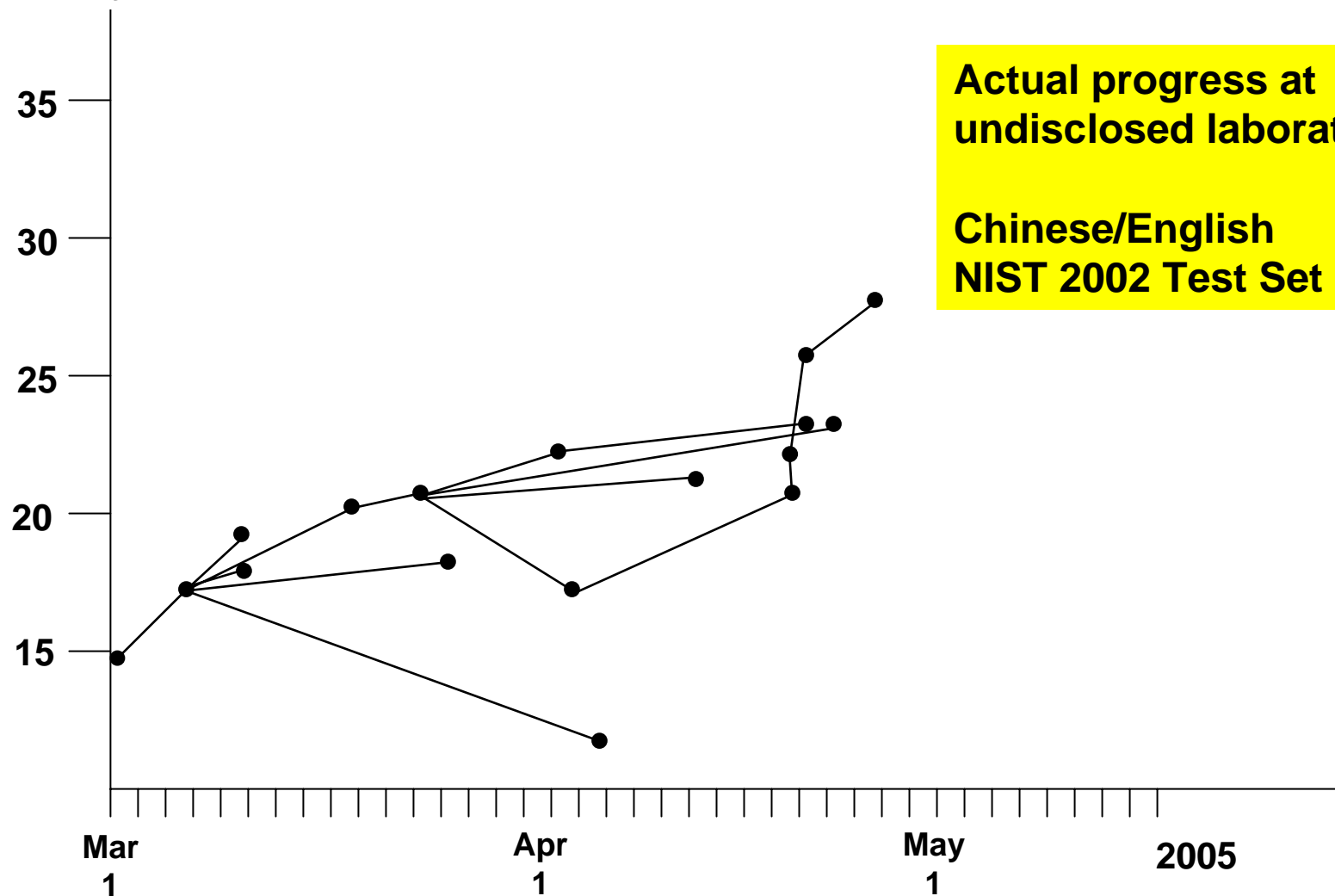
Two Ways to Improve Statistical MT Systems

Quality of resulting translation system

better algorithms

more data

Amount of bilingual training data
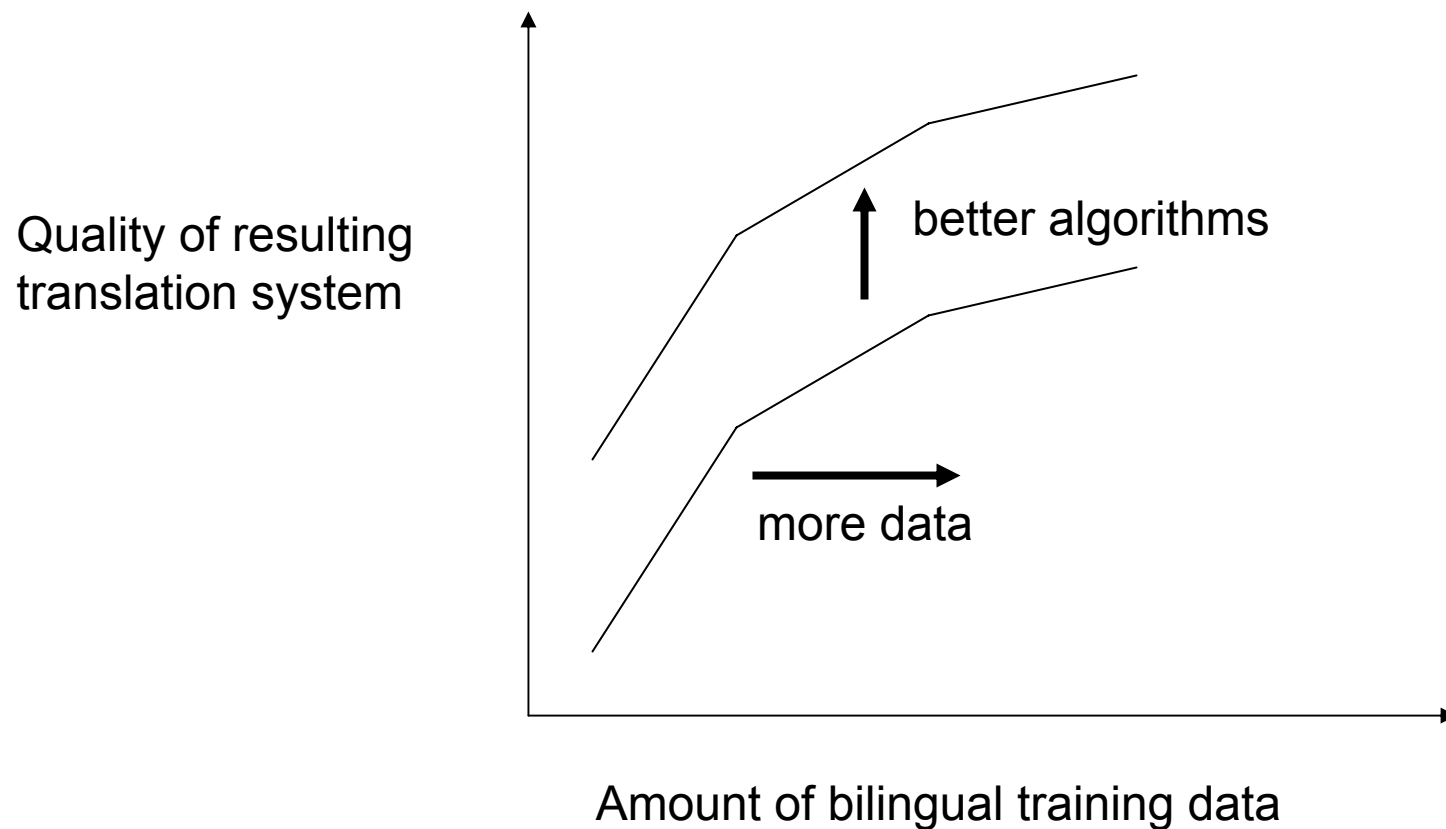
Can a machine translate between Chinese and English without knowing what a verb is?

- Of course
- But the output is often bad

    "Frequent high-tech exports are bright spots for foreign trade growth of Guangdong has made important contributions."

- Our phrase-based story might need some work

Syntax

Maybe we need some grammar?

MT Research Landscape

Syntax will never work!
We're better off without syntax!
Syntax has been *shown* to make things worse!
It has never worked in speech recognition!
You are crazy!

Syntax will never work!
You need *semantics*!
Language is about the world!
You are crazy!
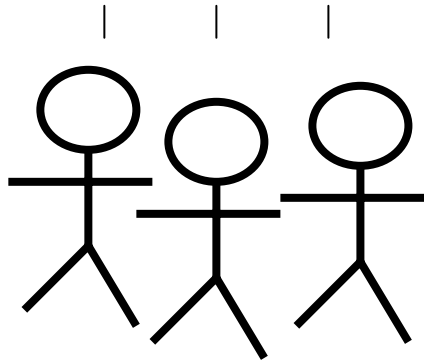
ACL Language Engineers
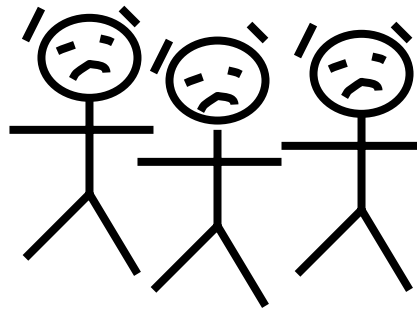
AAAI Fellows
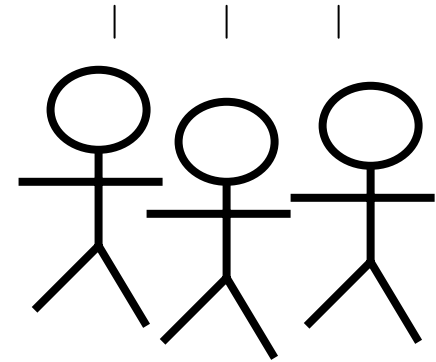
Working on syntax-based approach
to translation (nouns, verbs,
prepositional phrases…)

MT Progress

Translation Quality (BLEU)

Phrase-based MT Progress

70
60
50
40
30
20
10

syntax didn't work

syntax didn't work

syntax didn't work

syntax didn't work

2002   2003   2004   2005

NIST Common Evaluations
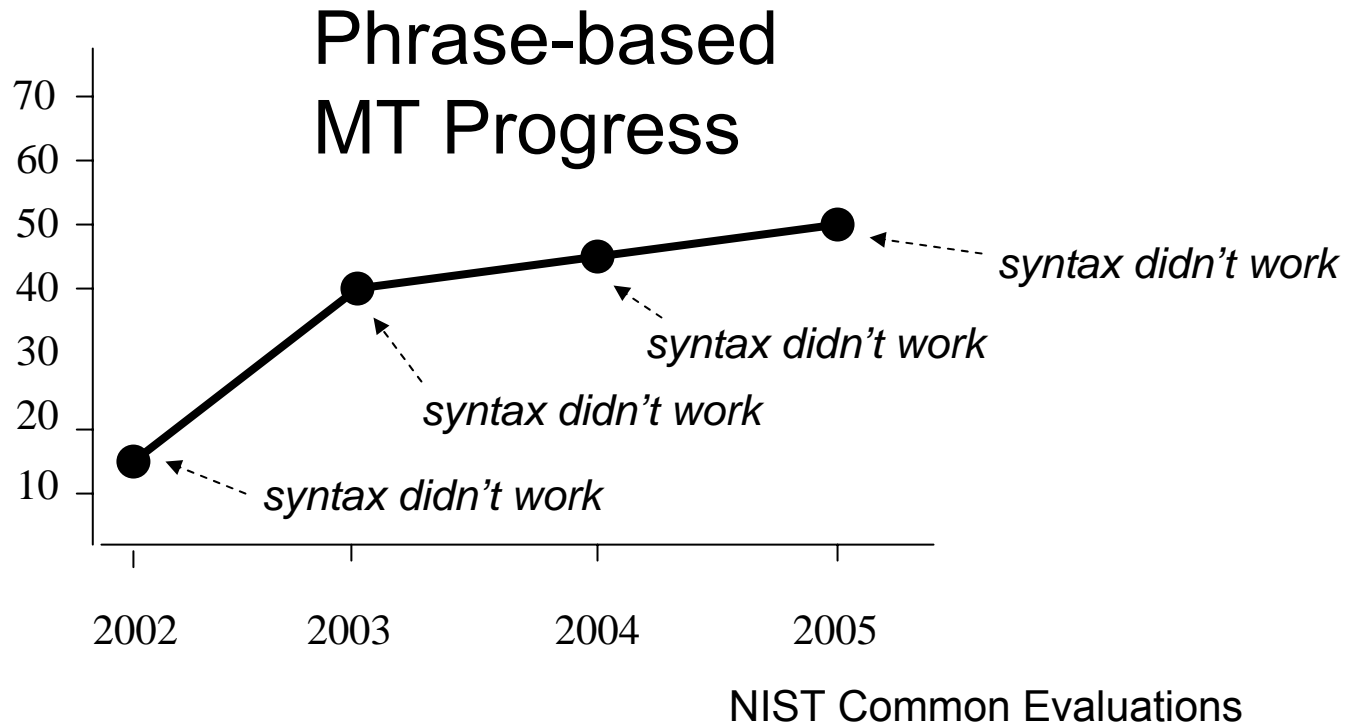
Syntax Started to Be Helpful in 2006

Translation Accuracy

Chinese/English

Phrase-based ══════

sentences < 16 words
(NIST-03/04)

Phrase-based ══════

all sentences
(NIST-2003)
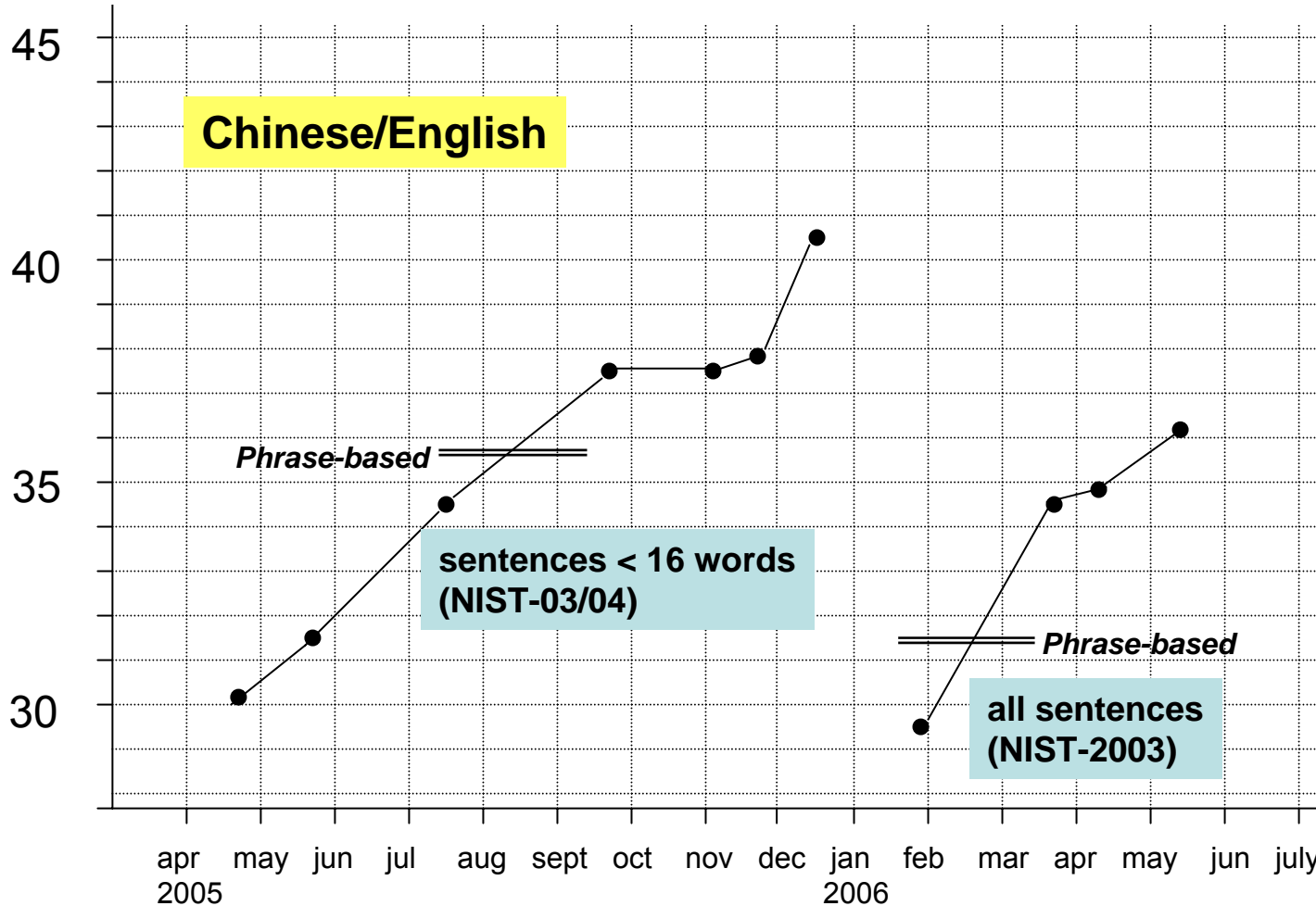
How to Add Syntax?

- Automatically parse training data
  - Many parsers are available: (Collins 97, Charniak 01, etc)
- Then many approaches are possible
  - Add **syntactic features** to phrase-based system
    - many references
  - Syntactically **re-order source sentences** into target-like word order (for training and decoding)
    - (Berger et al 94, Xia & McCord 04, Collins et al 05, etc)
  - Build **tree-to-tree** translation systems
    - (Eisner 03, Gildea 03, Melamed 04, Riezler & Maxwell 06, Cowan et al 06, etc)
  - Build **tree-to-string** translation systems
    - (Quirk et al 05, Huang et al AMTA-06, Liu et al 06, etc)
  - Build **string-to-tree** translation systems
    - (Yamada & Knight 01, Galley et al 04, Venugopal & Zollmann 06, etc)
- Let's just look at one approach & investigate

Phrase-Based Output

枪手　被　警方　击毙 ．

Gunman of police killed ．
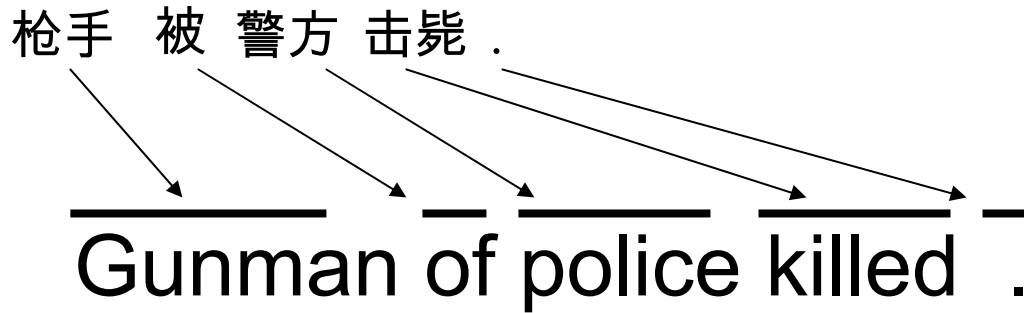
*Decoder
Hypothesis #1*

Phrase-Based Output

枪手 被 警方 击毙 .

Gunman of police attack .

*Decoder
Hypothesis #7*

Phrase-Based Output

枪手　被　警方　击毙 .

Gunman by police killed .
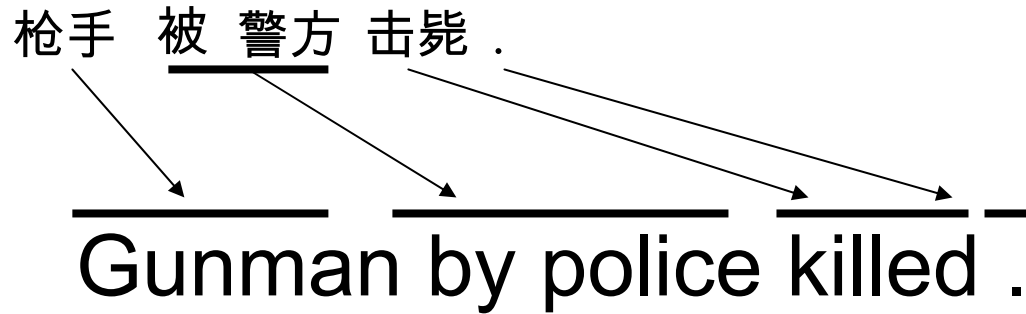
Decoder
Hypothesis #12

Phrase-Based Output

枪手　被　警方　击毙 .

Killed gunman by police .

Decoder
Hypothesis #134

Phrase-Based Output

枪手 被 警方 击毙 .

_____ _____ ____ _____ __
Gunman killed the police .

*Decoder
Hypothesis #9,329*

Phrase-Based Output

枪手　被　警方　击毙 .

Gunman killed by police   .

highest scoring output, phrase-based model

Decoder Hypothesis #50,654

Problematic:
- VBD "killed" needs a direct object
- VBN "killed" needs an auxiliary verb ("was")
- countable "gunman" needs an article ("a", "the")
- "passive marker" in Chinese controls re-ordering

Can't enforce/encourage any of this!

Syntax-Based Output

枪手　被　警方　击毙 .

The gunman killed by police .

Decoder Hypothesis #1

DT　NN　VBD IN　NN

NPB　　　　　　PP

NP-C　　　　VP

S

Syntax-Based Output

枪手　被　警方　击毙 .

Gunman　by police shot .

NN　　IN　NN　VBD

NPB　　　　PP

NP-C　　　　　　VP

S

Decoder
Hypothesis #16

Syntax-Based Output

枪手 被 警方 击毙 .

The gunman was killed by police .

DT   NN   AUX VBN IN   NN

NPB

NP-C

VP

PP

S

Decoder Hypothesis #1923

highest scoring output, syntax-based model

Syntax-Based Output

- Better modeling of target language structure
  - Always a verb
  - Verb is always in the right place

- Better handling of function words
  - They often don't translate
  - But they control how the translation goes

- Better generalization in translation patterns

Syntax-Based Statistical MT

- Terminology
- Mathematical Framework
- Translation Model
- Language Model
- Decoder

word alignment

estring

These 7 people include astronauts coming from France and Russia .

这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

cstring

Mathematical Framework

- ## String-based system

$$\text{argmax}_{e,a} \quad P(e, a, c)^\alpha \cdot P(e)^\beta \cdot |e|^\gamma \cdot \ldots$$

- ## Tree-based system

$$\text{argmax}_{etree,a} \quad P(etree, a, c)^\alpha \cdot P(etree)^\beta \cdot |etree|^\gamma \cdot \ldots$$

translation model   language model   length bonus

String-to-Tree

- Mathematically, we want a weighted relation with pairs drawn from:
  - (the infinite) set of Chinese strings
  - (the infinite) set of English trees

- Good pairs should have a high weight

- Bad pairs should have a low weight

- Probabilistic generative modeling approach
  - How does a Chinese string become an English tree (or vice-versa)?

An Early Syntactic Model of Translation

[Yamada & Knight 01]

Parse (E)

[Yamada & Knight 01]
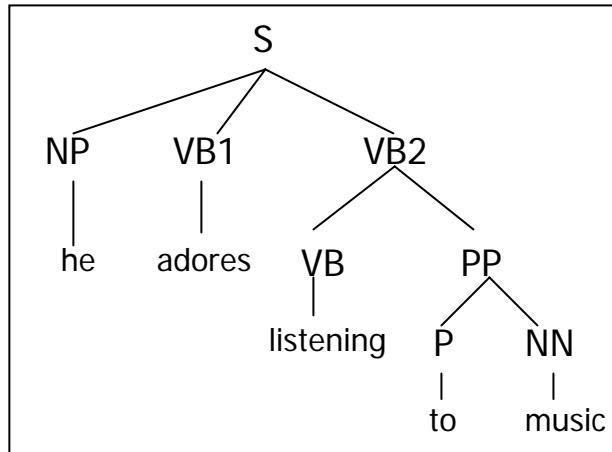
Reorder
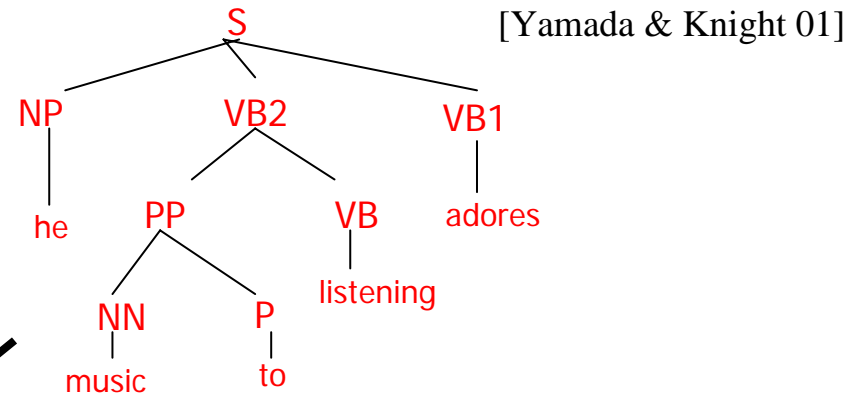
Insert
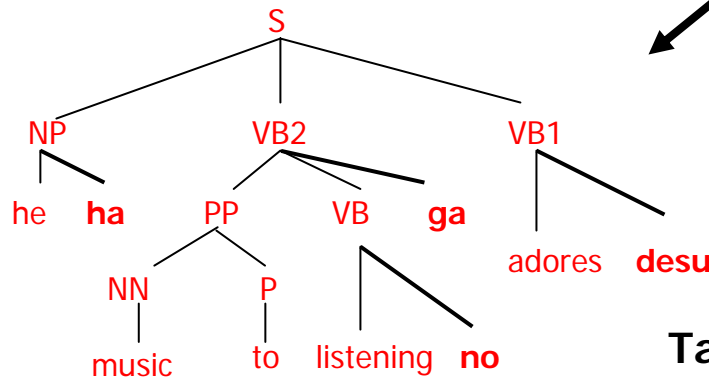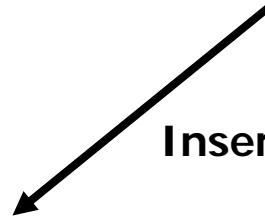
Translate
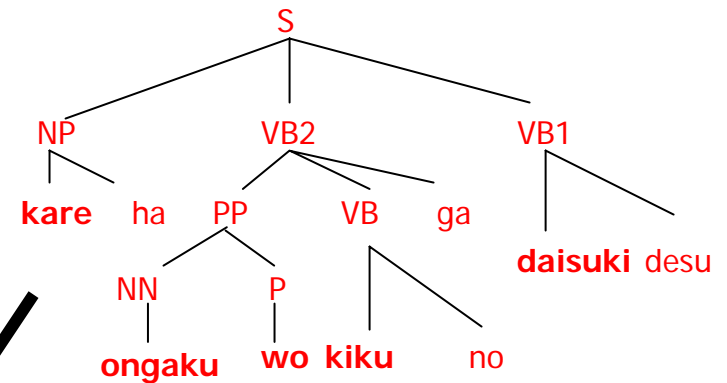
Take Leaves

Sentence (J)   **Kare ha ongaku wo kiku no ga daisuki desu**

Phrase-Based

Finite-state acceptor

Finite-state transducer

Language Model → English string → Translation Model → Chinese string

English string ← Decoder ← Chinese string
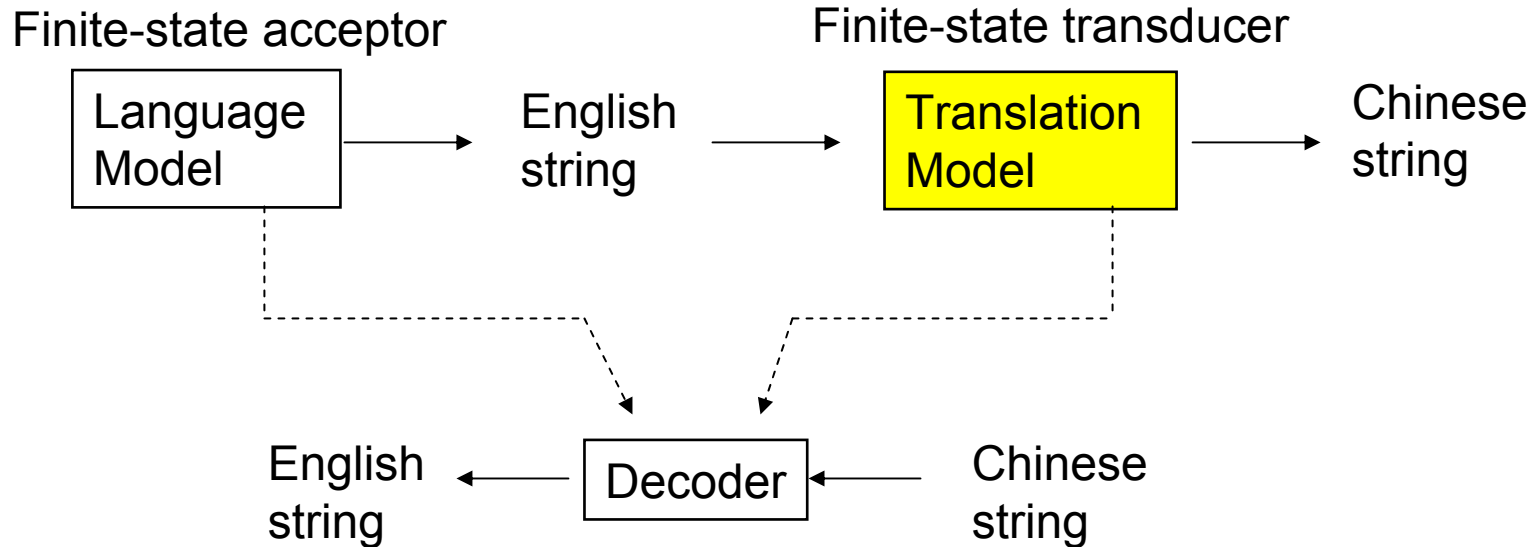
- – Grab a chunk of English string
- – Decide how to translate it (using phrase pair inventory)
- – Recurse on remaining input
  - Can be modeled by finite-state string transducer
  - [Mealy, 1959] → [Kumar & Byrne, 2003, HLT]

Syntax-Based

Tree Acceptor

Tree Transducer

```
Language        →   English   →   Translation   →   Chinese
Model               tree          Model             string
```

```
English   ←   Decoder   ←   Chinese
tree                        string
```
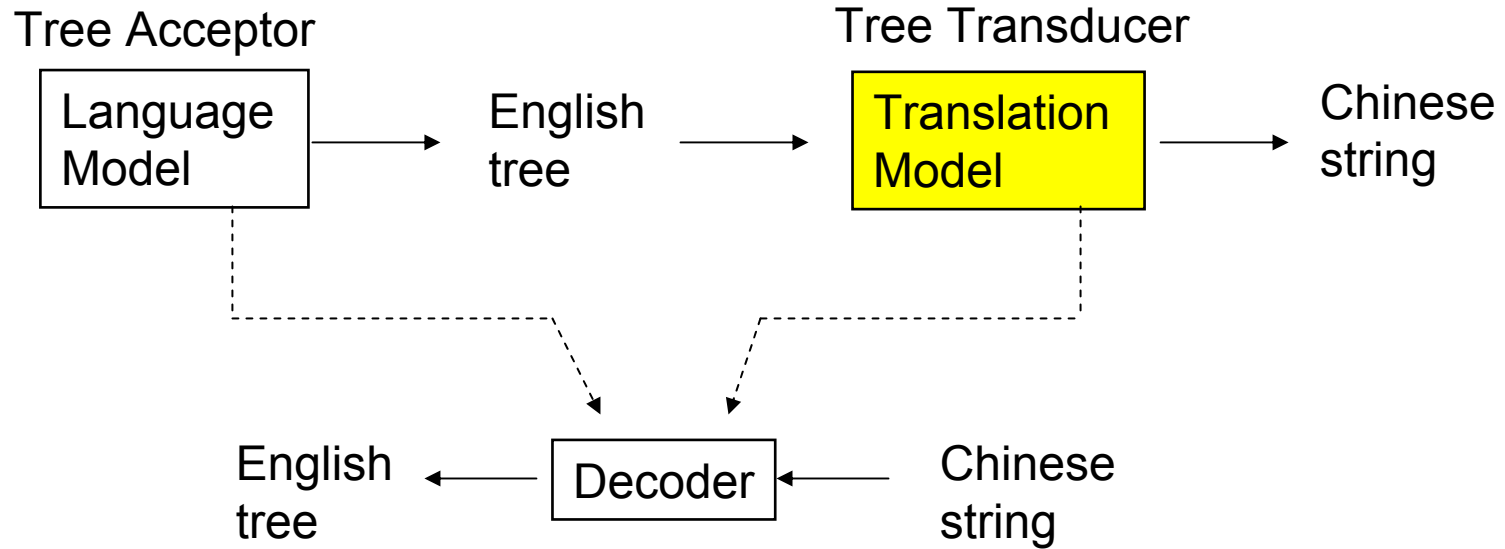
– Grab a chunk of English input tree
– Decide how to translate it
– Recurse of remaining subtrees
  • Can be modeled by tree transducer
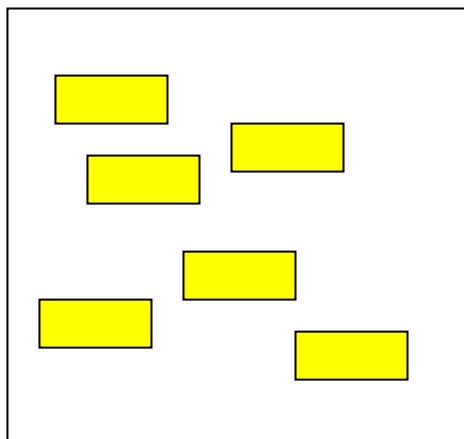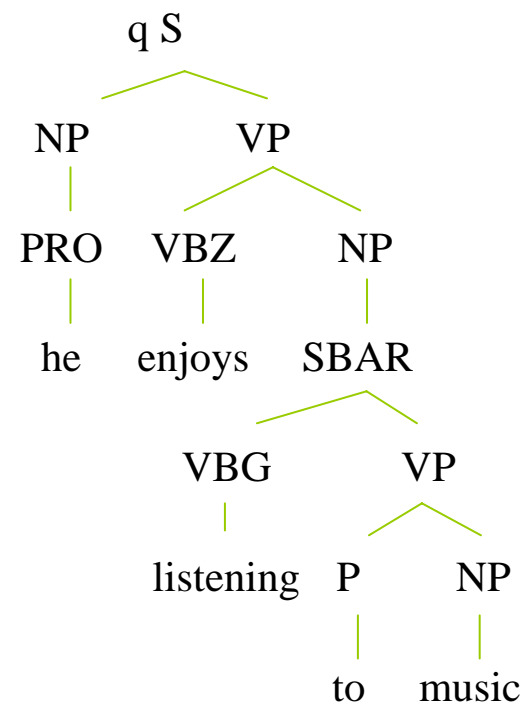 • [Rounds, 1970] → [Graehl & Knight, 2004, HLT]

Top-Down Tree Transducer
(W. Rounds 1970; J. Thatcher 1970)
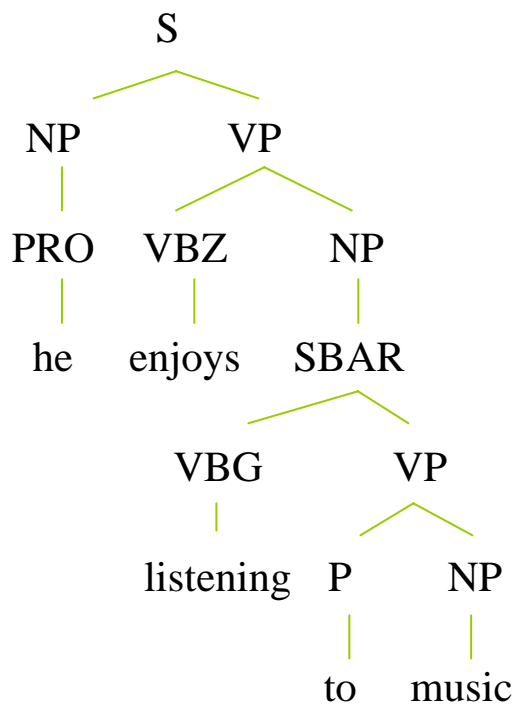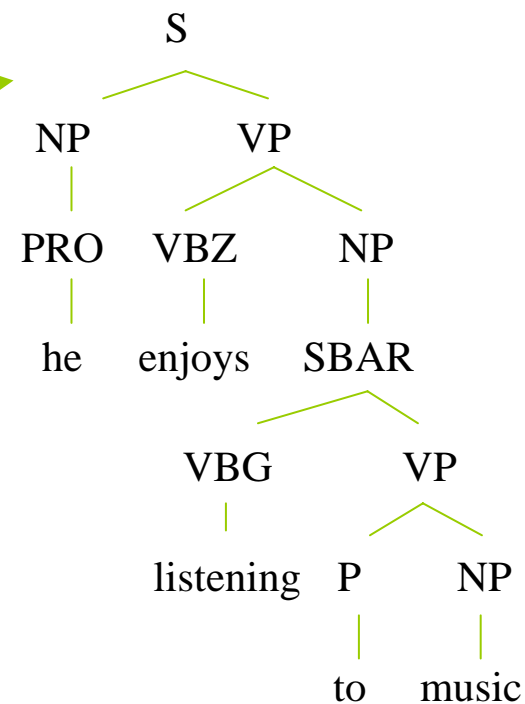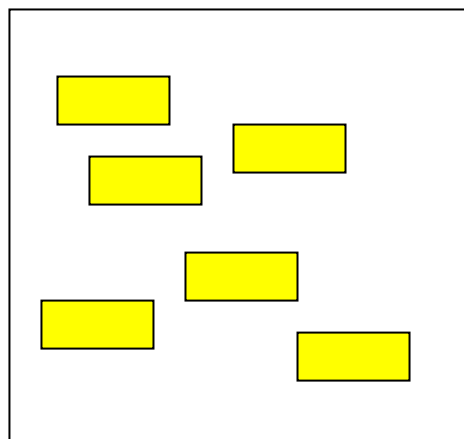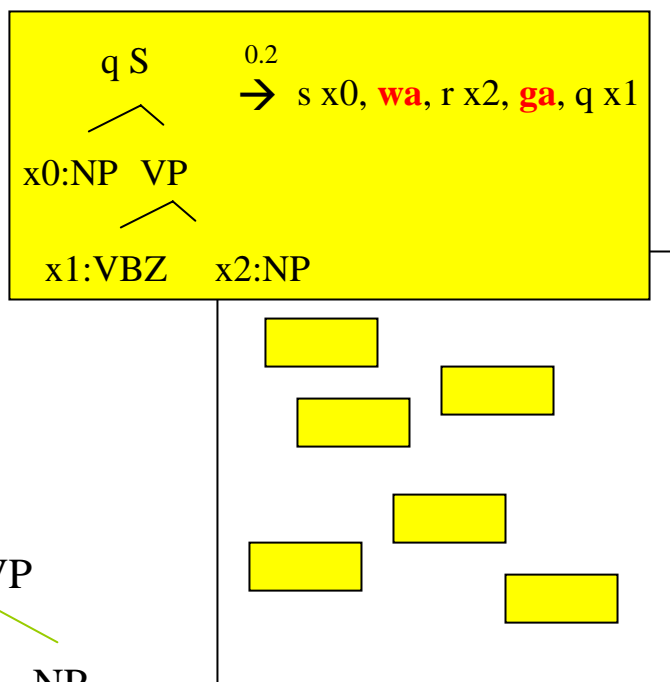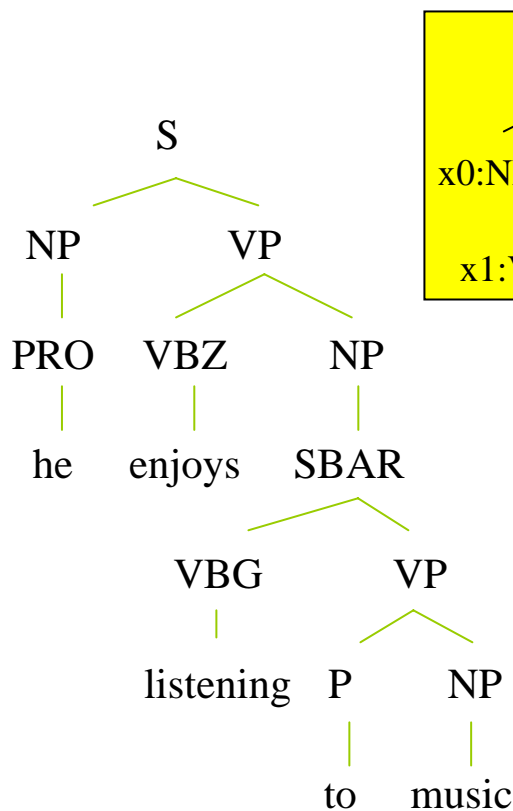
Original input:

Transformation:

Top-Down Tree Transducer
(W. Rounds 1970; J. Thatcher 1970)

Original input:

Transformation:

Top-Down Tree Transducer
(W. Rounds 1970; J. Thatcher 1970)
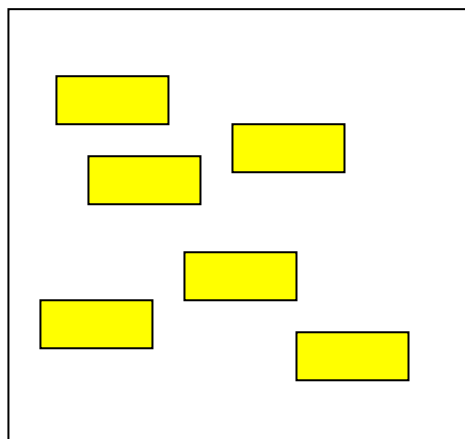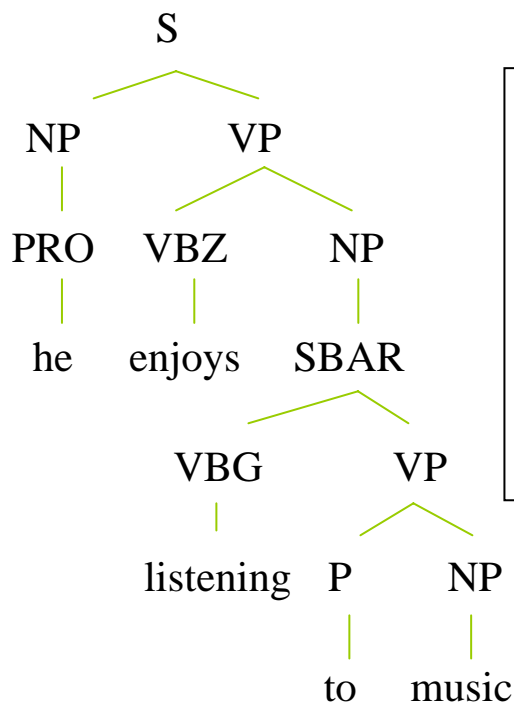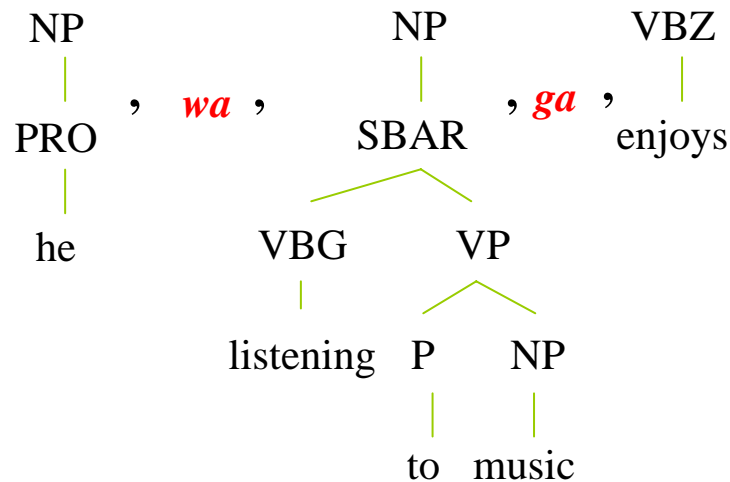
Top-Down Tree Transducer
(W. Rounds 1970; J. Thatcher 1970)
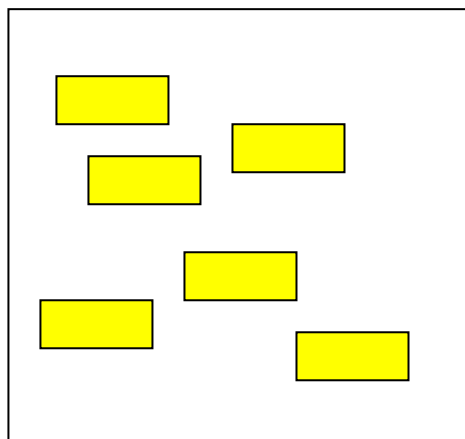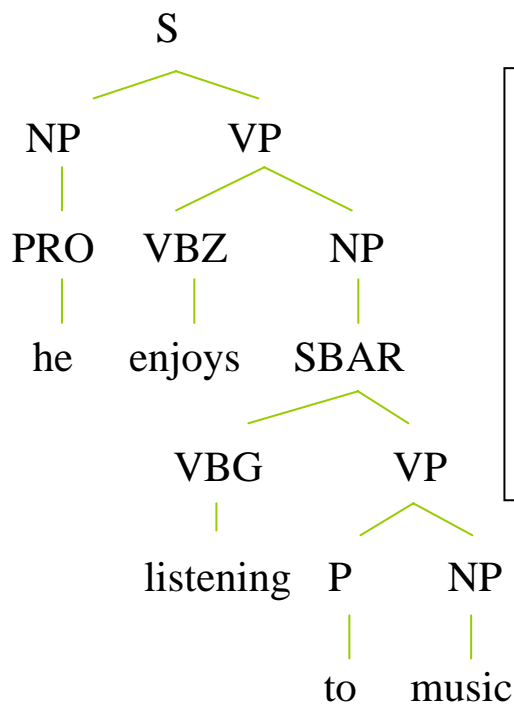
Original input: **Transformation:**
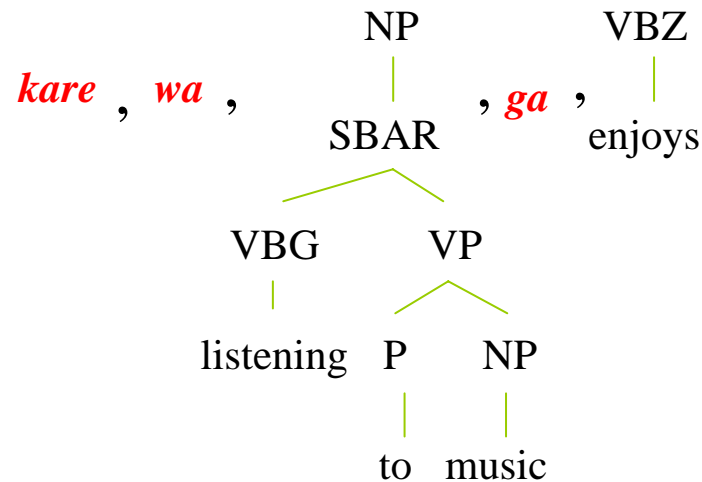
Top-Down Tree Transducer
(W. Rounds 1970; J. Thatcher 1970)
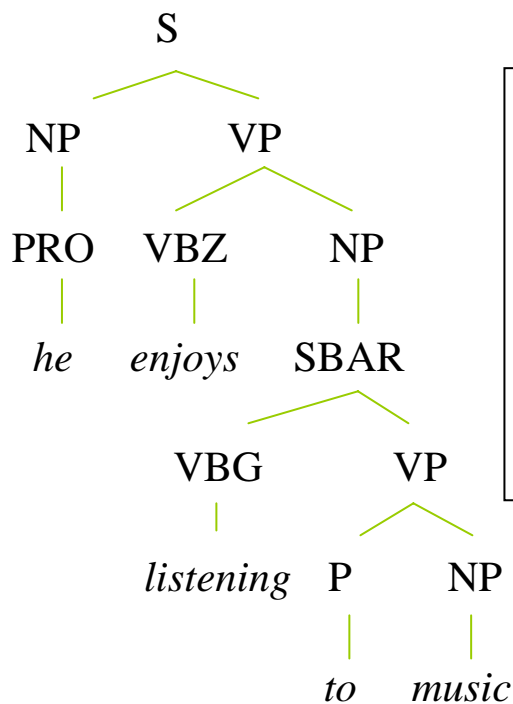
Original input:

Final output:
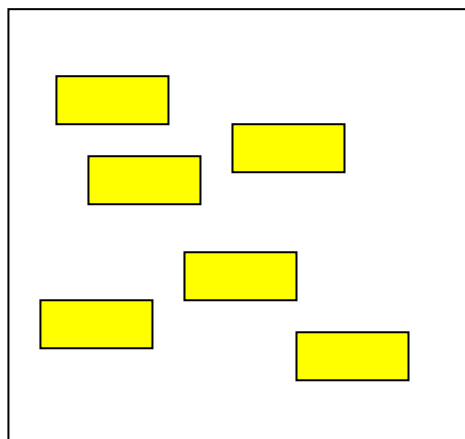
kare , *wa* , *ongaku* , *o* , *kiku* , *no* , *ga* , *daisuki* , *desu*

Top-Down Tree Transducer
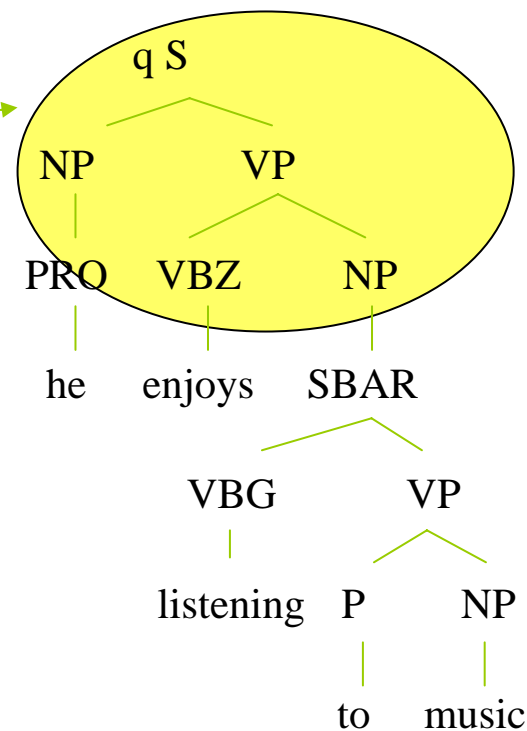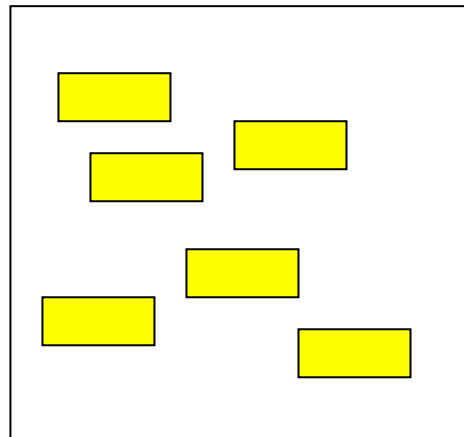## (W. Rounds 1970; J. Thatcher 1970)
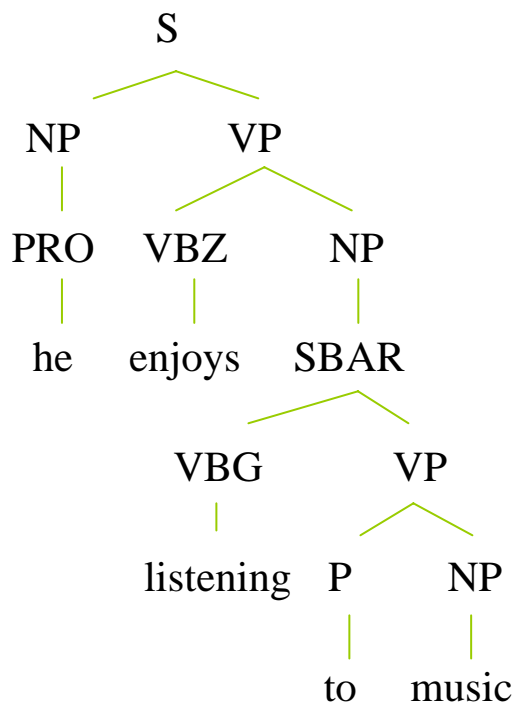
Top-Down Tree Transducer
(W. Rounds 1970; J. Thatcher 1970)

Original input:

Transformation:

q S
0.2
→ s x0, **wa**, r x2, **ga**, q x1

x0:NP VP

x1:VBZ   x2:NP

Top-Down Tree Transducer
(W. Rounds 1970; J. Thatcher 1970)

Top-Down Tree Transducer
(W. Rounds 1970; J. Thatcher 1970)
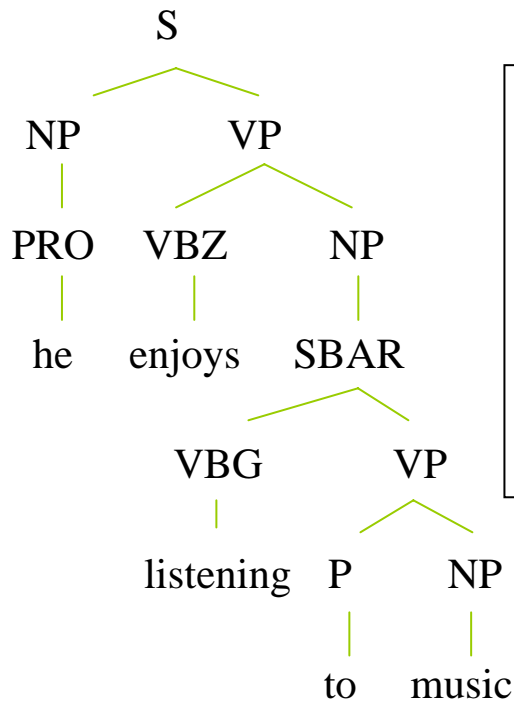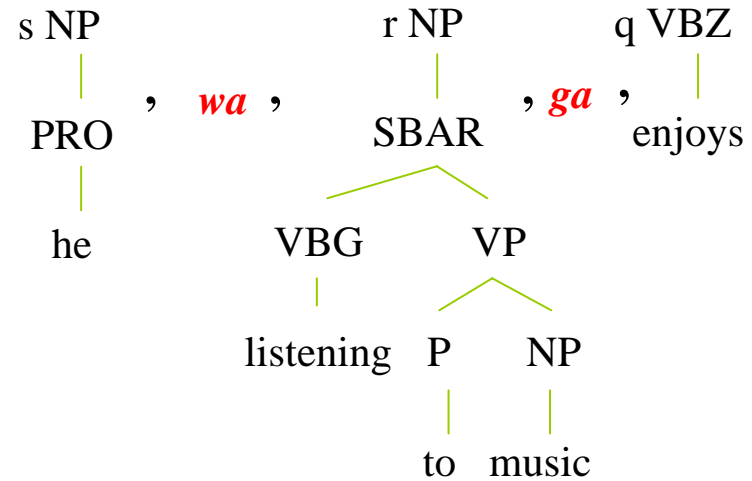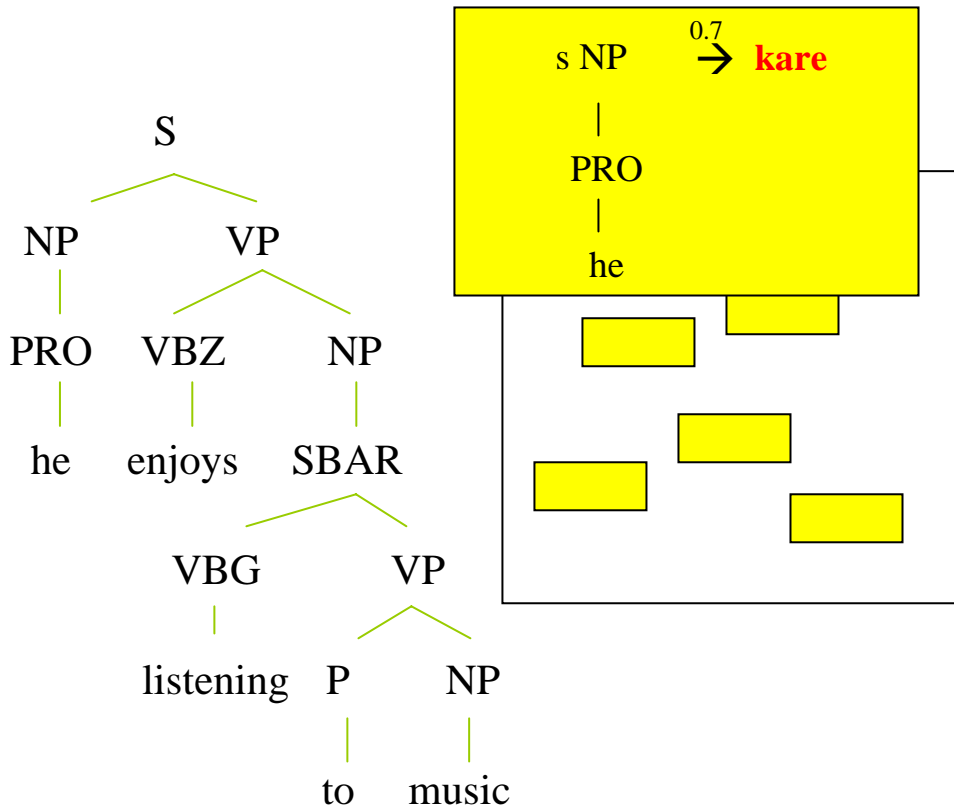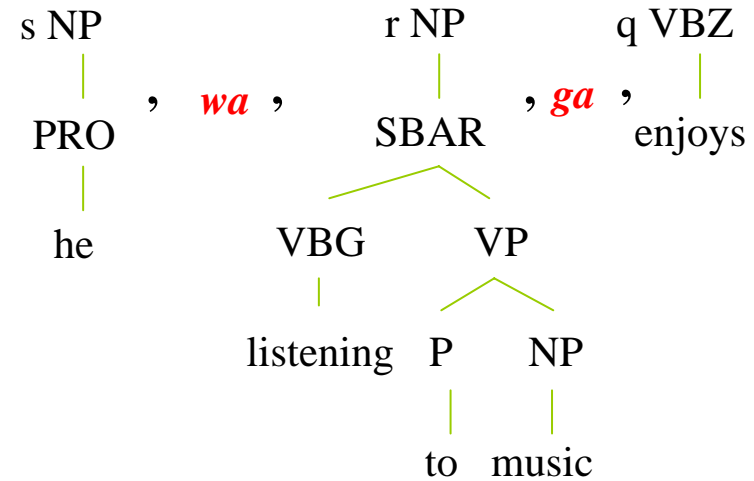
Original input:

Transformation:

Top-Down Tree Transducer
(W. Rounds 1970; J. Thatcher 1970)

Original input:

Transformation:

Top-Down Tree Transducer
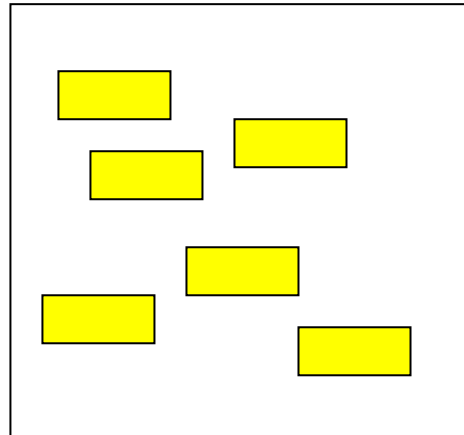## (W. Rounds 1970; J. Thatcher 1970)

Original input:
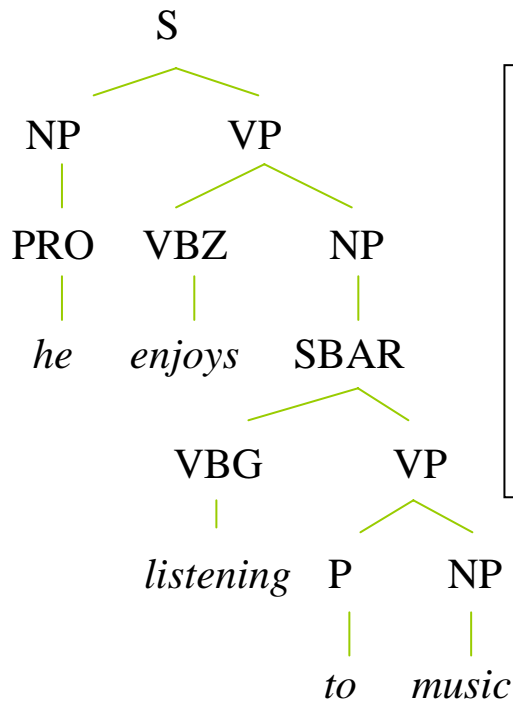
Final output:
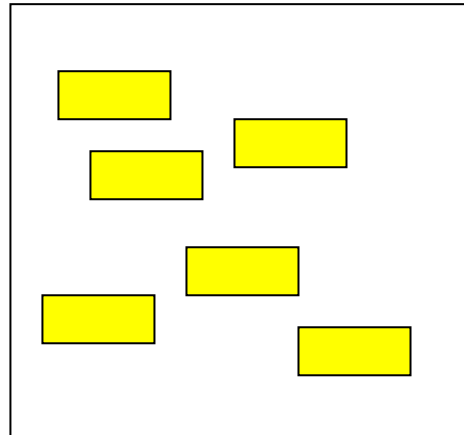
kare , *wa* , *ongaku* , *o* , *kiku* , *no* , *ga* , *daisuki* , *desu*

To get total probability, multiply probabilities of the individual steps.

An Early Syntactic Model of Translation
[Yamada & Knight 01]

Parse (E)
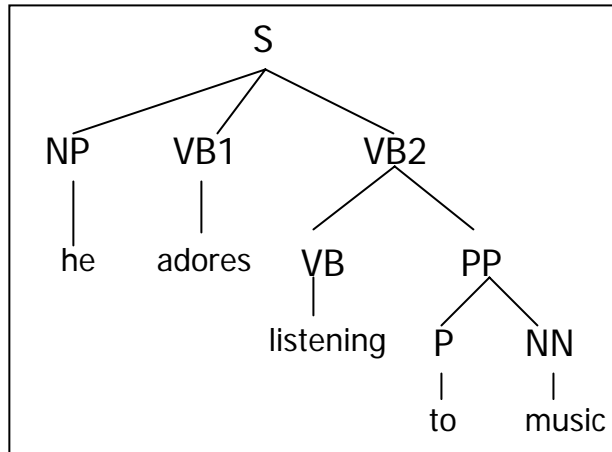
Reorder
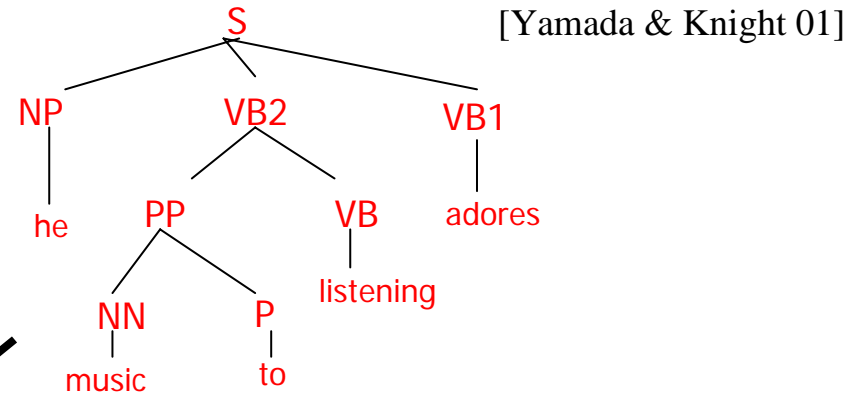
[Yamada & Knight 01]

Insert
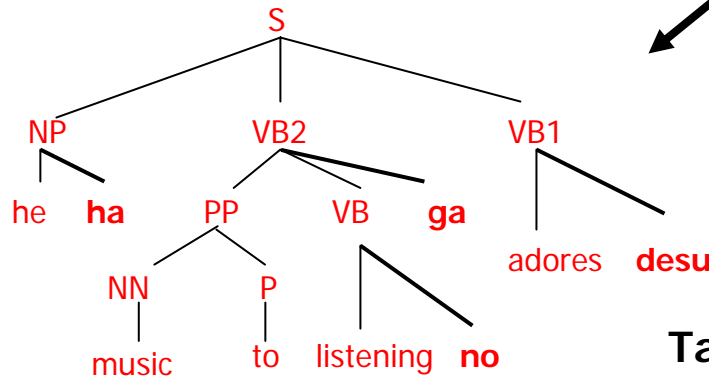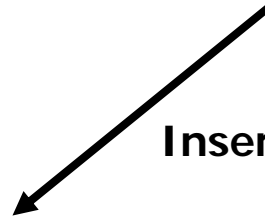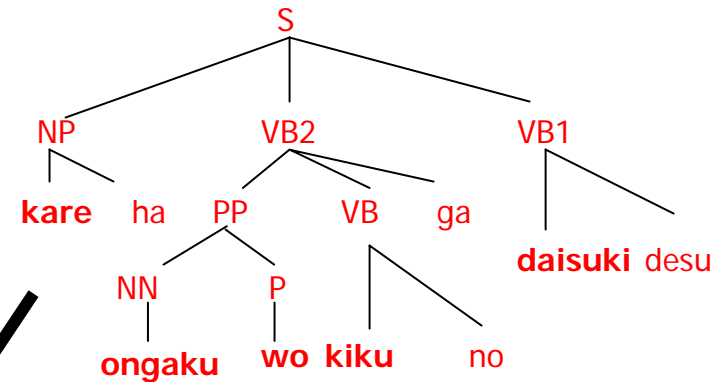
Translate

Take Leaves

Sentence(J) **Kare ha ongaku wo kiku no ga daisuki desu**

An Early Syntactic Model of Translation
[Yamada & Knight 01]

Parse (E)

[Yamada & Knight 01]

Reorder
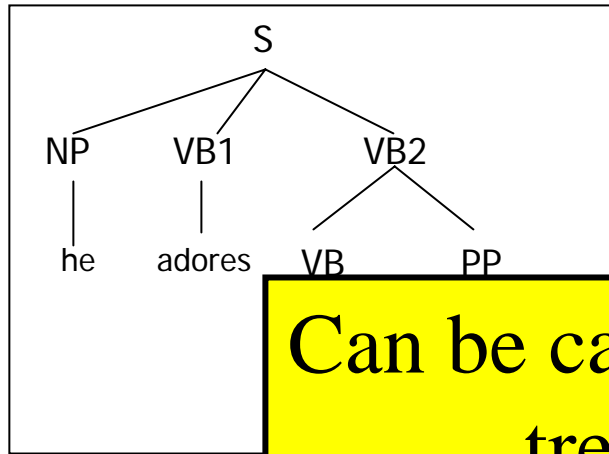
Translate
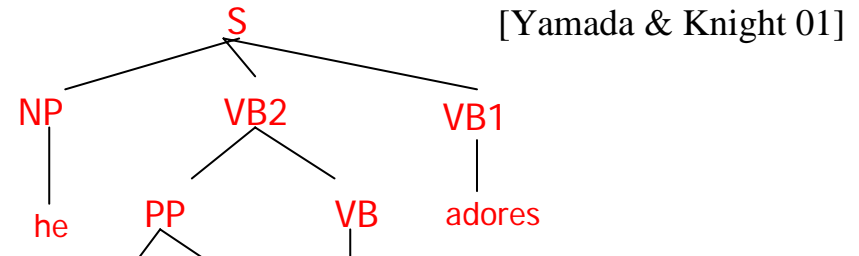
Take Leaves

Can be cast as a single 4-state
tree transducer.
[Graehl & Knight 04; Graehl, Knight & May 08]
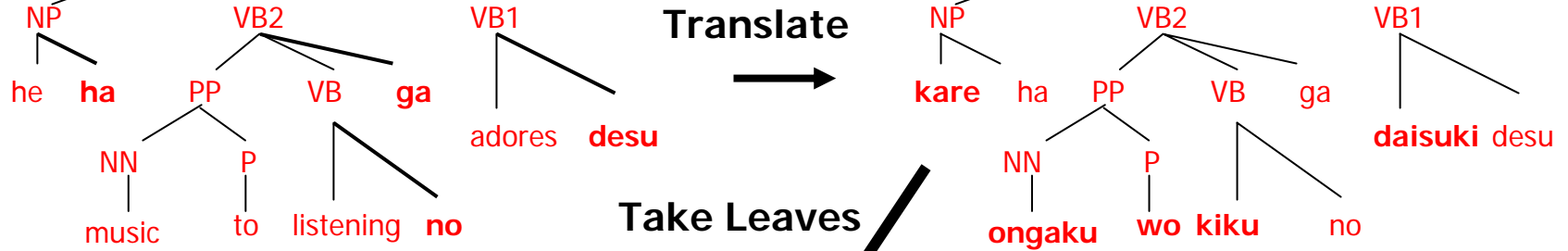
Sentence(J)   **Kare ha ongaku wo kiku no ga daisuki desu**

Tree Transducers are Expressive

Phrasal Translation
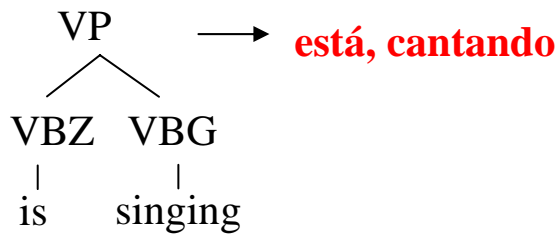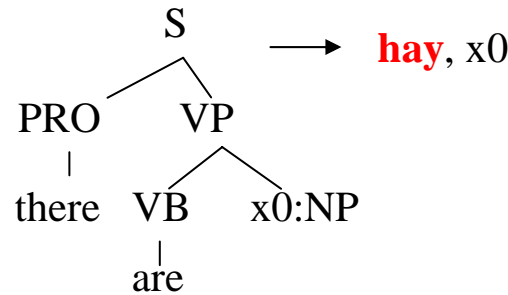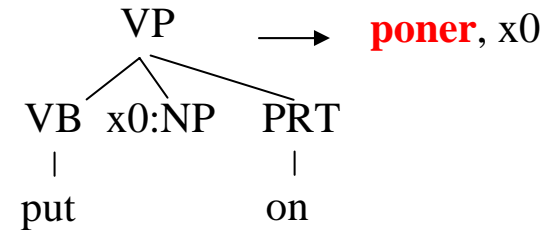
VP → **está, cantando**
  VBZ VBG
   | |
   is singing

Non-constituent Phrases

S → **hay**, x0
  PRO   VP
   |
  there VB   x0:NP
          |
         are

Non-contiguous Phrases

VP → **poner**, x0
  VB x0:NP PRT
   | |
  put   on

Context-Sensitive Word Insertion

NPB → x0
  DT   x0:NNS
   |
  the

Multilevel Re-Ordering

S → x1, x0, x2
  x0:NP   VP
       x1:VB   x2:NP2

Lexicalized Re-Ordering

NP → x1, 的 , x0
  x0:NP   PP
        P   x1:NP
        |
        of
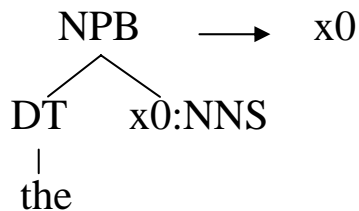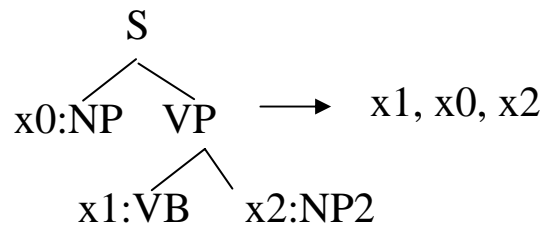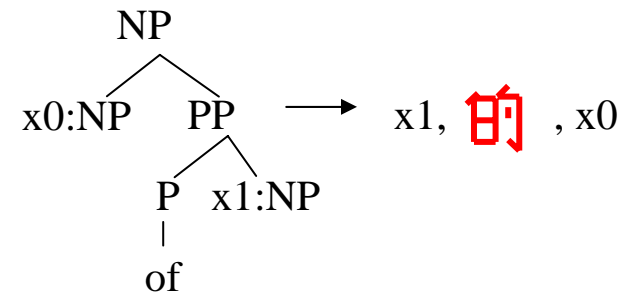
also QA, compression, paraphrasing, etc
most probabilistic tree-based models proposed 2000-2005 can be so cast

Limitations of the Top-Down Transducer Model

Who does John think Mary believes I saw? → John thinks Mary believes I saw *who*?
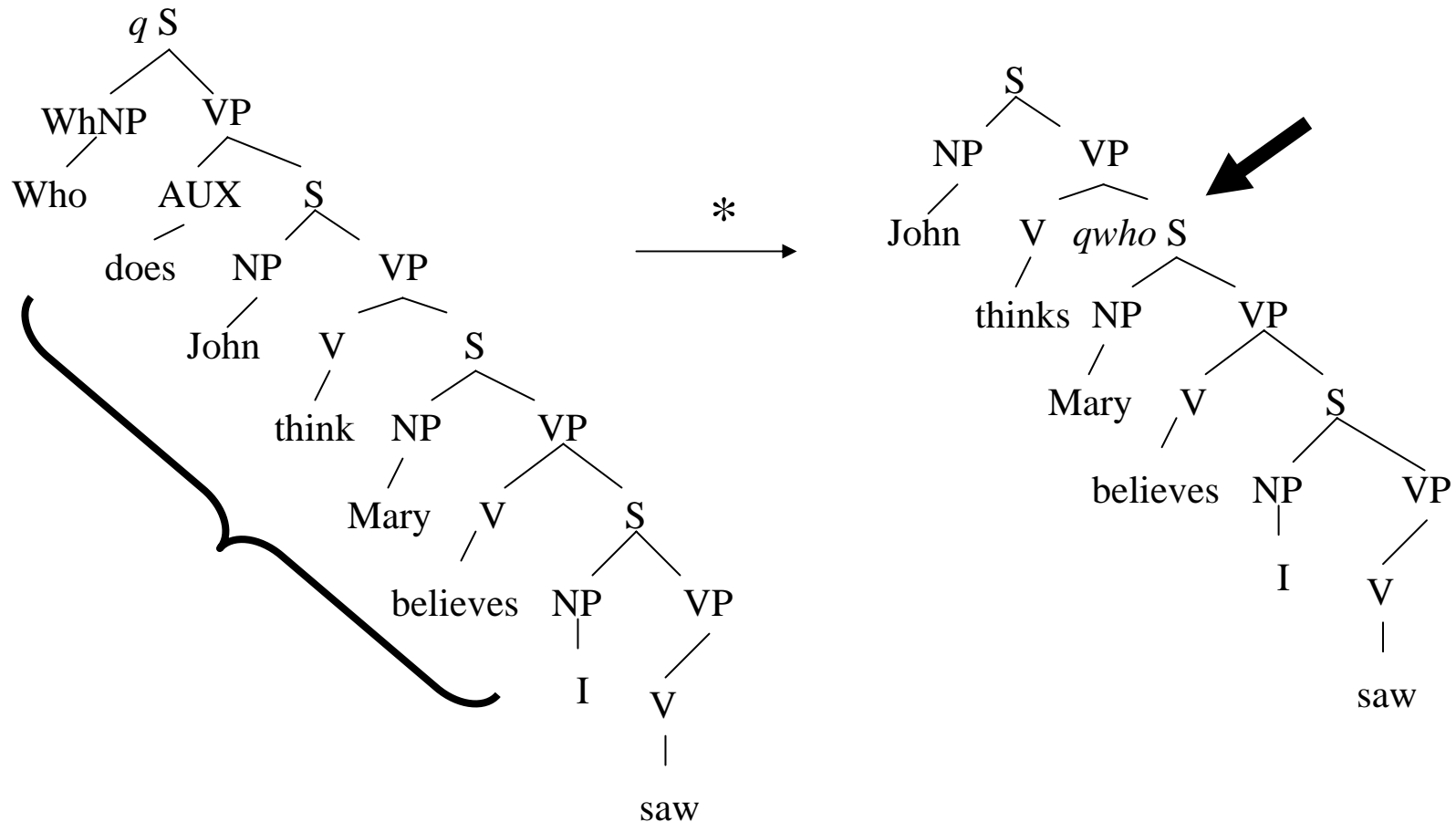
Limitations of the Top-Down Transducer Model

Who does John think Mary believes I saw? → John thinks Mary believes I saw *who*?
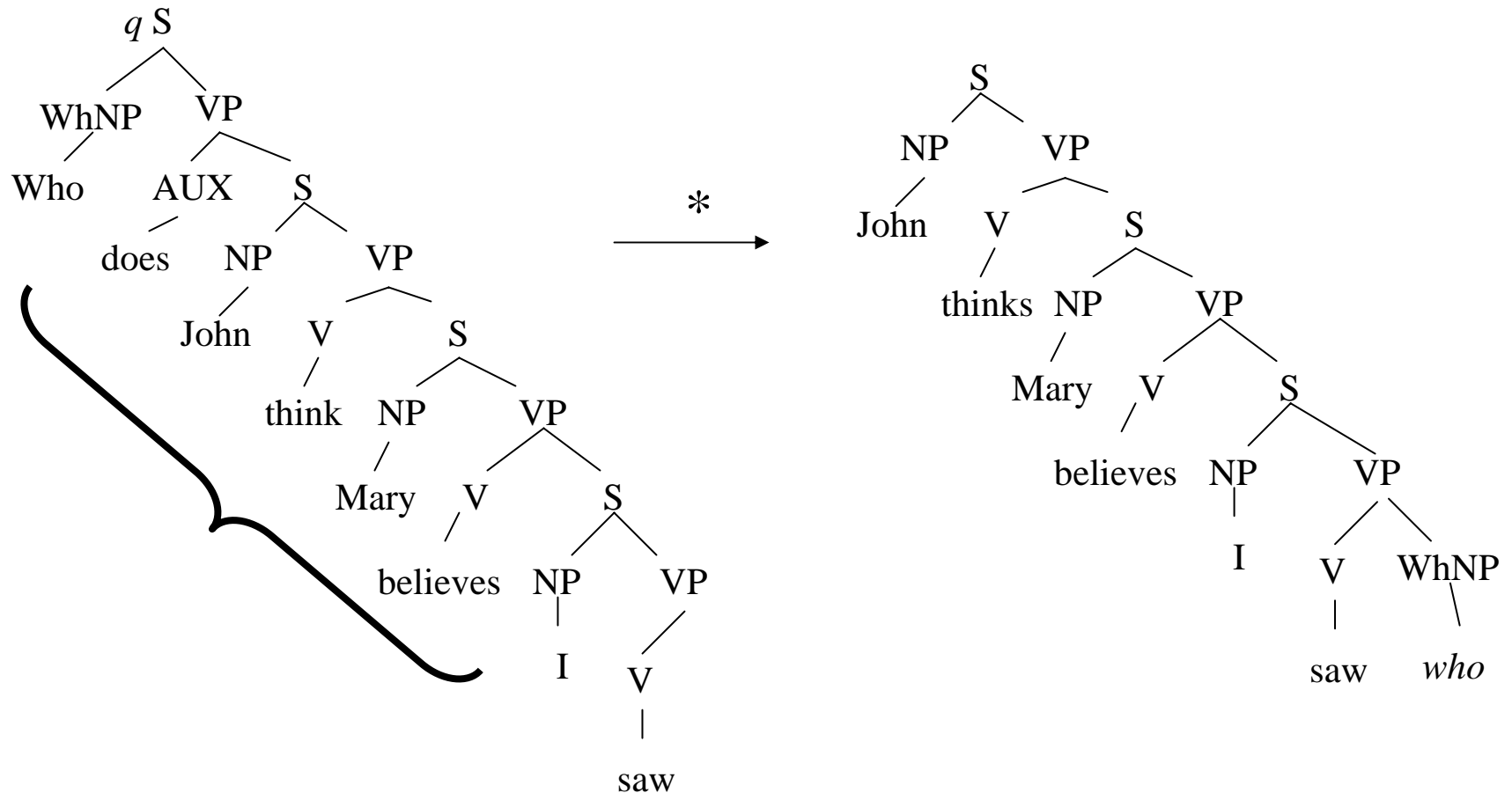
Limitations of the Top-Down Transducer Model

Whose blue dog does John think Mary believes I saw**?** → John thinks Mary believes I saw *whose blue dog*?
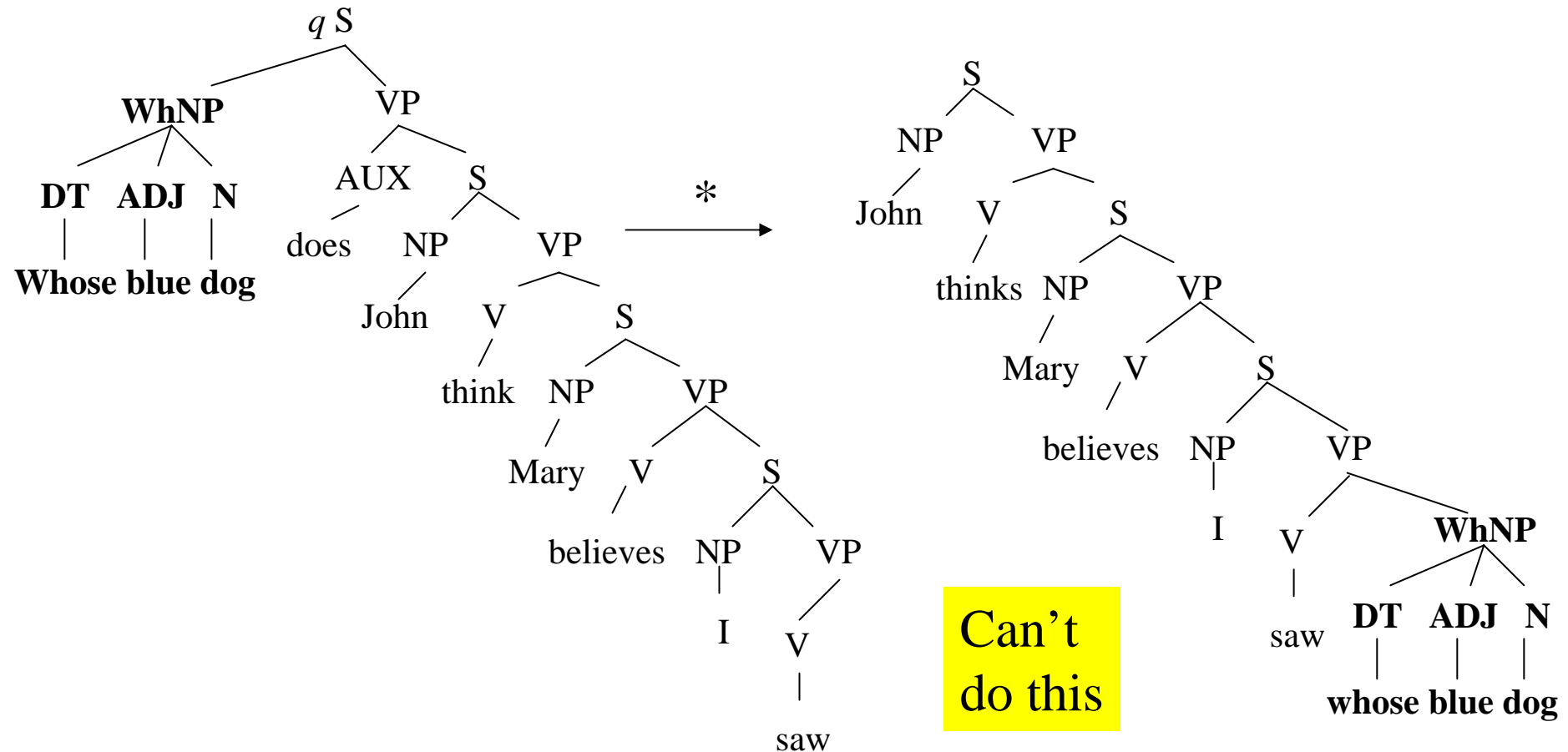
Computer-Friendly Format for Tree Transducer Rules

Phrasal Translation

```
        VP        ──→  está, cantando
       /  \
    VBZ   VBG
     |     |
     is  singing
```
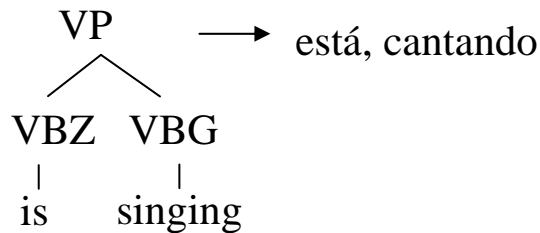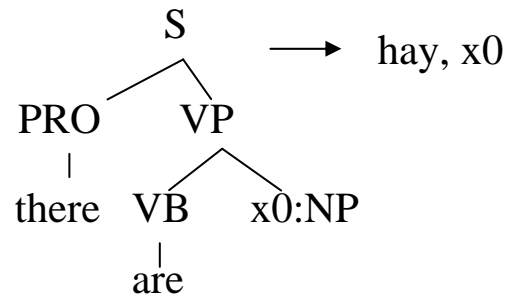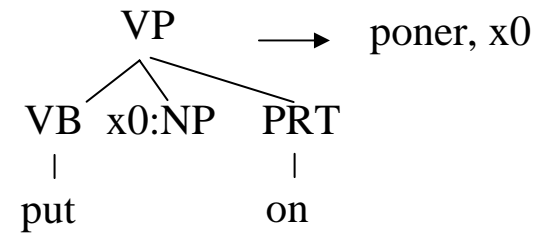
VP(VBZ(is), VBG(singing)) → está, cantando

Non-constituent Phrases

```
         S        ──→  hay, x0
        / \
     PRO   VP
      |    /  \
    there VB   x0:NP
          |
         are
```

S(PRO(there), VP(VB(are), x0:NP)) → hay, x0

Non-contiguous Phrases

```
        VP        ──→  poner, x0
       / | \
     VB x0:NP PRT
      |       |
     put      on
```

VP(VB(put), x0:NP, PRT(on)) → poner, x0

etree (includes estring)
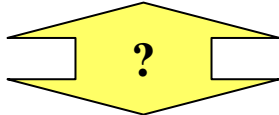
S
- VP
  - NP
    - VP
      - PP
        - NP
          - NP
          - NP

NP
- DT CD NNS
  - These 7 people
- VBP
  - include
- NNS
  - astronauts
- VBG
  - coming
- IN
  - from
- NNP
  - France
- CC
  - and
- NNP
  - Russia
- PUNC
  - .

?

cstring

这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

Tree Transformations

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP(includes) → 中包括
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航，员
9. PUNC(.) → .
10. NP(x0:DT, CD(7), NNS(people) → x0 , 7人
11. VP(VBG(coming), PP(IN(from), x0:NP)) → 来自 , x0
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0 , x1 , x2
14. VP(x0:VBP, x1:NP) → x0 , x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0 , x1, x2
16. NP(x0:NP, x1:VP) → x1 , 的 , x0
17. NP(DT("the"), x0:JJ, x1:NN) → x0 , x1

I made these rules up – they capture what is really happening in this Chinese sentence.

Contiguous phrase pair substitution rules (alignment templates)

Higher-level rules

Tree Transformations

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP(includes) → 中包括
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航，身
9. PUNC(.) → .
10. NP(x0:DT, CD(7), NNS(people) ...
11. VP(VBG(coming), PP(IN(from), x0:NP)) → 来自，x0
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0，x1，x2
14. VP(x0:VBP, x1:NP) → x0，x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0，x1，x2
16. NP(x0:NP, x1:VP) → x1，的，x0
17. NP(DT("the"), x0:JJ, x1:NN) → x0，x1

Both VBP("include") and VBP("includes") will translate to "中包括" in Chinese.

In decoding Chinese, "中包括" is ambiguous and can translate back as either VBP("include") or VBP("includes").

pair

tes)

Higher-level rules

Phrase pairs learned by alignment-templates that are relevant to this particular Chinese input sentence.

这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

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	included	by france		and the	the russian		international astronautical	of rapporteur .	
this	7 out	including the	from	the french	and the	russian	the fifth		.	
these	7 among	including from		the french and		of the russian	of	space	members	.
that	7 persons	including from	the	of france	and to	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 included		come from	france	and russia		astronautical	in	personnel	;
	7 philtrum	including those from	france and		russia	a space	member	member		
		including representatives from	france and the		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				

lattice

Table 1: #11# the seven - member crew includes astronauts from france and russia .

Only top 5 translations-per-Chinese-phrase are shown here – there are many more.

Phrase pairs learned by alignment-templates that are relevant to this particular Chinese input sentence.

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	and the russian	the fifth		.		
these	7 among	including from		the french and		of the russian	of	space	members	.
that	7 persons	including from the	of france	and to	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 included	come from	france	and russia		in	astronautical	personnel	;	
	7 philtrum	including those from	france and	russia		a space	astronaut	member		
		including representatives from	france and the	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				

lattice

Table 1: #11# the seven - member crew includes astronauts from france and russia .

Only top 5 translations-per-Chinese-phrase are shown here – there are many more.

Tree Transformations

1.   DT(these) → 这
2.   VBP(include) → 中包括
3.   VBP(includes) → 中包括
4.   NNP(France) → 法国
5.   CC(and) → 和
6.   NNP(Russia) → 俄罗斯
7.   IN(of) → 的
8.   NP(NNS(astronauts)) → 宇航，员
9.   PUNC(.) → .
10.   NP(x0:DT, CD(7), NNS(people) → x0，7人
11.   VP(VBG(coming), PP(IN(from), x0:NP)) → 来自，x0
12.   IN(from) → 来自
13.   NP(x0:NNP, x1:CC, x2:NNP) → x0，x1，x2
14.   VP(x0:VBP, x1:NP) → x0，x1
15.   S(x0:NP, x1:VP, x2:PUNC) → x0，x1, x2
16.   NP(x0:NP, x1:VP) → x1，的，x0
17.   NP(DT("the"), x0:JJ, x1:NN) → x0，x1

The phrase "coming from" translates to "来自" only if followed by an NP (whose translation is then placed to the right of "来自").

...rase pair

...es

...plates)

Higher-level rules

Tree Transformations

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP(includes) → 中包括
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航
9. PUNC(.) → .
10. NP(x0:DT, CD(7), NNS(peopl
11. VP(VBG(coming), PP(IN(from
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0 , x1 , x2
14. VP(x0:VBP, x1:NP) → x0 , x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0 , x1, x2
16. NP(x0:NP, x1:VP) → x1 , 的 , x0
17. NP(DT("the"), x0:JJ, x1:NN) → x0 , x1

Translate an English NP ("astronauts") modified by a gerund VP ("coming from France and Russia") as follows:
(1) translate the gerund VP,
(2) type the Chinese word "的",
(3) translate the NP.

In decoding Chinese, if we analyze
(1) some Chinese into an English NP &
(2) some other Chinese into an English VP and these two bits are separated by "的", then create an English NP(NP, VP) structure.

Higher-level rules

Tree Trans...

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP(includes) → 中包括
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航，员
9. PUNC(.) → .
10. NP(x0:DT, CD(7), NNS(people) → x0 ...
11. VP(VBG(coming), PP(IN(from), x0:NF...
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0 , x1 , x2
14. VP(x0:VBP, x1:NP) → x0 , x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0 , x1, x2
16. NP(x0:NP, x1:VP) → x1 , 的 , x0
17. NP(DT("the"), x0:JJ, x1:NN) → x0 , x1

To translate "the JJ NN", just translate the JJ and then translate the NN (drop "the").

When we are decoding Chinese, if we create an English JJ and an adjacent English NN, we can hook these together into an NP, and also insert the word "the."

Most frequent deficiency of lattices is the lack of critical English function words!

Higher-level rules

Tree Transformations

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP(includes) → 中包括
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航，员
9. PUNC(.) → .
10. NP(x0:DT, CD(7), NNS(people) → x0 , 7人
11. VP(VBG(coming), PP(IN(from), x0:NP)) → 来自 , x0
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0 , x1 , x2
14. VP(x0:VBP, x1:NP) → x0 , x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0 , x1, x2
16. NP(x0:NP, x1:VP) → x1 , 的 , x0
17. NP(DT("the"), x0:JJ, x1:NN) → x0 , x1

Note that this rule goes ahead and makes "astronauts" a full NP. Might be better to have two rules:

NNS(astronauts) → 宇航，员
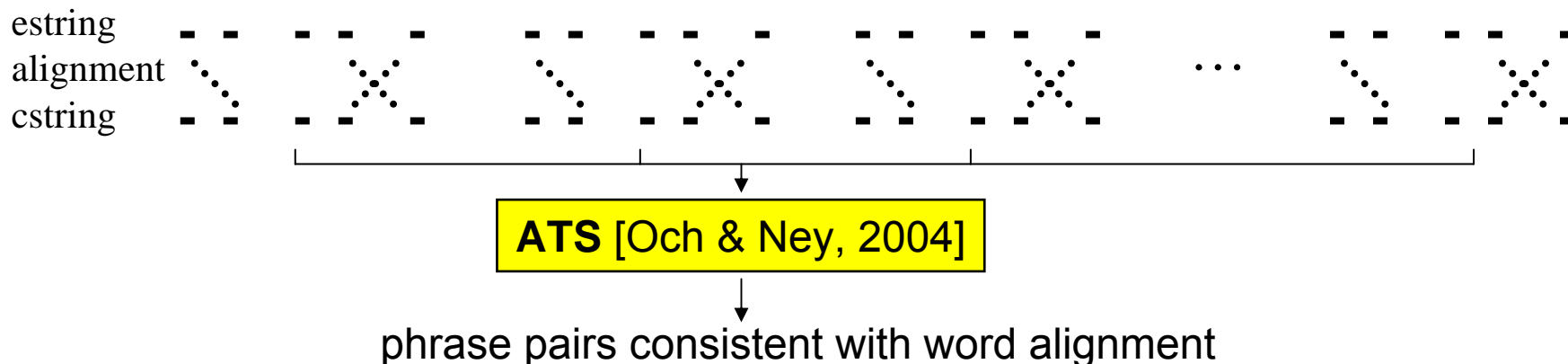NP(x0:NNS) → x0

Higher-level rules

Tree Transformations

1. DT(these) → 这
2. VBP(include) → 中包括
3. VBP(includes) → 中包括
4. NNP(France) → 法国
5. CC(and) → 和
6. NNP(Russia) → 俄罗斯
7. IN(of) → 的
8. NP(NNS(astronauts)) → 宇航，员
9. PUNC(.) → .
10. NP(x0:DT, CD(7), NNS(people) → x0 , 7人
11. VP(VBG(coming), PP(IN(from), x0:NP)) → 来自 , x0
12. IN(from) → 来自
13. NP(x0:NNP, x1:CC, x2:NNP) → x0 , x1 , x2
14. VP(x0:VBP, x1:NP) → x0 , x1
15. S(x0:NP, x1:VP, x2:PUNC) → x0 , x1, x2
16. NP(x0:NP, x1:VP) → x1 , 的 , x0
17. NP(DT("the"), x0:JJ, x1:NN) → x0 , x1

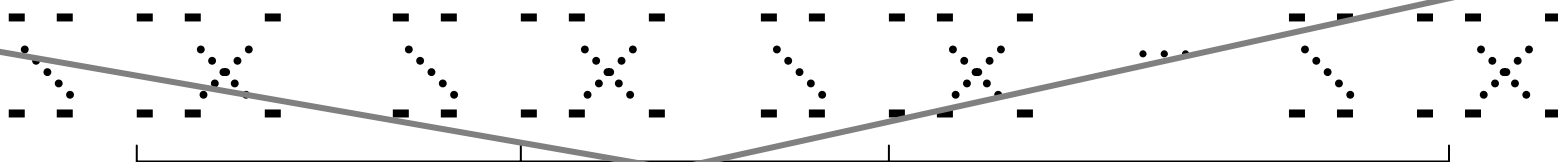Okay, these rules look interesting.

It would be cool if we could acquire rules like these from data!!

Phrase-Based and Syntax-Based Pattern Extraction

estring
alignment
cstring

ATS [Och & Ney, 2004]

phrase pairs consistent with word alignment

Phrase-Based and Syntax-Based Pattern Extraction

estring
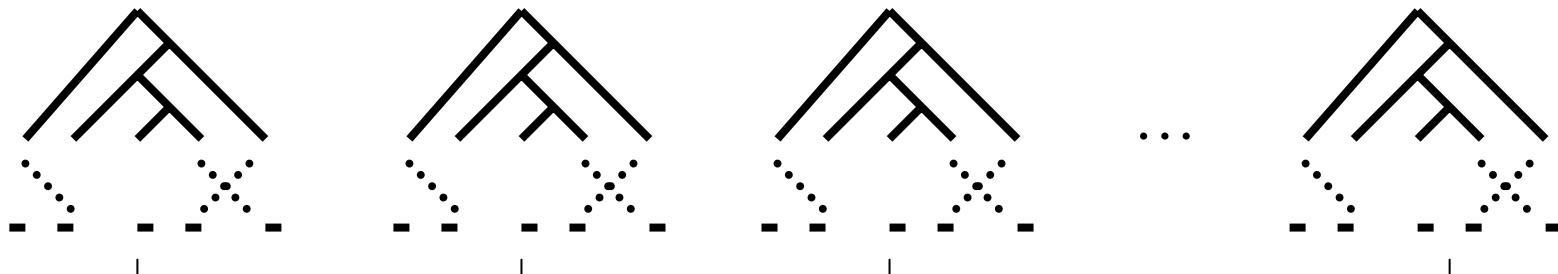
alignment

cstring

ATS [Och & Ney, 2004]

phrase pairs consistent with word alignment

etree

alignment

cstring

GHKM [Galley et al 2004, 2006]

syntax transformation rules consistent with word alignment

Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)

i felt obliged to do my part

我 有 责任 尽 一份 力

RULES ACQUIRED:

VBD(felt)   → 有
VBN(obliged)   → 责任
VB(do)   → 尽
NN(part)   → 一份
NN(part)   → 一份 力
VP-C(x0:VBN x1:SG-C) → x0 x1
VP(TO(to) x0:VP-C)   → x0
  …
S(x0:NP-C x1:VP)   → x0 x1

Tree Transducers Can be Extracted from Data
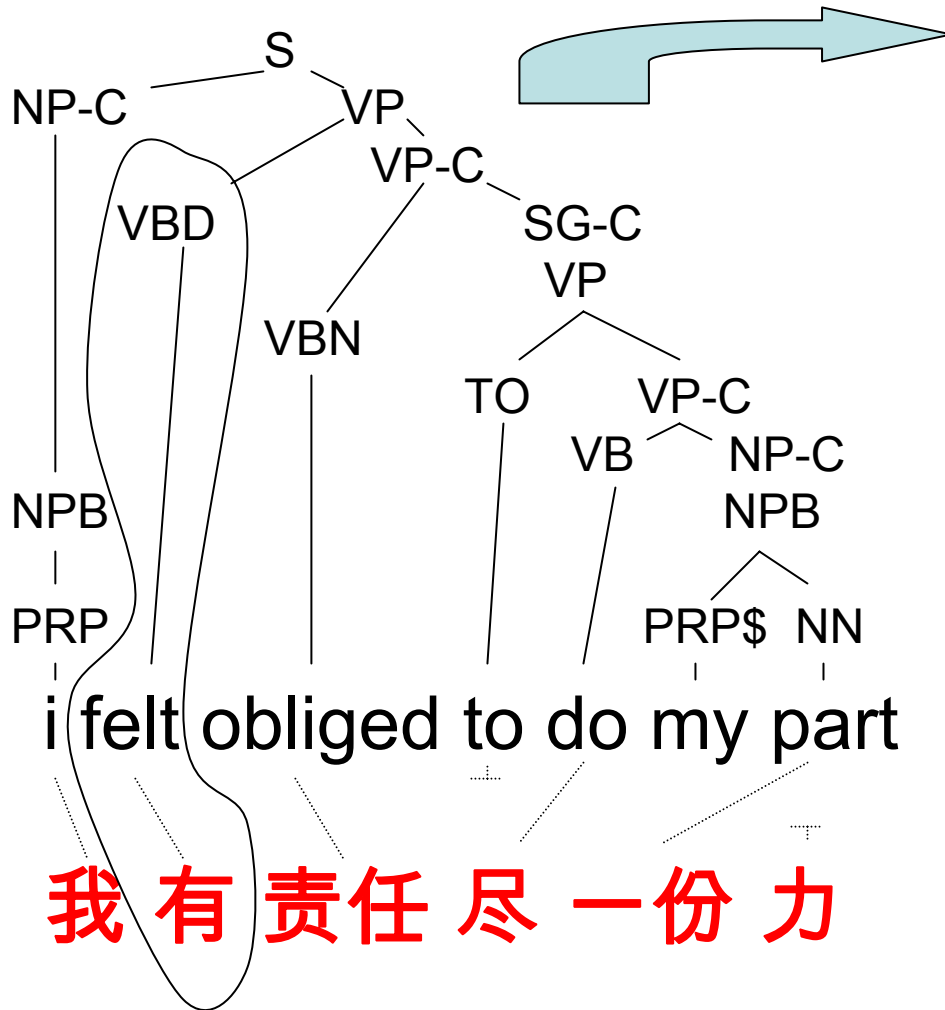## (Galley, Hopkins, Knight, Marcu, 2004)

RULES ACQUIRED:

VBD(felt)	→	有
VBN(obliged)	→	责任
VB(do)	→	尽
NN(part)	→	一份
NN(part)	→	一份 力
VP-C(x0:VBN x1:SG-C)	→	x0 x1
VP(TO(to) x0:VP-C)	→	x0
…		
S(x0:NP-C x1:VP)	→	x0 x1

i felt obliged to do my part

我 有 责任 尽 一份 力

Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)

RULES ACQUIRED:

VBD(felt)	→	有
VBN(obliged)	→	责任
VB(do)	→	尽
NN(part)	→	一份
NN(part)	→	一份 力
VP-C(x0:VBN x1:SG-C)	→	x0 x1
VP(TO(to) x0:VP-C)	→	x0
…		
S(x0:NP-C x1:VP)	→	x0 x1

Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)
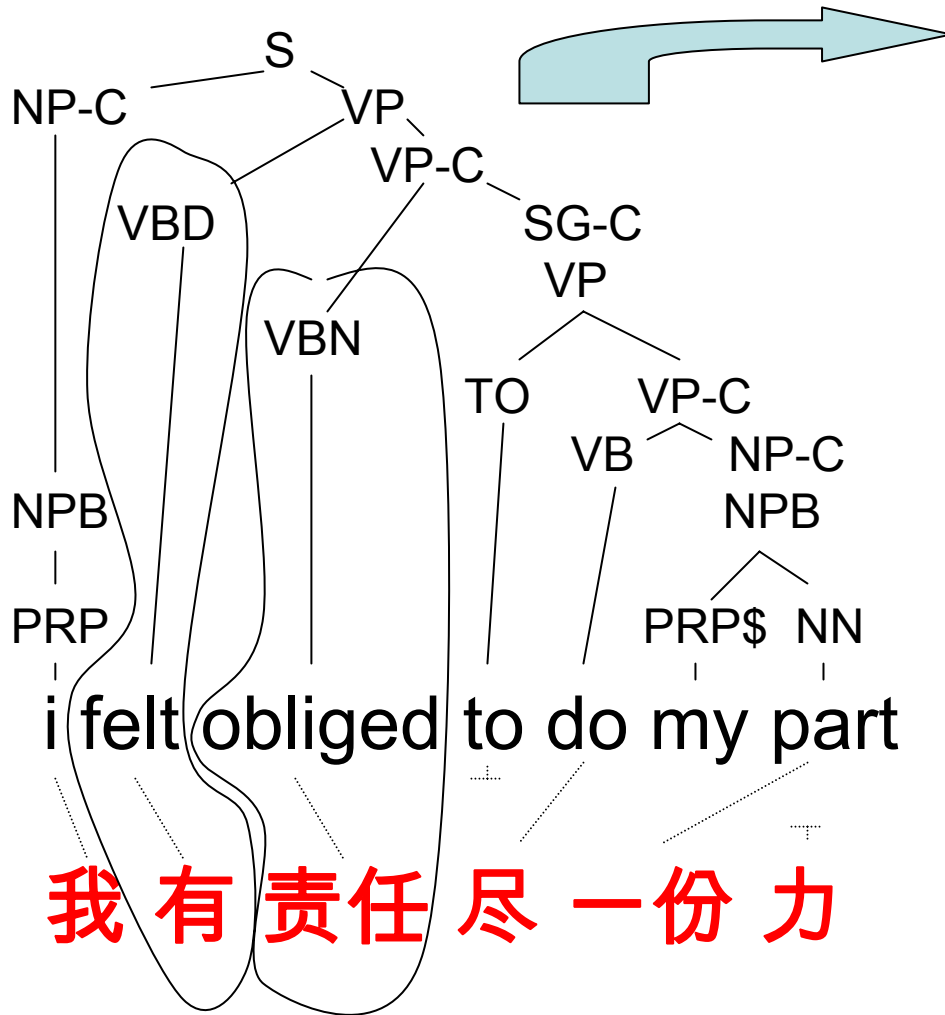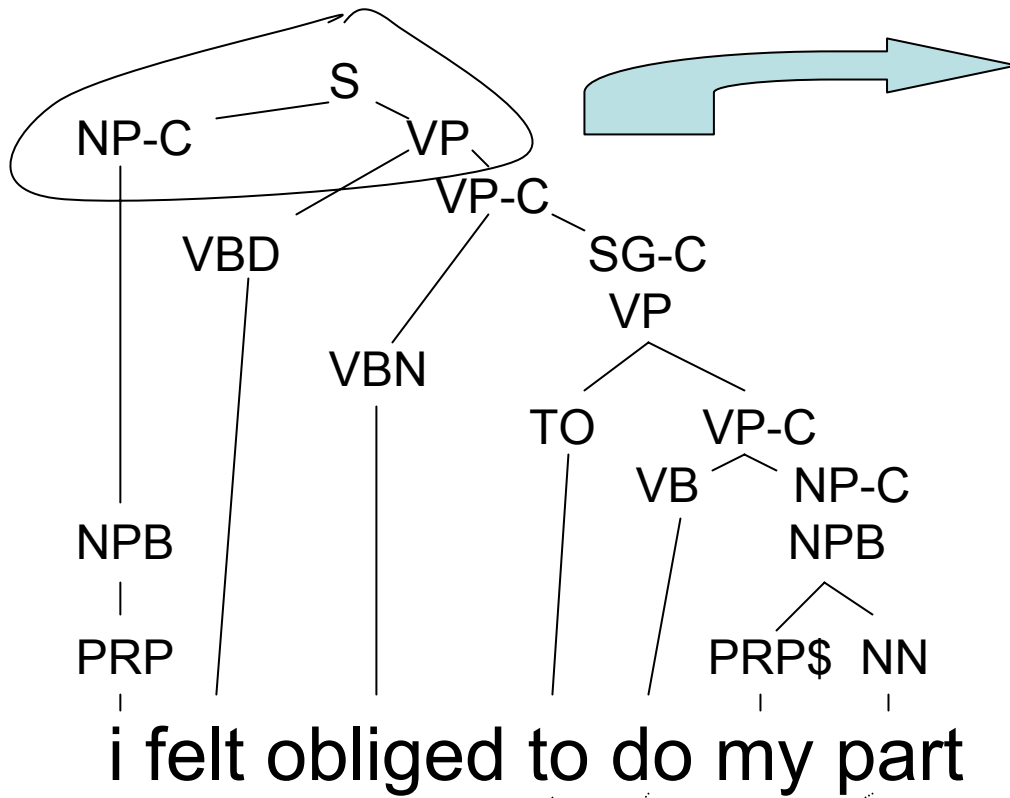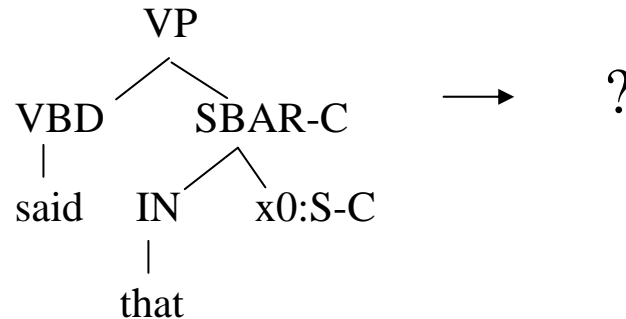

RULES ACQUIRED:

VBD(felt)   → 有
VBN(obliged)   → 责任
VB(do)   → 尽
NN(part)   → 一份
NN(part)   → 一份 力
VP-C(x0:VBN x1:SG-C) → x0 x1
VP(TO(to) x0:VP-C)   → x0
  …
S(x0:NP-C x1:VP)   → x0 x1

i felt obliged to do my part

我 有 责任 尽 一份 力

There is a unique tiling that
identifies minimal translation units.

Sample "said that" rules


```
0.57    VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 说 , x0
0.09    VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 说 x0
0.02    VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 他 说 , x0
0.02    VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 指出 , x0
0.02    VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> x0
0.01    VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 表示 x0
0.01    VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 说 , x0 的
```

Sample "NP-from-NP" rules

NP-C
x0:NPB   PP $\longrightarrow$?
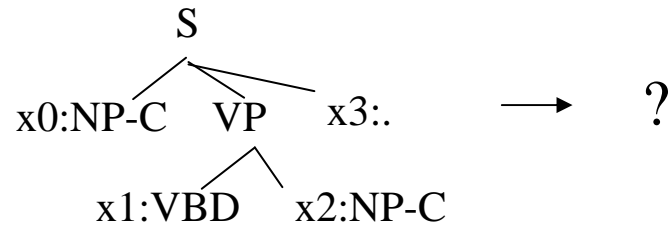   IN   x1:NP-C
        |
       from

0.27	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x1 x0
0.15	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 来自 x1 x0
0.06	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x1 的 x0
0.06	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 从 x1 x0
0.06	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 来自 x1 的 x0
0.02	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x0 从 x1
0.01	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 自 x1 x0
0.01	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x1 x0 ,

Sample SVO rules

```
           S
         / | \
   x0:NP-C  VP  x3:.         ——→      ?
          /  \
     x1:VBD  x2:NP-C
```

CHINESE / ENGLISH

0.82   S(x0:NP-C VP(x1:VBD x2:NP-C) x3:.) -> x0 x1 x2 x3
0.02   S(x0:NP-C VP(x1:VBD x2:NP-C) x3:.) -> x0 x1 , x2 x3
0.01   S(x0:NP-C VP(x1:VBD x2:NP-C) x3:.) -> x0 , x1 x2 x3

ARABIC / ENGLISH

0.54   S(x0:NP-C VP(x1:VBD x2:NP-C) x3:.) -> x0 x1 x2 x3
0.44   S(x0:NP-C VP(x1:VBD x2:NP-C) x3:.) -> x1 x0 x2 x3

Extensions to Rule Extraction from Data [Galley et al 06]

Enumerate all ways of dealing with unaligned Chinese words.

Generate rule counts which can be normalized into probabilities.
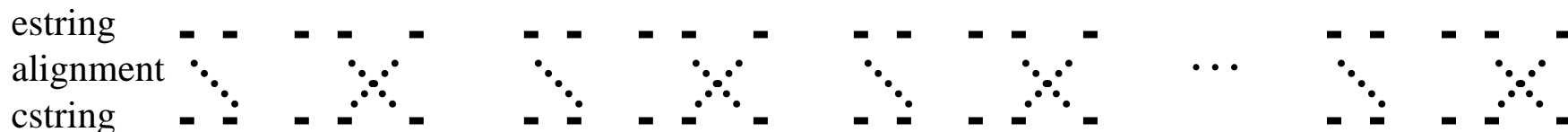
Language Models

- Syntax-based Language Model
  - Assigns P(tree)
    - [Collins 97; Charniak 01]
  - NOTE: Unlike parser, must be trained on domain data

- Ngram Language Model
  - Standard trigram model
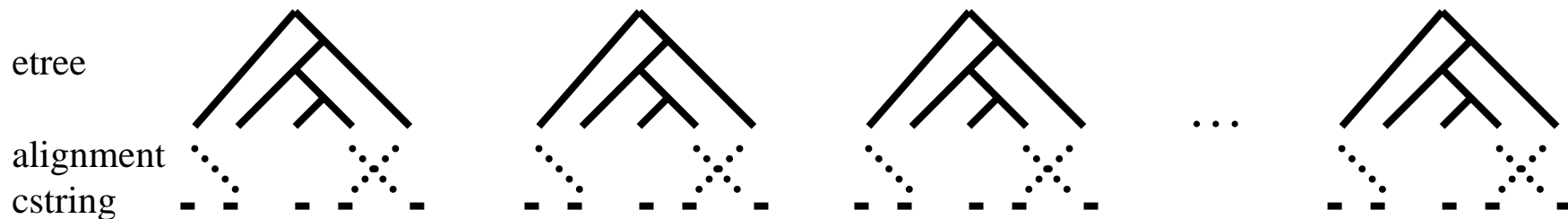 - Only judges a tree by its leaves

BREAK

- When We Come Back:
  - Review of syntax-based translation models
  - Syntax-based decoding
  - Is syntax harmful?
    - yes
    - what can be done
  - Sample outputs
  - Open problems
  - Connections to automata
  - Conclusions
  - Discussion

Phrase-Based and Syntax-Based Pattern Extraction

estring
alignment
cstring

ATS [Och & Ney, 2004]

phrase pairs consistent with word alignment

etree

alignment
cstring

GHKM [Galley et al 2004, 2006]
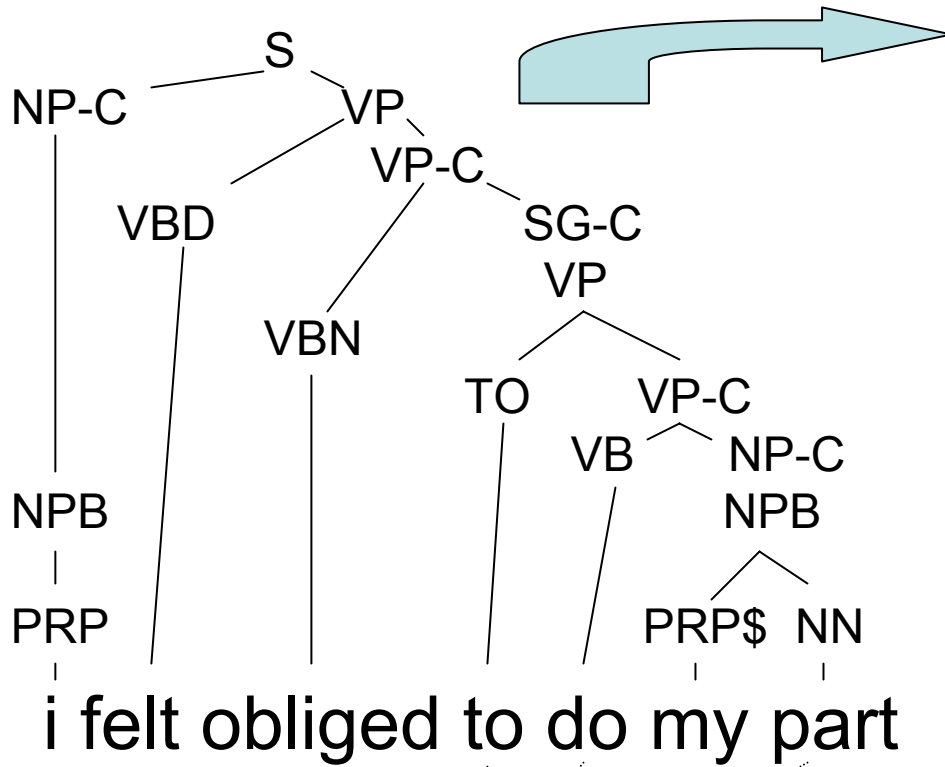
syntax transformation rules consistent with word alignment

Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)
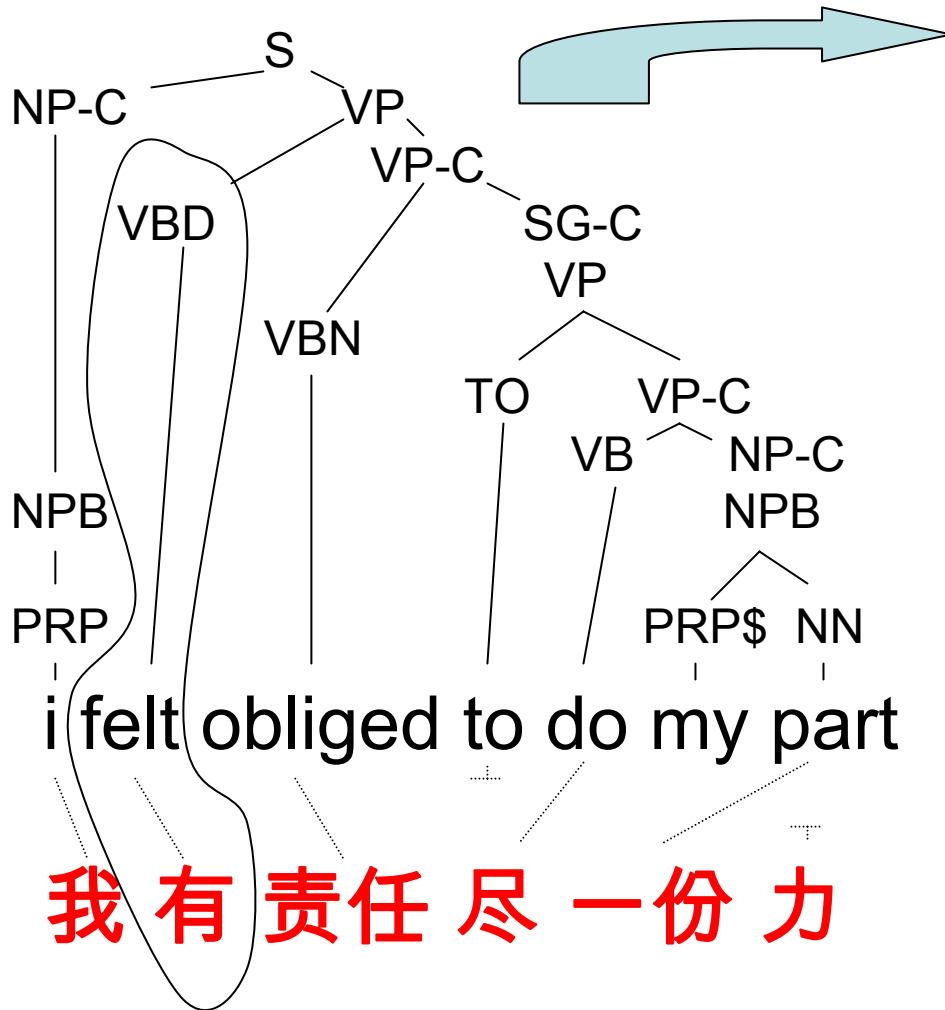

RULES ACQUIRED:

VBD(felt)   → 有
VBN(obliged)   → 责任
VB(do)   → 尽
NN(part)   → 一份
NN(part)   → 一份 力
VP-C(x0:VBN x1:SG-C) → x0 x1
VP(TO(to) x0:VP-C)   → x0
 …
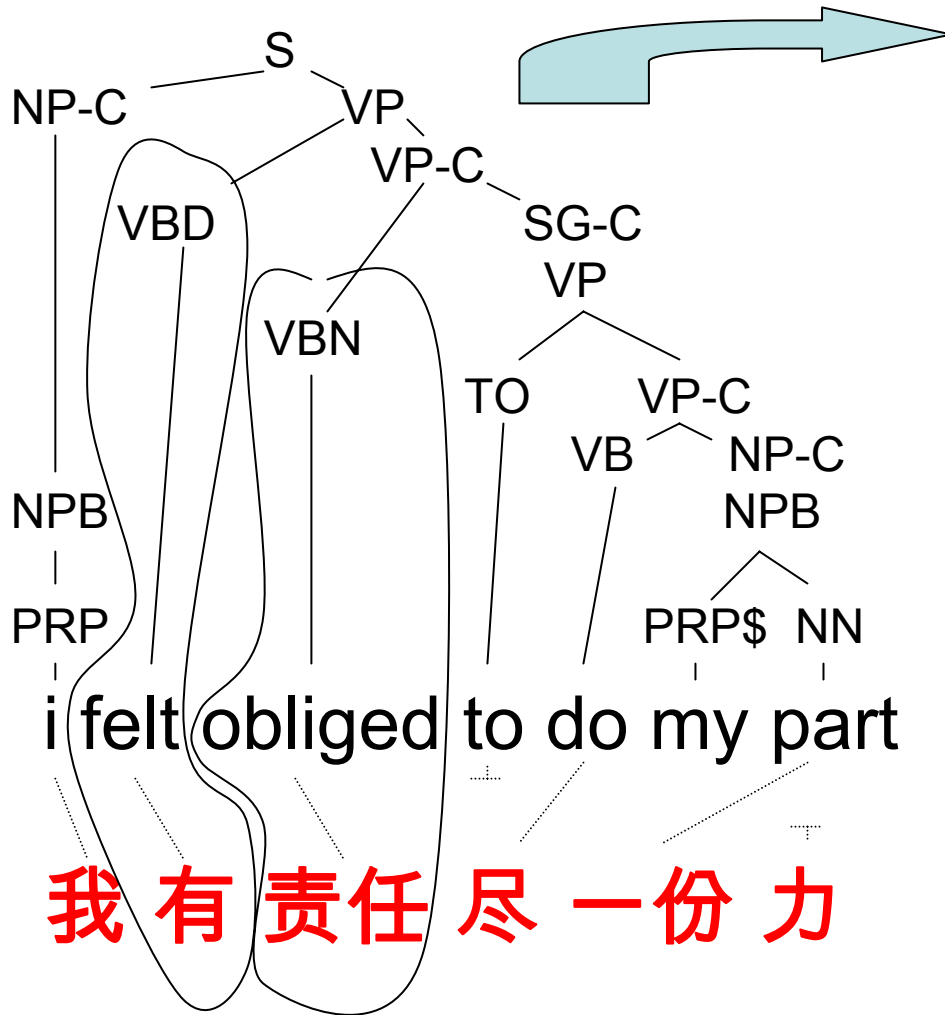S(x0:NP-C x1:VP)   → x0 x1

Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)

RULES ACQUIRED:

VBD(felt) → 有
VBN(obliged) → 责任
VB(do) → 尽
NN(part) → 一份
NN(part) → 一份 力
VP-C(x0:VBN x1:SG-C) → x0 x1
VP(TO(to) x0:VP-C) → x0
 …
S(x0:NP-C x1:VP) → x0 x1

i felt obliged to do my part

我 有 责任 尽 一份 力

Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)

RULES ACQUIRED:

VBD(felt)	→	有
VBN(obliged)	→	责任
VB(do)	→	尽
NN(part)	→	一份
NN(part)	→	一份 力
VP-C(x0:VBN x1:SG-C)	→	x0 x1
VP(TO(to) x0:VP-C)	→	x0
…		
S(x0:NP-C x1:VP)	→	x0 x1

Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)

S

NP-C   VP

VP-C

VBD   SG-C

VP

VBN

TO   VP-C

VB   NP-C

NPB   NPB

PRP   PRP$ NN

i felt obliged to do my part
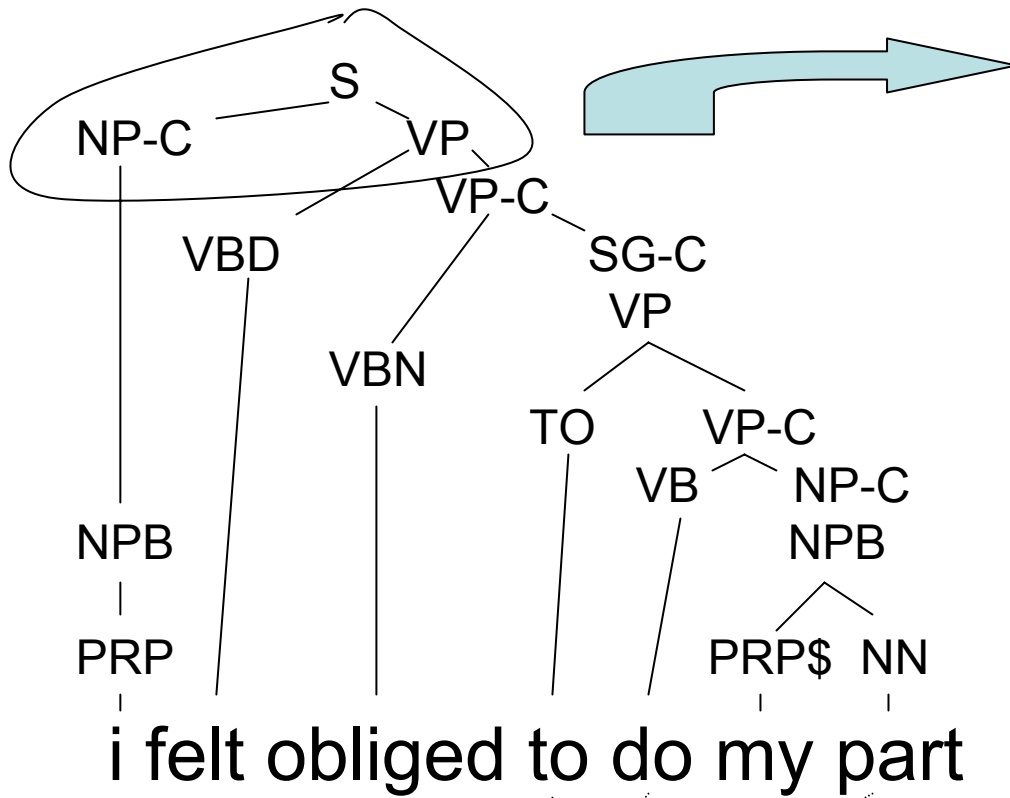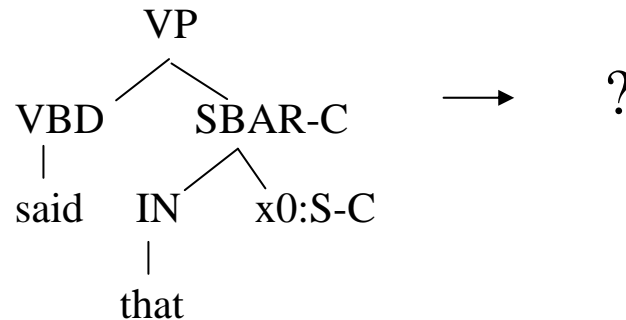
我 有 责任 尽 一份 力

RULES ACQUIRED:

VBD(felt)   → 有
VBN(obliged)   → 责任
VB(do)   → 尽
NN(part)   → 一份
NN(part)   → 一份 力
VP-C(x0:VBN x1:SG-C) → x0 x1
VP(TO(to) x0:VP-C)   → x0
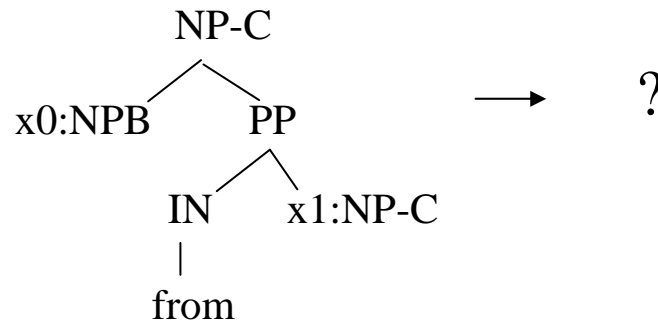 …
S(x0:NP-C x1:VP)   → x0 x1

Sample "said that" rules

```
        VP
       /  \
    VBD    SBAR-C              →        ?
     |    /     \
   said  IN    x0:S-C
         |
        that
```

0.57   VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 说 , x0

0.09   VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 说 x0

0.02   VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 他 说 , x0

0.02   VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 指出 , x0
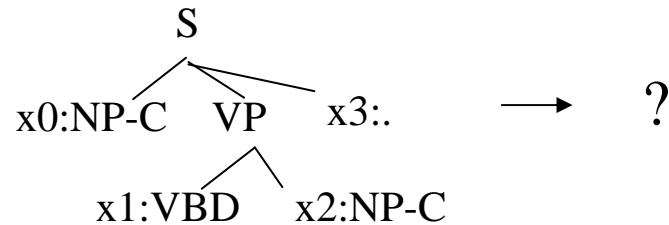
0.02   VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> x0

0.01   VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 表示 x0

0.01   VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 说 , x0 的

Sample "NP-from-NP" rules

NP-C
x0:NPB   PP
          IN   x1:NP-C
          |
          from
          →   ?

0.27	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x1 x0
0.15	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 来自 x1 x0
0.06	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x1 的 x0
0.06	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 从 x1 x0
0.06	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 来自 x1 的 x0
0.02	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x0 从 x1
0.01	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 自 x1 x0
0.01	NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x1 x0 ,

Sample SVO rules

S
x0:NP-C   VP   x3:.   →   ?
x1:VBD   x2:NP-C

CHINESE / ENGLISH

0.82   S(x0:NP-C VP(x1:VBD x2:NP-C) x3:.) -> x0 x1 x2 x3
0.02   S(x0:NP-C VP(x1:VBD x2:NP-C) x3:.) -> x0 x1 , x2 x3
0.01   S(x0:NP-C VP(x1:VBD x2:NP-C) x3:.) -> x0 , x1 x2 x3

ARABIC / ENGLISH

0.54   S(x0:NP-C VP(x1:VBD x2:NP-C) x3:.) -> x0 x1 x2 x3
0.44   S(x0:NP-C VP(x1:VBD x2:NP-C) x3:.) -> x1 x0 x2 x3

Hiero (Chiang 05, 07)

- Phrase pairs with variables
  - e.g., "of X ←→ X de"

- Hierarchical decoding
  - the X itself could be created via other phrase pairs with variables

- Only one syntactic symbol in rules
  - X

- Translation patterns can be extracted without syntactically parses of the training data

Hiero Grammar Extraction
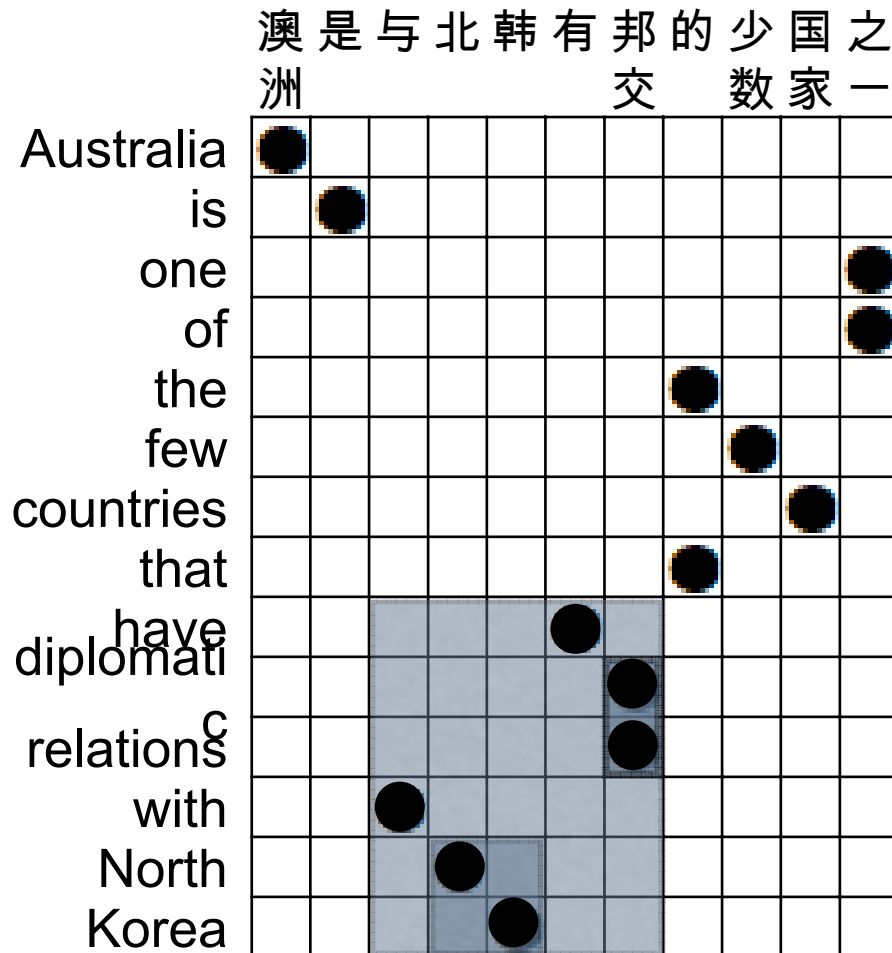
X(diplomatic relations)
→ 邦交

X(North Korea)
→ 北 韩

X(have diplomatic relations with North Korea)
→ 与 北 韩 有 邦交

X(have x0:X with x1:X)
→ 与 x1 有 x0

In Hiero literature, this rule is written in *synchronous grammar* format:
(X → 与 X1 有 X2,
X → have X2 with X1)

Sample Hiero rules
(in tree transducer format)

X('s) → 的

X(the x0:X of x1:X) → x1 的 x0

X(the x0:X that x1:X) → x1 的 x0

X(in) → 在

X(under x0:X) → 在 x0 下

X(before x0:X) → 在 x0 前

X(x0:X this year) → 今年 x0

X(one of x0:X) → x0 之一

X(president x0:X) → x0 总统

Decoding
Reminder: 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 ut	including the	from	the french	and the russian		the fifth		.	
these	7 among	including from		the french and		of the russian	of	space	members	.
that	7 persons	including from the		of france	and to	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 included	come from	france		and russia		in	astronautical	personnel	;
	7 philtrum	including those from	france and	russia		a space	astronaut	member	
		including representatives from	france and the		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			

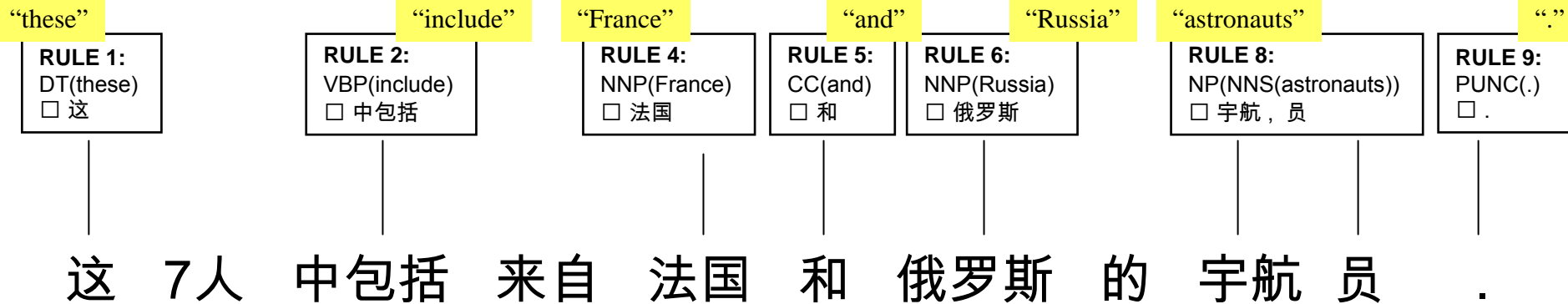Table 1: #11# the seven - member crew includes astronauts from france and russia .

Scoring: Try to use phrase pairs that have been frequently observed.
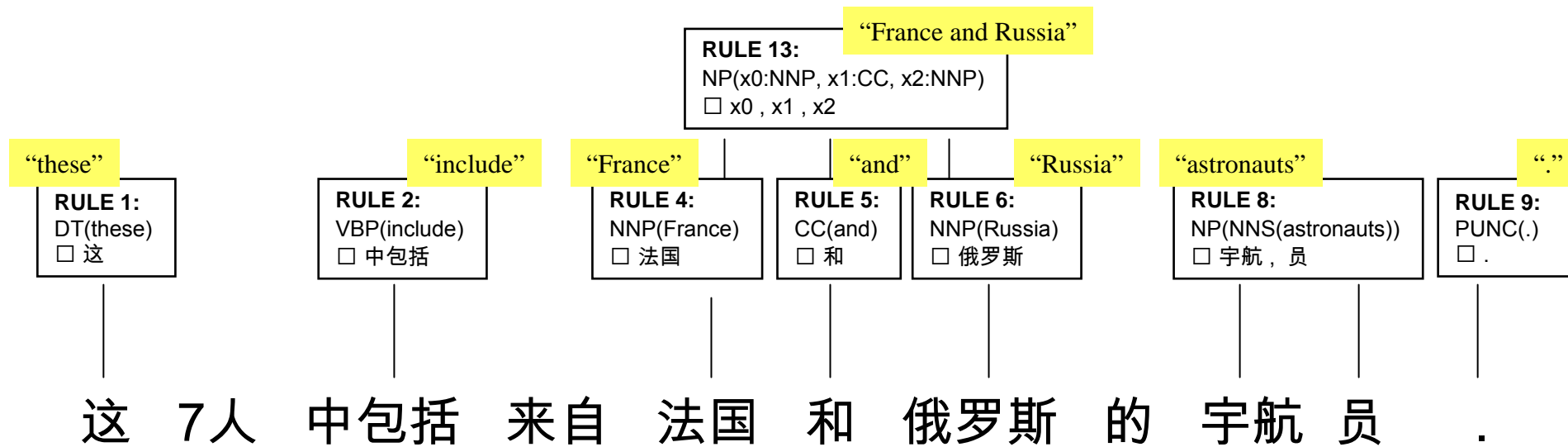Try to output a sentence with frequent English word sequences.

Syntax-Based Decoding

- Bottom-up CKY parser
- Builds English constituents on top of Chinese spans
- Record of rule applications (the derivation) provides information to construct English tree
- Returns k-best trees
- Same decoder can handle syntax translation rules and Hiero rules
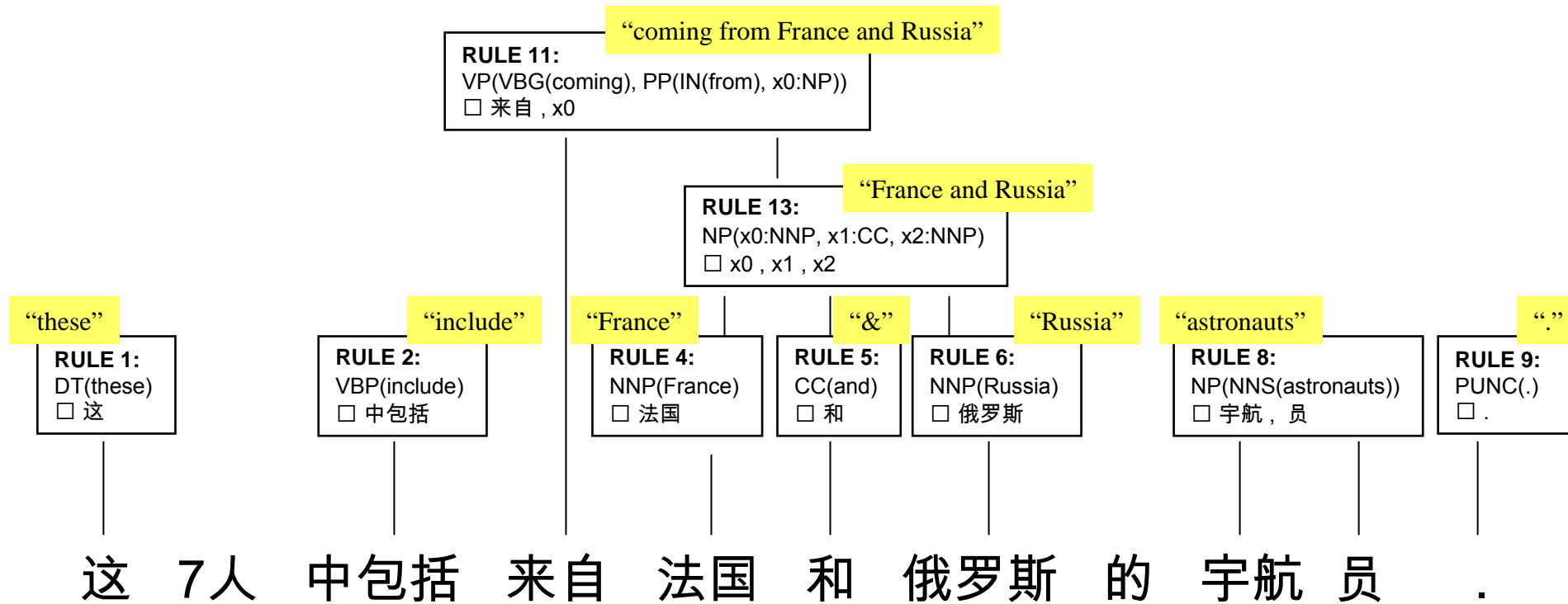
Syntax-Based Decoding

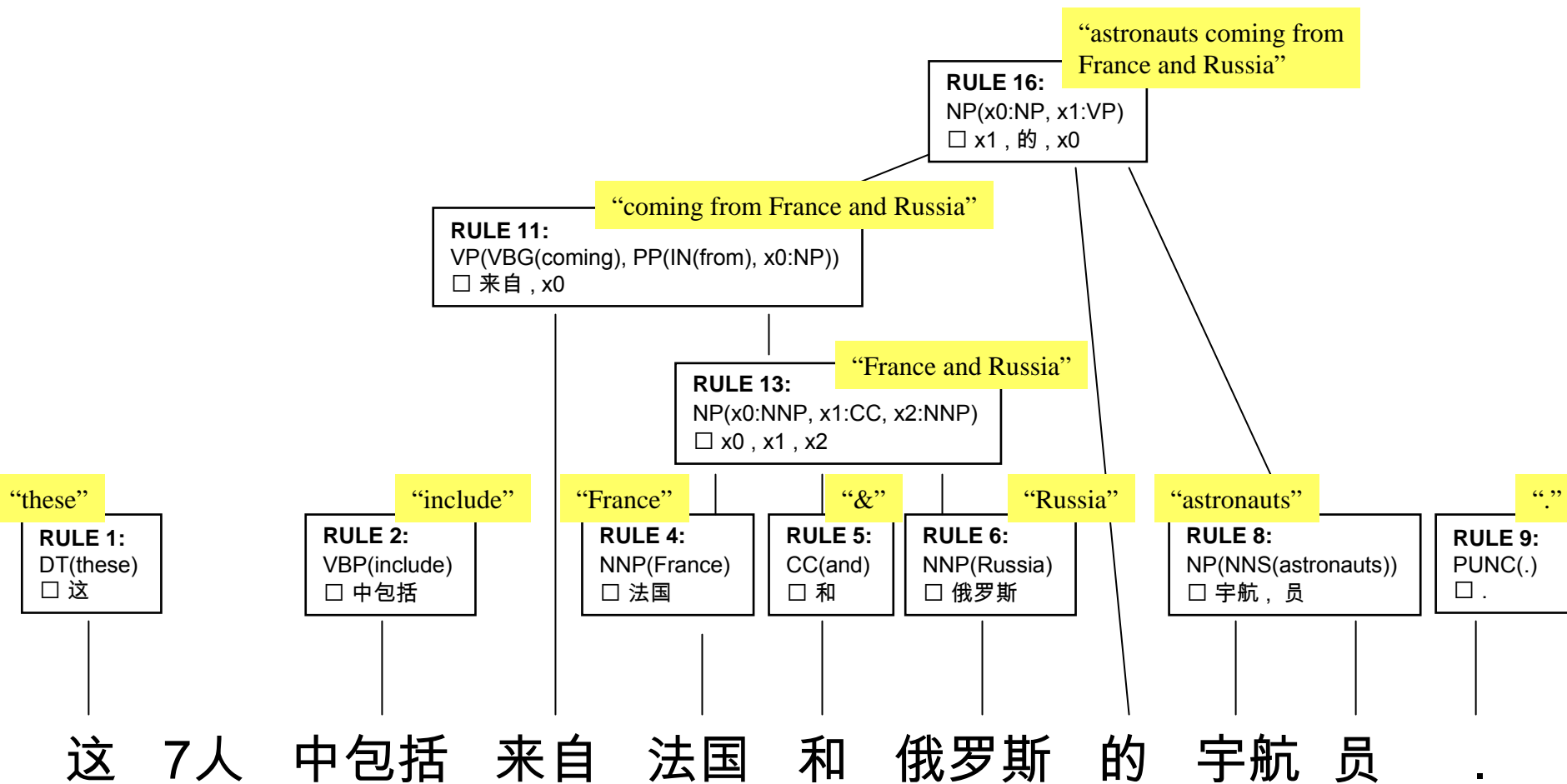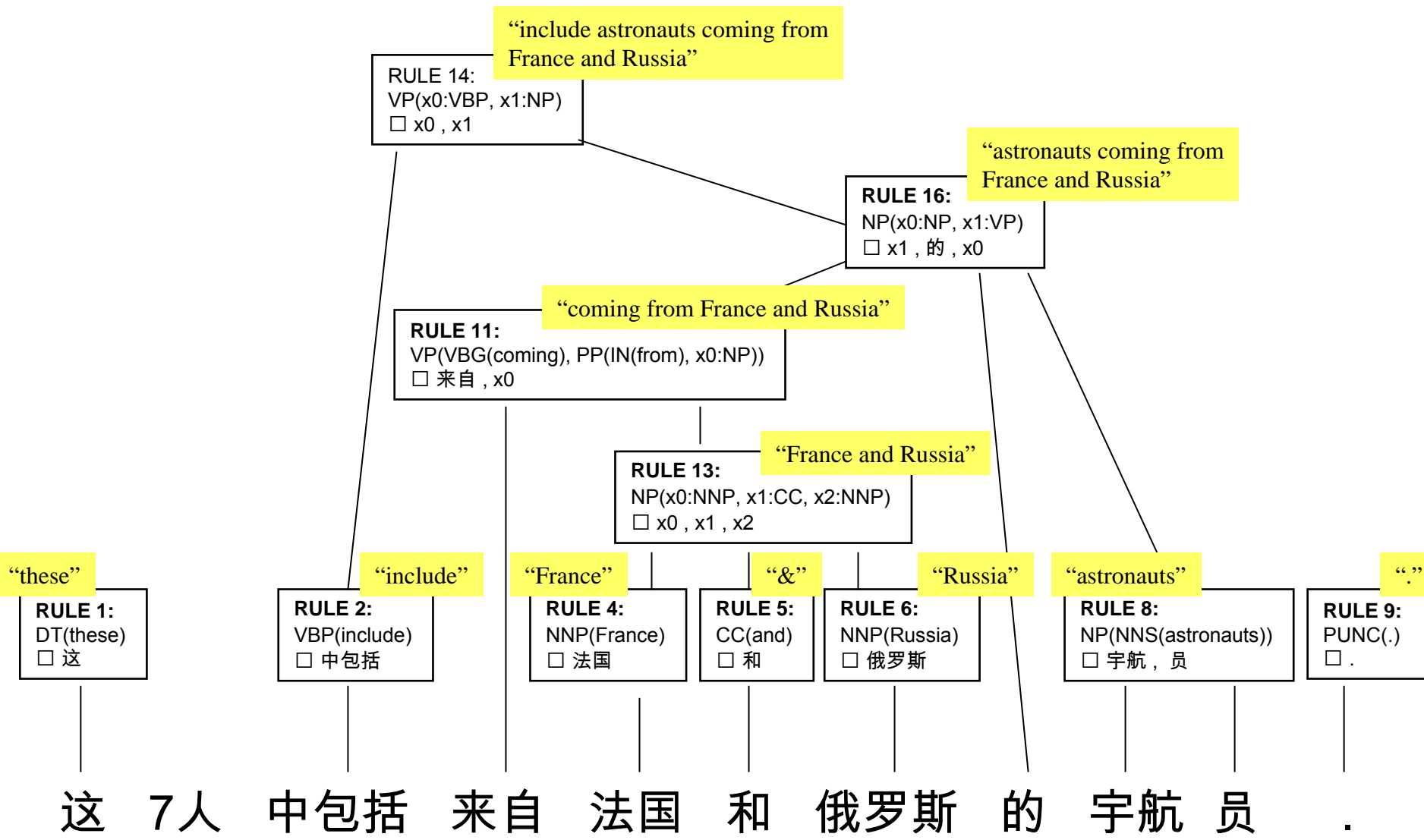Rules apply when their right-hand sides (RHS) match some portion of the input.

"these"
RULE 1:
DT(these)
□ 这

"include"
RULE 2:
VBP(include)
□ 中包括

"France"
RULE 4:
NNP(France)
□ 法国

"and"
RULE 5:
CC(and)
□ 和

"Russia"
RULE 6:
NNP(Russia)
□ 俄罗斯

"astronauts"
RULE 8:
NP(NNS(astronauts))
□ 宇航，员

"."
RULE 9:
PUNC(.)
□ .

这　7人　中包括　来自　法国　和　俄罗斯　的　宇航　员　　.

Syntax-Based Decoding

"France and Russia"

RULE 13:
NP(x0:NNP, x1:CC, x2:NNP)
□ x0 , x1 , x2

"these"

RULE 1:
DT(these)
□ 这

"include"

RULE 2:
VBP(include)
□ 中包括

"France"

RULE 4:
NNP(France)
□ 法国

"and"

RULE 5:
CC(and)
□ 和

"Russia"

RULE 6:
NNP(Russia)
□ 俄罗斯

"astronauts"

RULE 8:
NP(NNS(astronauts))
□ 宇航 , 员

"."

RULE 9:
PUNC(.)
□ .

这　7人　中包括　来自　法国　和　俄罗斯　的　宇航　员　　.

Syntax-Based Decoding

"coming from France and Russia"

RULE 11:
VP(VBG(coming), PP(IN(from), x0:NP))
□ 来自 , x0

"France and Russia"

RULE 13:
NP(x0:NNP, x1:CC, x2:NNP)
□ x0 , x1 , x2

"these"

RULE 1:
DT(these)
□ 这

"include"

RULE 2:
VBP(include)
□ 中包括

"France"

RULE 4:
NNP(France)
□ 法国

"&"

RULE 5:
CC(and)
□ 和

"Russia"

RULE 6:
NNP(Russia)
□ 俄罗斯

"astronauts"

RULE 8:
NP(NNS(astronauts))
□ 宇航 , 员

"."

RULE 9:
PUNC(.)
□ .

这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

Syntax-Based Decoding

"astronauts coming from France and Russia"

RULE 16:
NP(x0:NP, x1:VP)
□ x1 , 的 , x0

"coming from France and Russia"

RULE 11:
VP(VBG(coming), PP(IN(from), x0:NP))
□ 来自 , x0

"France and Russia"

RULE 13:
NP(x0:NNP, x1:CC, x2:NNP)
□ x0 , x1 , x2

"these"

RULE 1:
DT(these)
□ 这

"include"

RULE 2:
VBP(include)
□ 中包括

"France"

RULE 4:
NNP(France)
□ 法国

"&"

RULE 5:
CC(and)
□ 和

"Russia"

RULE 6:
NNP(Russia)
□ 俄罗斯

"astronauts"

RULE 8:
NP(NNS(astronauts))
□ 宇航 , 员

"."

RULE 9:
PUNC(.)
□ .

这　7人　中包括　来自　法国　和　俄罗斯　的　宇航　员　　.

RULE 14:
VP(x0:VBP, x1:NP)
□ x0 , x1

"include astronauts coming from France and Russia"

RULE 16:
NP(x0:NP, x1:VP)
□ x1 , 的 , x0

"astronauts coming from France and Russia"

RULE 11:
VP(VBG(coming), PP(IN(from), x0:NP))
□ 来自 , x0

"coming from France and Russia"

RULE 13:
NP(x0:NNP, x1:CC, x2:NNP)
□ x0 , x1 , x2

"France and Russia"

"these"

RULE 1:
DT(these)
□ 这

"include"

RULE 2:
VBP(include)
□ 中包括

"France"

RULE 4:
NNP(France)
□ 法国

"&"

RULE 5:
CC(and)
□ 和

"Russia"

RULE 6:
NNP(Russia)
□ 俄罗斯

"astronauts"

RULE 8:
NP(NNS(astronauts))
□ 宇航 , 员

"."

RULE 9:
PUNC(.)
□ .

这　7人　中包括　来自　法国　和　俄罗斯　的　宇航　员　　.

Derivation Tree

"These 7 people include astronauts coming from France and Russia"

RULE 15:
S(x0:NP, x1:VP, x2:PUNC)
→ x0 , x1 , x2

RULE 14:
VP(x0:VBP, x1:NP)
→ x0 , x1

"include astronauts coming from France and Russia"

"astronauts coming from France and Russia"

RULE 16:
NP(x0:NP, x1:VP)
→ x1 , 的 , x0

"coming from France and Russia"

RULE 11:
VP(VBG(coming), PP(IN(from), x0:NP))
→ 来自 , x0

"these 7 people"

RULE 10:
NP(x0:DT, CD(7), NNS(people)
→ x0 , 7人

"France and Russia"

RULE 13:
NP(x0:NNP, x1:CC, x2:NNP)
→ x0 , x1 , x2

"these"

RULE 1:
DT(these)
→ 这

"include"

RULE 2:
VBP(include)
→ 中包括

"France"

RULE 4:
NNP(France)
→ 法国

"&"

RULE 5:
CC(and)
→ 和

"Russia"

RULE 6:
NNP(Russia)
→ 俄罗斯

"astronauts"

RULE 8:
NP(NNS(astronauts))
→ 宇航 , 员

"."

RULE 9:
PUNC(.)
→ .

这　7人　中包括　来自　法国　和　俄罗斯　的　宇航　员　　.

Derived Tree

Derived Tree

```
                                    S
                         VP
                              NP
                                   VP
                                        PP
        NP                                    NP
                         NP                 NP        NP
   DT  CD   NNS   VBP        NNS   VBG  IN  NNP  CC  NNP  PUNC

 These  7  people include astronauts coming from France and Russia  .
```

Derivation Tree

S → NP VP PUNC

VP → VBP NP

NP → NP VP

VP → VBG PP

PP → IN NP

NP → NP CC NP

NP → DT CD NNS

NP → NNS

NP → NNP

NP → NNP

DT→These

NNS→people

CD→7

VBP→include

NNS→astronauts

VBG→coming

IN→from

CC→and

PUNC→.

NNP→France

NNP→Russia

Binarization for Decoding

- For CKY decoding, all rules must be *binarized.*
- Rule with |RHS| > 2 must be split into rules with |RHS| = 2
  - S(x0:NP VP(x1:VBD x2:NP)) → x1 x0 x2
    Z(x0:NP x1:VBD) → x1 x0
    S(x0:Z x1:NP) → x0 x1
- Similar to putting a CFG into Chomsky normal form.
- A rule can be binarized in different ways: must pick best!
- Some translation rules cannot be binarized at all…
  - A(x0:B x1:C x2:D x3:E) → x1 x3 x0 x2   [Wu 96]
- We just delete these.

- Binarization details: [Zhang, Huang, Knight, Gildea, 2006]

DerivTool 0.6

New Corpus   Corpus=f-arabic.plain (81 lines) 48   Go to line   Load   Save   Print   Print to File   Redraw   Close

Derivation Tree   Freeform Notes

S(NP-C(NPB(PRP("he"))) VP(VBD("said") x0:SBAR-C) x1:.)
-> "وقال" x0 x1
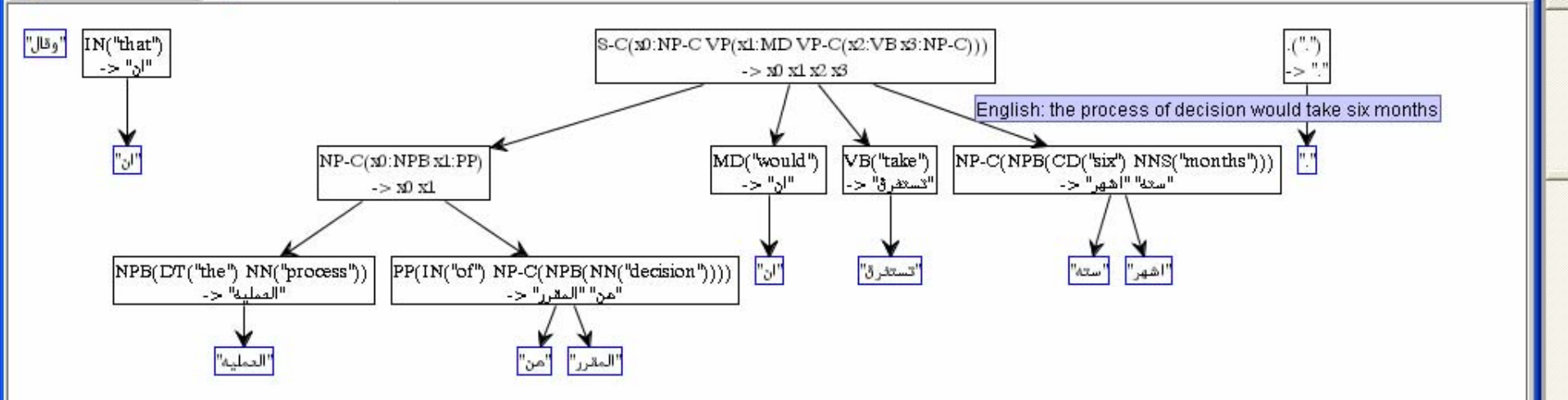
English: he said that the process of decision would take six months .

وقال

SBAR-C(x0:IN x1:S-C)
-> x0 x1

.(".")
-> "."

IN("that")
-> "ان"

S-C(x0:NP-C VP(x1:MD VP-C(x2:VB x3:NP-C)))
-> x0 x1 x2 x3

"."

ان

NP-C(x0:NPB x1:PP)   MD("would")   VB("take")   NP-C(NPB(CD("six") NNS("months")))

Click, Ctrl-click, or drag above to select. For adding rules, please select a contiguous span of top level nodes.   Modify   Delete

Phrase-based MT: he said that the process of decision to take six months .

Searching   Manual   Selecting Template

Lines = 100   Redraw   Click below to add rule.   (Red=no rules found, Blue=used by AT translation, Green=default translation, Purple=Red+Blue)

"وقال"	"ان"	"العملية"	"من"	"المقرر"	"ان"	"تستغرق"	"سعة"	"اشهر"
he said	that the process is			which would		take		four months
he		practical		would		takes		six weeks
that		the process		is to be		last		six months ' time
he		the operation		were to		took		six months ,
he said		operation		is		lasting	of six	@-@ mor
was		exercise	from	scheduled to		would take		month
had		operational		planned that		span		six @-@ months '
his		an operation		be		last for		6 month
while		the practical		the rapporteur		taken		of six months from

start   derivtool arabic   DerivTool 0.6   derivtool   7:53 PM

Why Might Syntax Help?

- Phrase-based MT output is "n-grammatical", not grammatical
  - Every sentence needs a subject and a verb

- Re-ordering is poorly explained as "distortion" -- better explained as syntactic transformation
  - Arabic to English, VSO → SVO

- Function words have syntactic effects even if they are not themselves translated

Why Might Syntax Hurt?

- Less freedom to glue pieces of output together -- search space has fewer output strings

- Search space is more difficult to navigate

- Rule extraction from bilingual text has limitations

available phrase-based translations
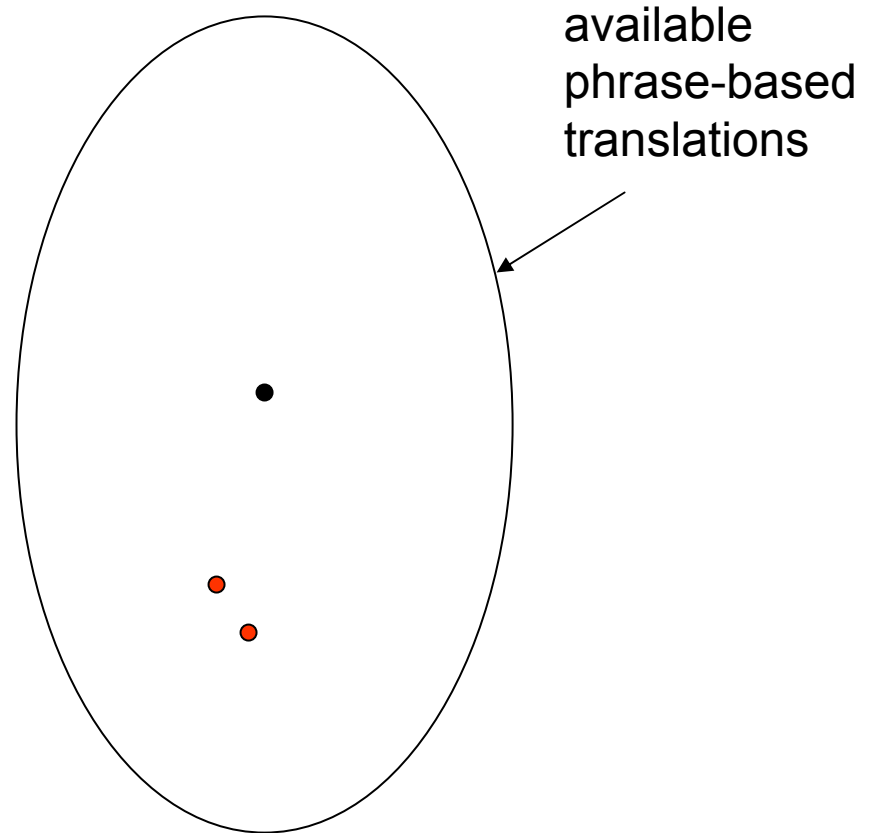
this section

Why Might Syntax Hurt?

- Less freedom to glue pieces of output together -- search space has fewer output strings

- Search space is more difficult to navigate

- Rule extraction from bilingual text has limitations

this section
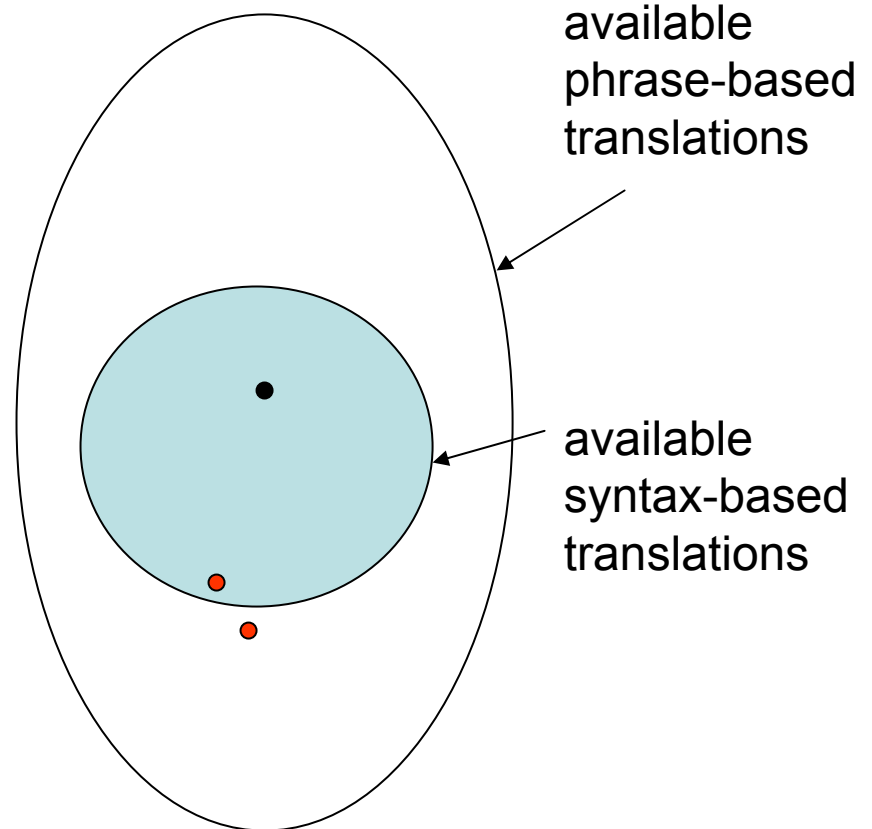
available phrase-based translations

available syntax-based translations

Why Might Syntax Hurt?

- Less freedom to glue pieces of output together -- search space has fewer output strings

- Search space is more difficult to navigate

- Rule extraction from bilingual text has limitations

available phrase-based translations

available syntax-based translations
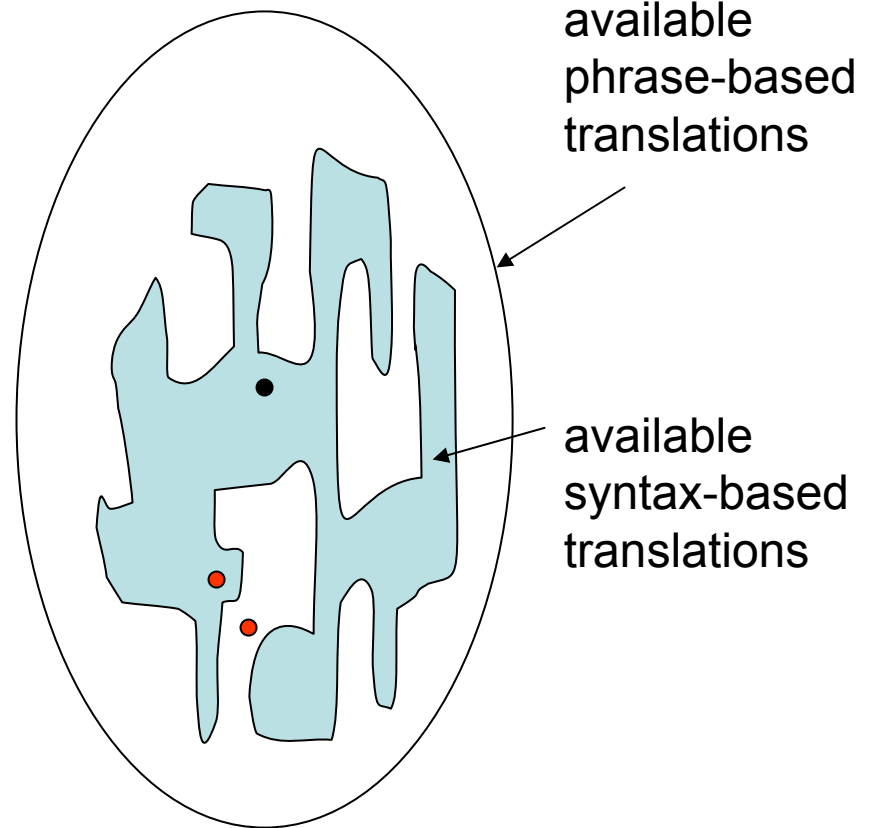
Comparing Phrase-Based Extraction with Syntax-Based Extraction

- Quantitatively compare
  - A typical phrase-based bilingual extraction algorithm (**ATS**, Och & Ney 2004)
  - A typical syntax-based bilingual extraction algorithm (**GHKM**, Galley et al 2004)
  - These algorithms picked from two good-scoring NIST-06 systems
- Identify areas of improvement for syntax-based rule coverage

Phrase-Based and Syntax-Based Pattern Extraction

estring
alignment
cstring

ATS [Och & Ney, 2004]

phrase pairs consistent with word alignment

etree
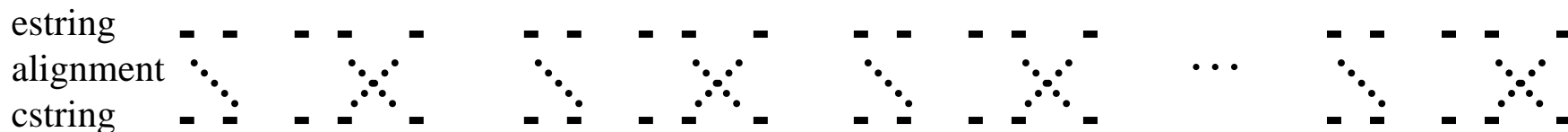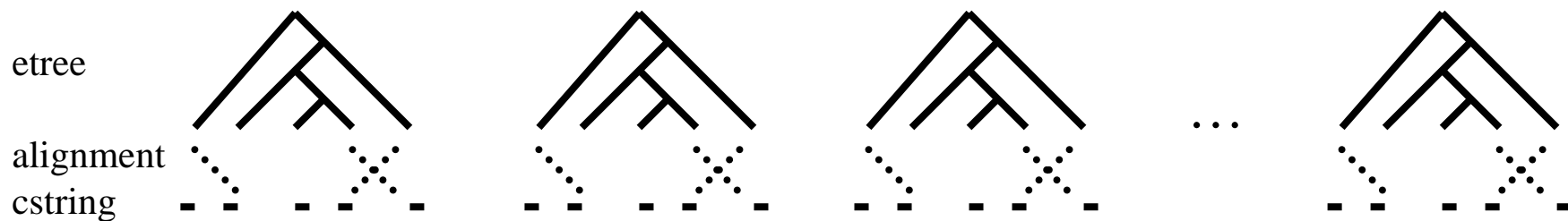
alignment
cstring

GHKM [Galley et al 2004]

syntax transformation rules consistent with word alignment

ATS (Och & Ney, 2004)

i felt obliged to do my part

我 有 责任 尽 一份 力

PHRASE PAIRS ACQUIRED:

felt	→ 有
felt obliged	→ 有 责任
felt obliged to do	→ 有 责任 尽
obliged	→ 责任
obliged to do	→ 责任 尽
do	→ 尽
part	→ 一份
part	→ 一份 力

ATS (Och & Ney, 2004)

i felt obliged to do my part

我 有 责任 尽 一份 力

PHRASE PAIRS ACQUIRED:

felt	→	有
felt obliged	→	有 责任
felt obliged to do	→	有 责任 尽
obliged	→	责任
obliged to do	→	责任 尽
do	→	尽
part	→	一份
part	→	一份 力

ATS (Och & Ney, 2004)

i felt obliged to do my part

我 有 责任 尽 一份 力

PHRASE PAIRS ACQUIRED:

felt → 有
felt obliged → 有 责任
felt obliged to do → 有 责任 尽
obliged → 责任
obliged to do → 责任 尽
do → 尽
part → 一份
part → 一份 力

ATS (Och & Ney, 2004)

i felt obliged to do my part

我 有 责任 尽 一份 力

PHRASE PAIRS ACQUIRED:

felt → 有
felt obliged → 有 责任
felt obliged to do → 有 责任 尽
obliged → 责任
obliged to do → 责任 尽
do → 尽
part → 一份
part → 一份 力

GHKM (Galley et al, 2004)

RULES ACQUIRED:

VBD(felt) → 有

VBN(obliged) → 责任

VP(x0:VBD
  VP-C(x1:VBN
    x2:SG-C) → x0 x1 x2

S(x0:NP-C x1:VP) → x0 x1

S
NP-C   VP
    VP-C
VBD   SG-C
    VP
VBN
  TO   VP-C
    VB   NP-C
        NPB
NPB
PRP   PRP$ NN

i felt obliged to do my part

我 有 责任 尽 一份 力

GHKM (Galley et al, 2004)

RULES ACQUIRED:

VBD(felt) → 有

VBN(obliged) → 责任

VP(x0:VBD
   VP-C(x1:VBN
      x2:SG-C) → x0 x1 x2

S(x0:NP-C x1:VP) → x0 x1

Tree content:

S
NP-C   VP
   VP-C
VBD   SG-C
   VP
VBN
  TO   VP-C
    VB   NP-C
NPB   NPB
PRP   PRP$ NN

i felt obliged to do my part

我 有 责任 尽 一 份 力

GHKM (Galley et al, 2004)

RULES ACQUIRED:

VBD(felt) → 有

VBN(obliged) → 责任

VP(x0:VBD
    VP-C(x1:VBN
        x2:SG-C) → x0 x1 x2

S(x0:NP-C x1:VP) → x0 x1

GHKM (Galley et al, 2004)

RULES ACQUIRED:

VBD(felt) → 有

VBN(obliged) → 责任

VP(x0:VBD
    VP-C(x1:VBN
        x2:SG-C) → x0 x1 x2

S(x0:NP-C x1:VP) → x0 x1

GHKM (Galley et al, 2004)

RULES ACQUIRED:

VBD(felt) → 有

VBN(obliged) → 责任

VP(x0:VBD
   VP-C(x1:VBN
       x2:SG-C) → x0 x1 x2

S(x0:NP-C x1:VP) → x0 x1

There is a unique tiling that identifies minimal translation units.

S
NP-C VP
VBD VP-C
VBN SG-C
VP
TO VP-C
VB NP-C
NPB NPB
PRP PRP$ NN

i felt obliged to do my part

我 有 责任 尽 一份 力

GHKM Syntax Rules

Phrasal Translation

VP → **está**, **cantando**

VBZ — VBG

is — **singing**

Non-constituent Phrases

S → **hay**, NP

PRO — VP

there — VB — NP

are

Non-contiguous Phrases

VP → **poner**, NP

VB — NP — PRT

put — **on**

Context-Sensitive Word Insertion

NPB → NNS

DT — NNS

the

Multilevel Re-Ordering

S → VB, NP1, NP2

NP1 — VP

VB — NP2

Lexicalized Re-Ordering

NP → NP1, 的 , NP2

NP2 — PP

P — NP1

of

GHKM Syntax Rules

Phrasal Translation

VP → **está**, **cantando**

VBZ VBG
| |
is **singing**

Non-constituent Phrases

S → **hay**, NP

PRO   VP
|
there VB   NP
|
are

Non-contiguous Phrases

VP → **poner**, NP

VB   NP   PRT
| |
put   **on**

Context-Sensitive Word Insertion

NPB → NNS

DT   NNS
|
the

Multilevel Re-Ordering

S → VB, NP1, NP2

NP1   VP

VB   NP2

Lexicalized Re-Ordering

NP → NP1, 的 , NP2

NP2   PP

P   NP1
|
of
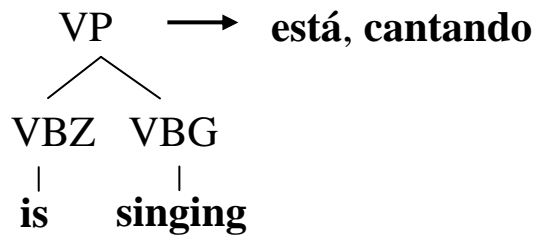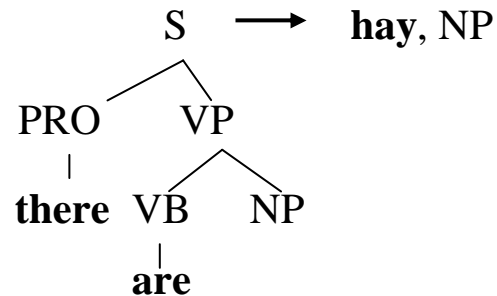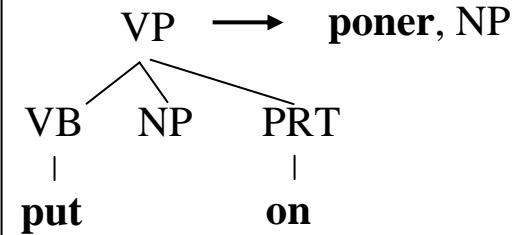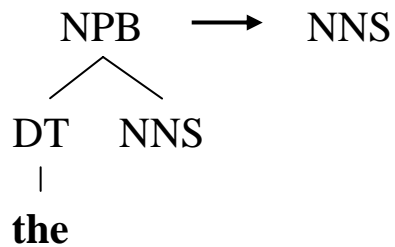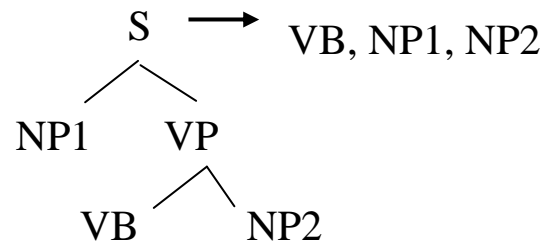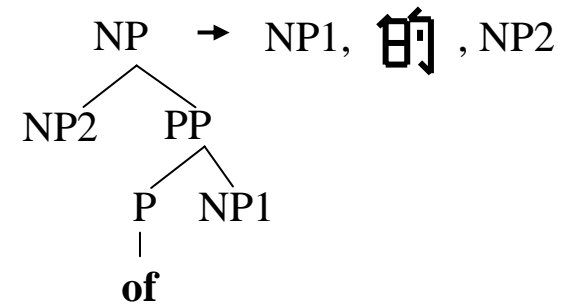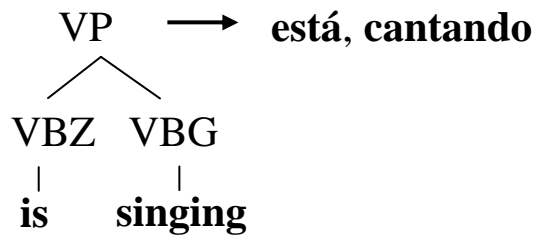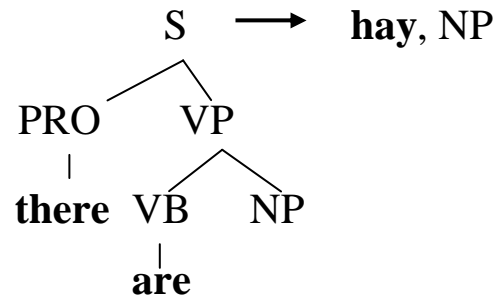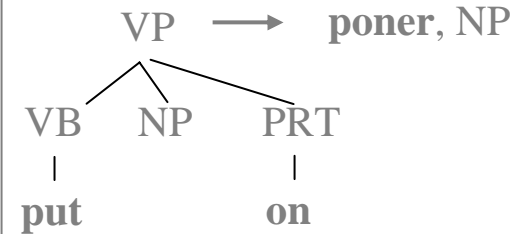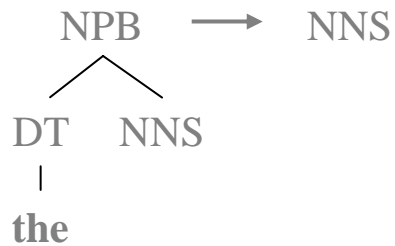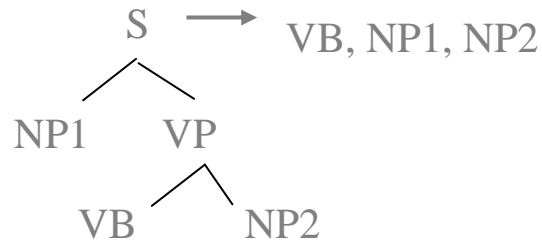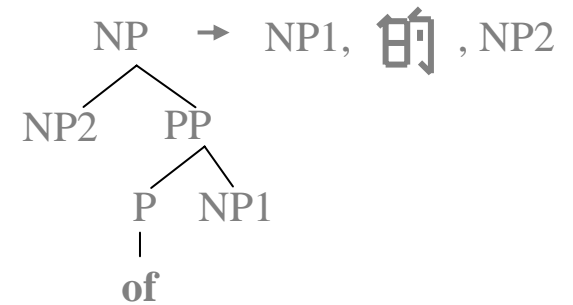
ATS and GHKM Methods Do Not Coincide

GHKM Phrase Pairs Relevant to NIST-02

ATS Phrase Pairs Relevant to NIST-02

43k

134k

161k

GHKM has no built-in phrase size limit -- ATS does.

GHKM misses phrases due to parse failures.

GHKM pulls unaligned English words into phrases.

GHKM forced to incorporate some unaligned foreign words into phrases.

GHKM only gets *minimal* rules to explain each segment pair.

GHKM forced to incorporate unaligned English words into phrases.

GHKM phrases come with applicability conditions.

ATS and GHKM Methods Overlap

GHKM Phrase Pairs Relevant to NIST-02

ATS Phrase Pairs <u>actually used</u> in 1-best decodings of NIST-02 (1,994 = 2 per sentence).

CAN WE REDUCE THIS NUMBER?

1,994

GHKM misses phrases due to parse failures.

GHKM phrases come with applicability conditions.

GHKM only gets *minimal* rules to explain each segment pair.

GHKM forced to incorporate unaligned English words into phrases.

GHKM forced to incorporate some unaligned foreign words into phrases.

Some Methods for Improving Syntax-Based Rule Extraction

- Acquire larger rules
  - Composed rules (Galley et al, 06)
  - Phrasal rules (Marcu et al, 06)
- Acquire more general rules
  - Re-structure English trees (Wang et al, 07)
  - Re-align tree/string pairs (May & Knight, 07)
- Expand syntactic category set
  - Slash categories (Zollmann & Venugopal 06)

Larger, Composed Rules

Minimal GHKM Rules:

B(e1 e2) → c1 c2
C(e3) → c3
A(x0:B x1:C) → x0 x1

Additional Composed Rules:

A(B(e1 e2) x0:C) -> c1 c2 x0
A(x0:B C(e3)) -> x0 c3
A(B(e1 e2) C(e3)) -> c1 c2 c3

"big phrasal rule"

Larger, Composed Rules

Minimal GHKM Rules:

B(e1 e2) → c1 c2
C(e3) → c3
A(x0:B x1:C) → x0 x1

Additional Composed Rules:

A(B(e1 e2) x0:C) -> c1 c2 x0
A(x0:B C(e3)) -> x0 c3
A(B(e1 e2) C(e3)) -> c1 c2 c3

"big phrasal rule"

Larger, Composed Rules

A

B

C

e1

e2

e3

c1   c2

c3

Minimal GHKM Rules:

B(e1 e2) → c1 c2
C(e3) → c3
A(x0:B x1:C) → x0 x1

Additional Composed Rules:

A(B(e1 e2) x0:C) -> c1 c2 x0
A(x0:B C(e3)) -> x0 c3
A(B(e1 e2) C(e3)) -> c1 c2 c3

"big phrasal rule"

Larger, Composed Rules

A

B

C

e1

e2

e3

c1

c2

c3

Minimal GHKM Rules:

B(e1 e2) → c1 c2
C(e3) → c3
A(x0:B x1:C) → x0 x1

Additional Composed Rules:

A(B(e1 e2) x0:C) -> c1 c2 x0
A(x0:B C(e3)) -> x0 c3
A(B(e1 e2) C(e3)) -> c1 c2 c3

"big phrasal rule"

GHKM (Galley et al, 2006)
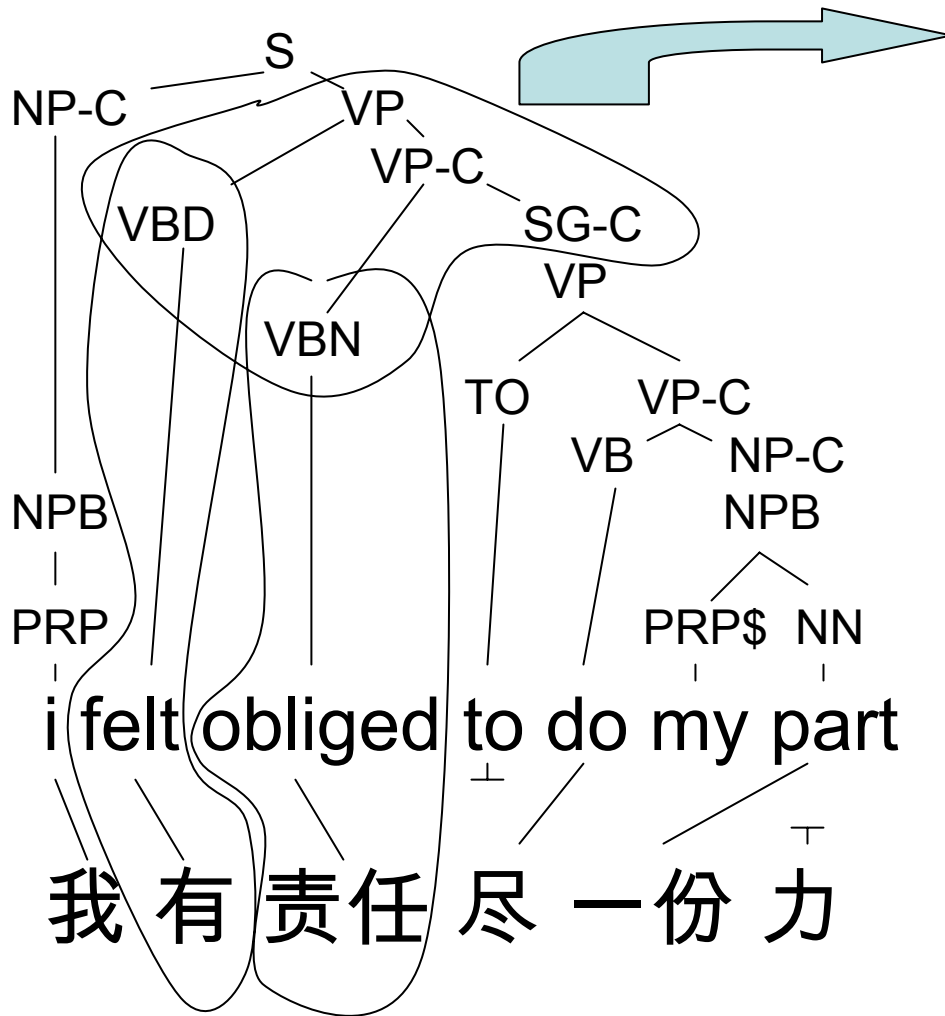
S

NP-C   VP

VBD   VP-C

SG-C

VBN   VP

TO   VP-C

VB   NP-C

NPB   NPB

PRP   PRP$ NN

i felt obliged to do my part

我 有 责任 尽 一份 力

RULES ACQUIRED:

VBD(felt)   → 有

VBN(obliged)   → 责任

VP(x0:VBD
    VP-C(x1:VBN
        x2:SG-C) → x0 x1 x2

VP(VBD(felt)
    VP-C(VBN(obliged))
        x0:SG-C) → 有 责任 x0

S(x0:NP-C x1:VP)   → x0 x1

GHKM (Galley et al, 2006)

S

NP-C

VP

VP-C

VBD

SG-C

VBN

VP

TO

VP-C

NPB

VB

NP-C

NPB

PRP

PRP$ NN

i felt obliged to do my part

⊥

⊤

我 有 责任 尽 一份 力

RULES ACQUIRED:

VBD(felt)   → 有

VBN(obliged)   → 责任

VP(x0:VBD
   VP-C(x1:VBN
        x2:SG-C) → x0 x1 x2

VP(VBD(felt)
   VP-C(VBN(obliged))
        x0:SG-C) → 有 责任 x0

S(x0:NP-C x1:VP)   → x0 x1

GHKM (Galley et al, 2006)

RULES ACQUIRED:

VBD(felt) → 有
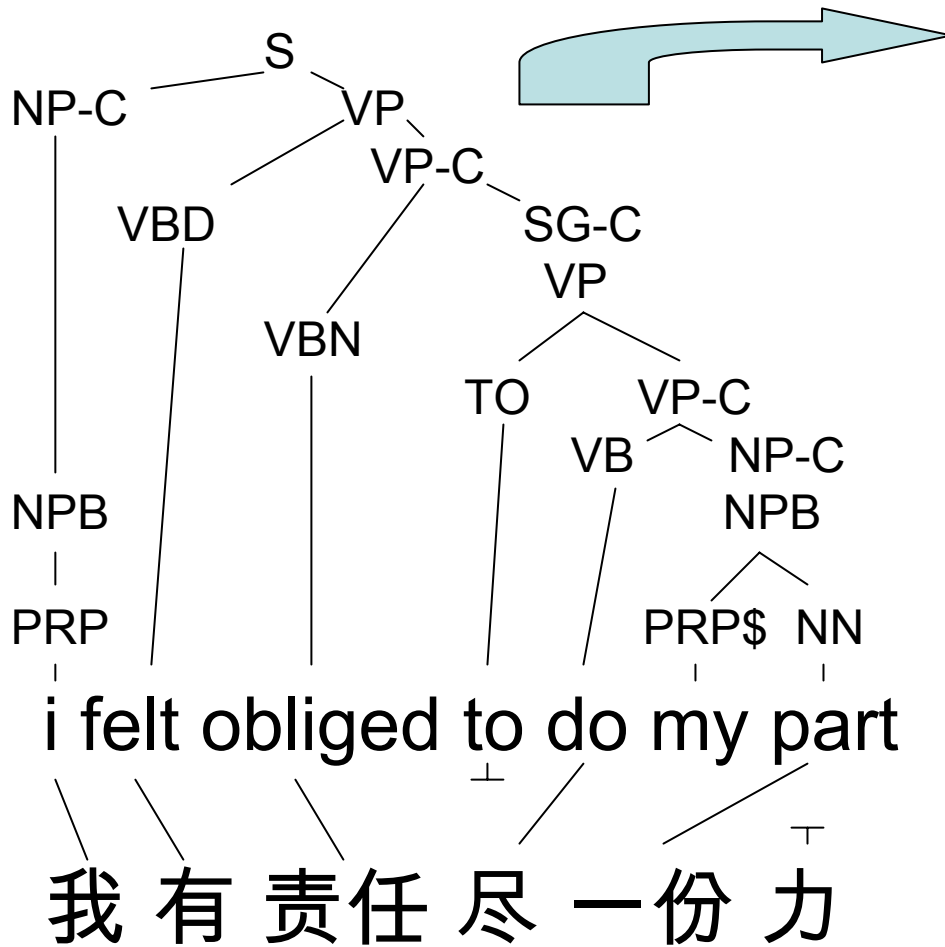
VBN(obliged) → 责任

VP(x0:VBD
    VP-C(x1:VBN
        x2:SG-C) → x0 x1 x2

**VP(VBD(felt)
    VP-C(VBN(obliged))
        x0:SG-C) → 有 责任 x0**

S(x0:NP-C x1:VP) → x0 x1

minimal rules tile the tree/string/alignment triple.
composed rules are made by combining those tiles.

Larger, Composed Rules

Composed limit (internal nodes in composed rule)	# of rules acquired	Unacquired phrase pairs used in ATS 1-best decodings
0 = minimal	2.5m	1994
2	12.4m	1478
3	26.9m	1096
4	**55.8m**	**900**

"Phrasal" Syntax Rules

- SPMT Model 1 (Marcu et al 2006)
  - consider each foreign phrase up to length L
  - extract smallest possible syntax rule that does not violate alignments

Method	Unacquired ATS Phrase Pairs
Minimal	1994
Composed 4	900
SPMT M1	**676**
Both	**663**

Restructuring English Training Trees
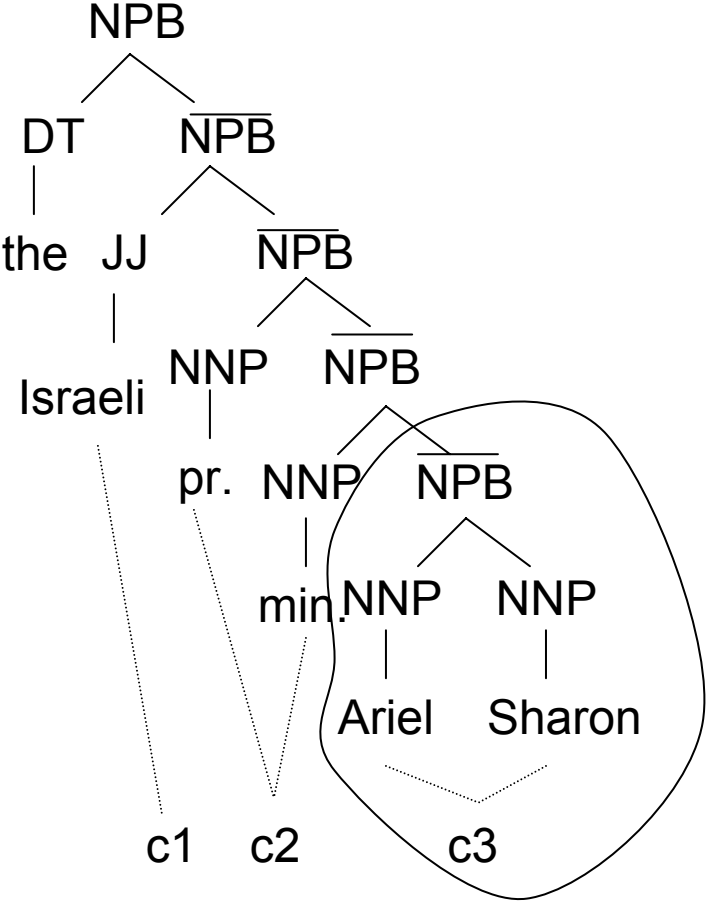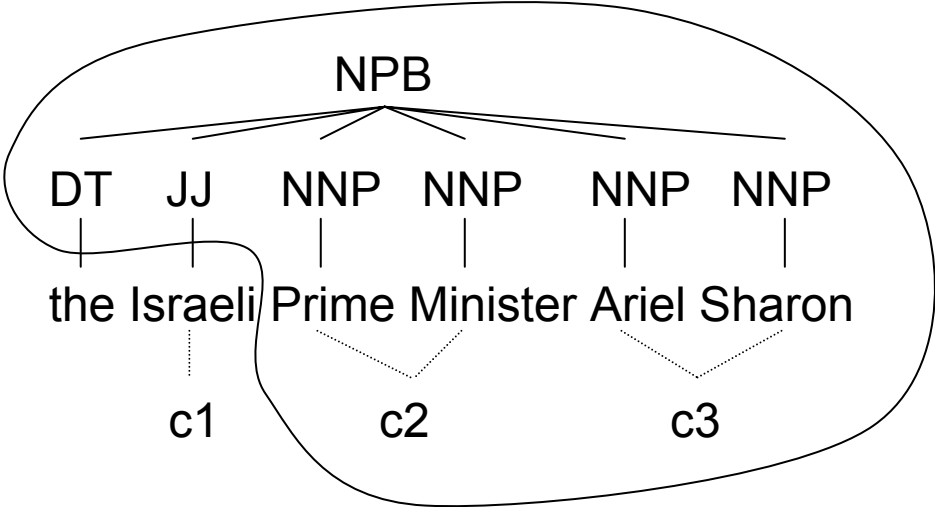
Restructuring English Training Trees

Method	Unacquired ATS Phrase Pairs
Minimal	1994
+ Composed 4	900
+ SPMT M1	663
+ Restructuring	**458**

Effects of Coverage Improvements on Syntax-Based MT Accuracy

	Chinese/English Trained on 9.8m words		Arabic/English Trained on 4.1m words	
	Dev-02	Test-03	Dev-02	Test-03
ATS	36.00	34.31	50.88	51.04
GHKM minimal	39.11	38.85	49.81	50.46
GHKM composed 2	41.59	40.90	51.18	51.52
GHKM composed 3	42.28	41.62	51.96	52.04
GHKM composed 4	42.63	41.82	52.05	52.26
GHKM minimal + SPMT	41.01	40.34	50.74	51.81
GHKM composed 4 + SPMT	43.30	42.17	52.15	52.12
+ Left binarization of etrees	43.45	42.41	52.86	52.42

NIST Bleu r4n4

Improved English Binarization
Why are Penn Treebank Trees Problematic for Translation?
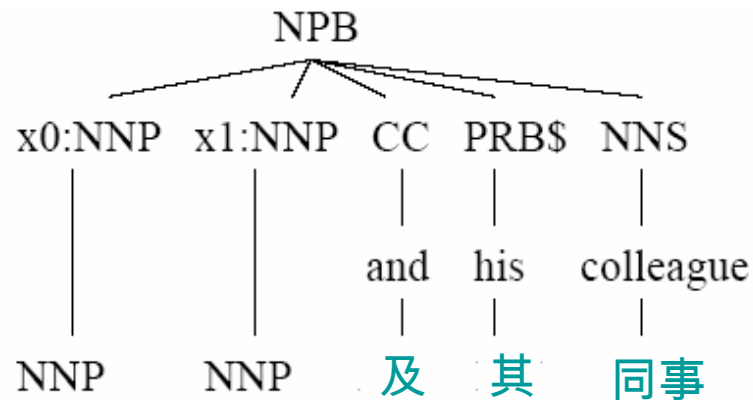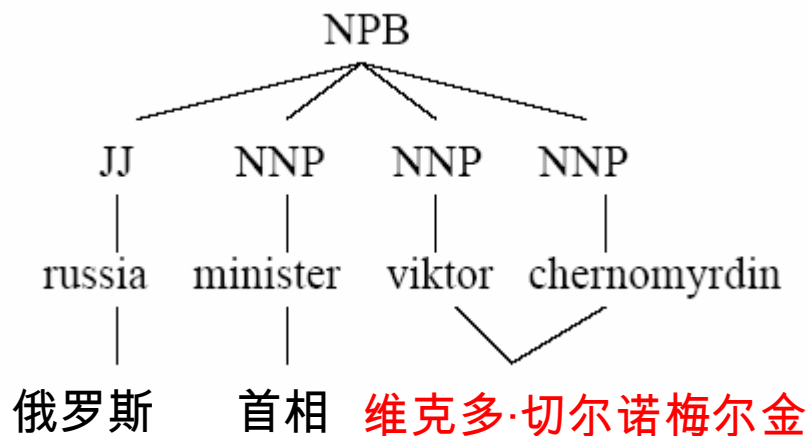
Improved English Binarization
Why are Penn Treebank Trees Problematic for Translation?

R1
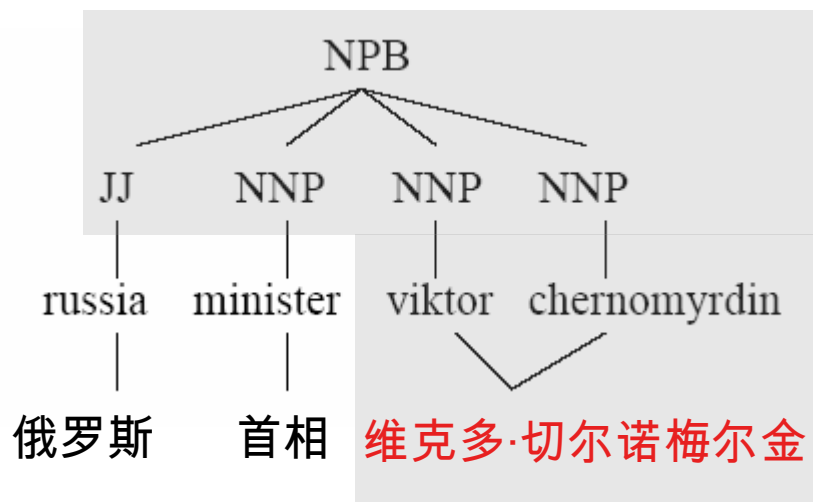
R2

?
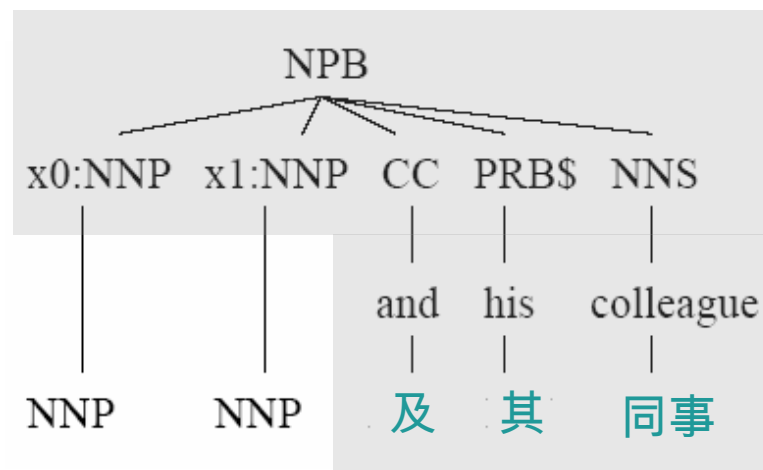
维克多·切尔诺梅尔金   及   其   同事

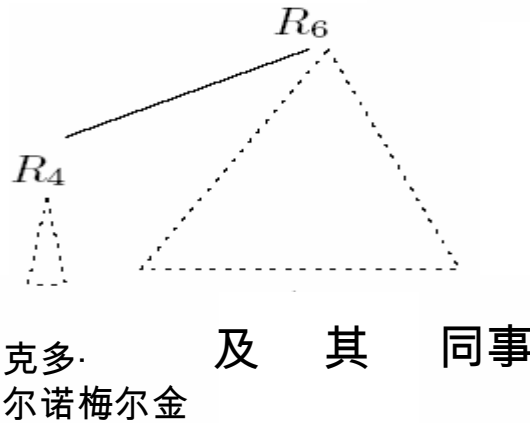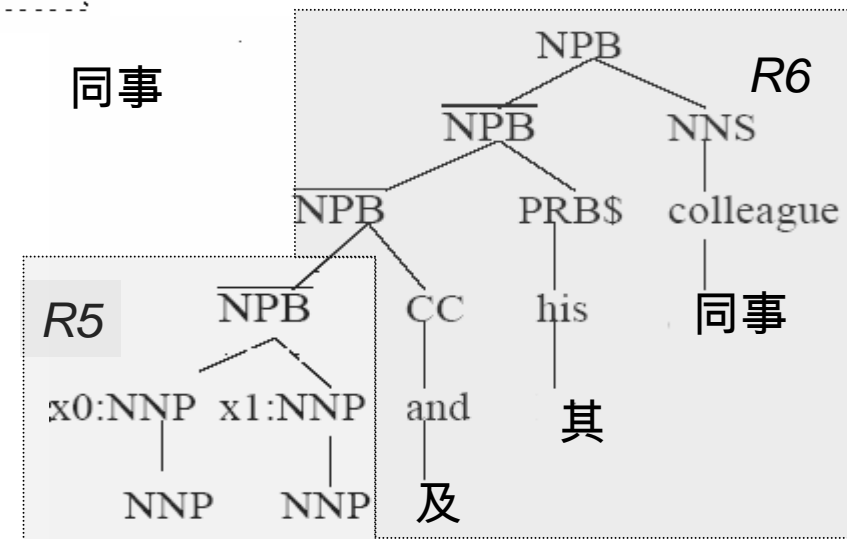# Binarizing English Trees
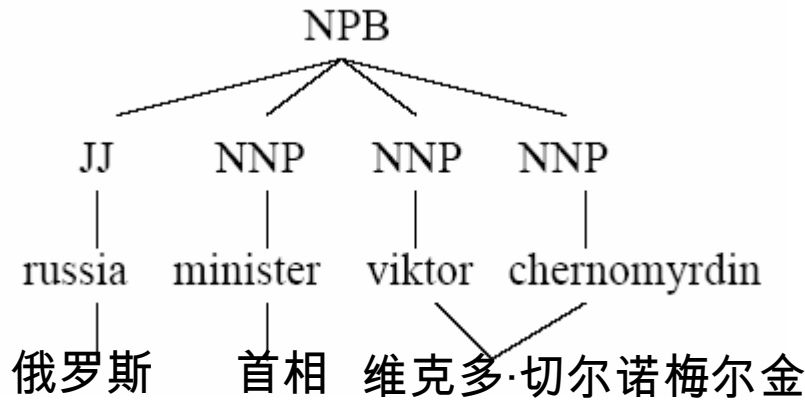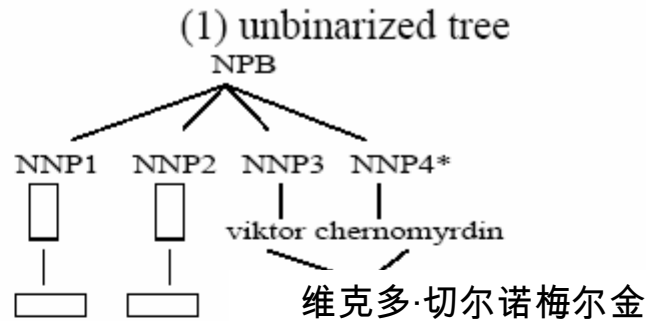

Right binarize

Left binarize

R3
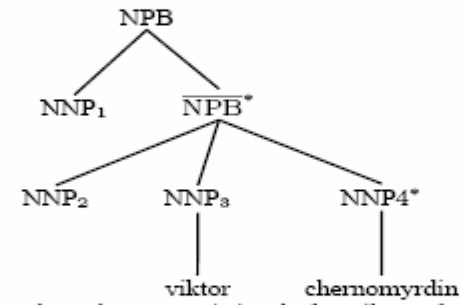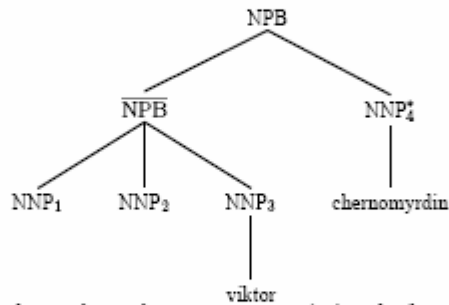
R4

R5

R6
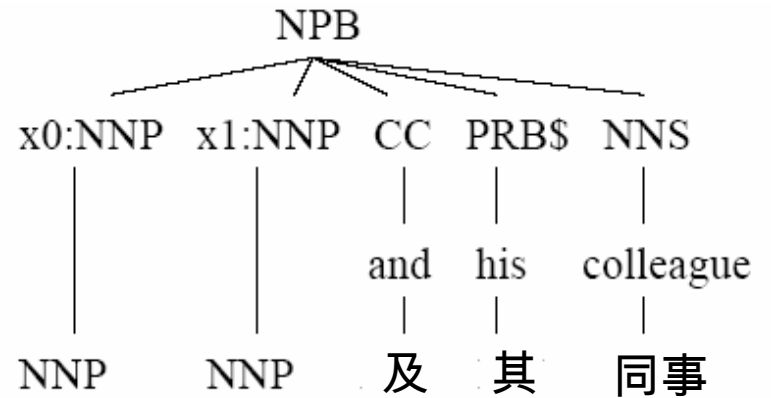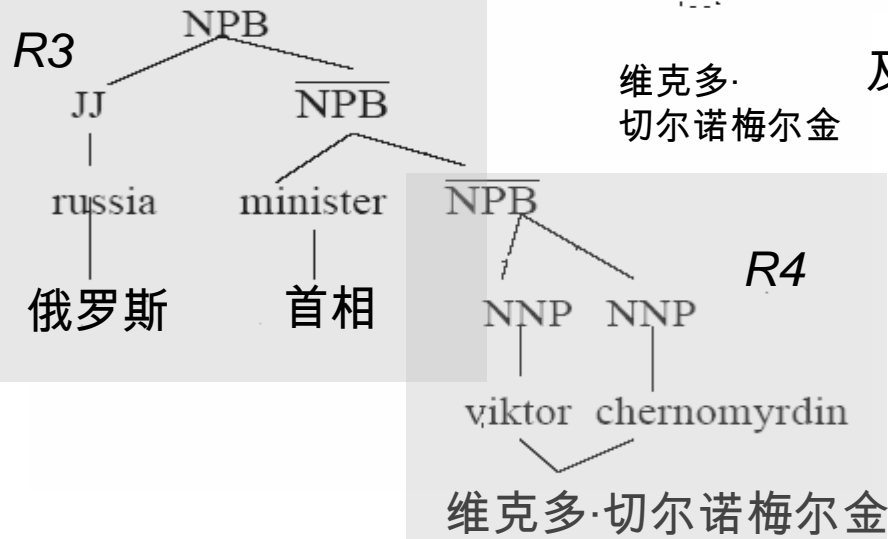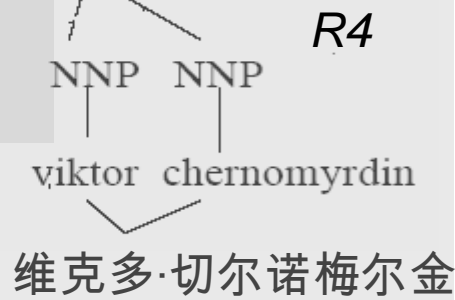
Simple Binarizations

(1) unbinarized tree

维克多·切尔诺梅尔金
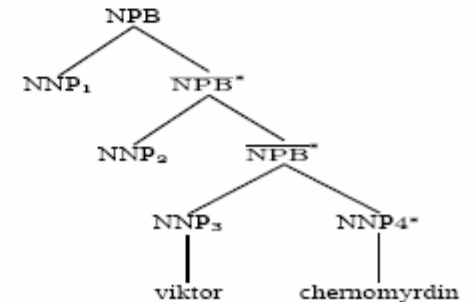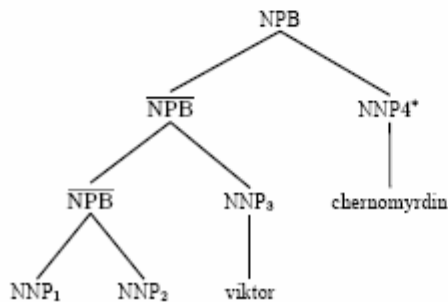
(2) left-binarization

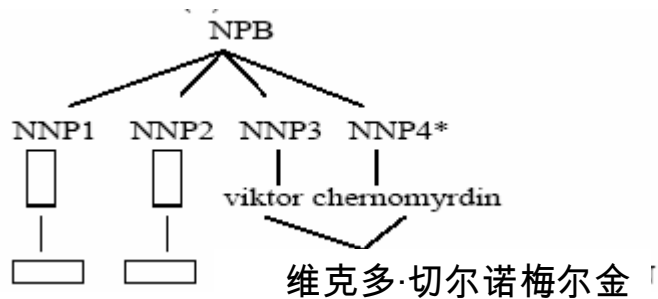(3) right-/head-binarization

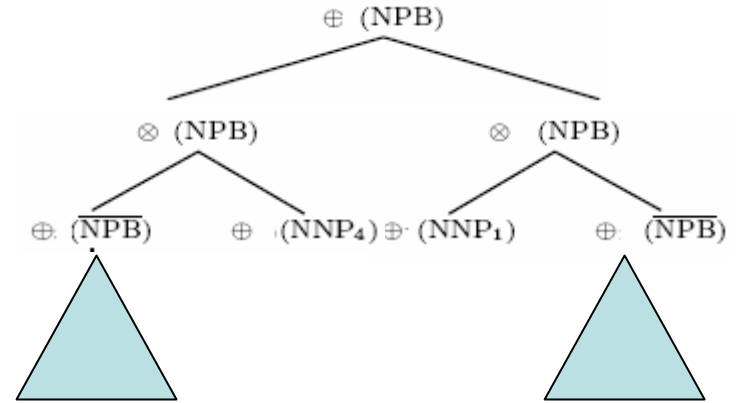(4) left-binarization   (5) right-binarization
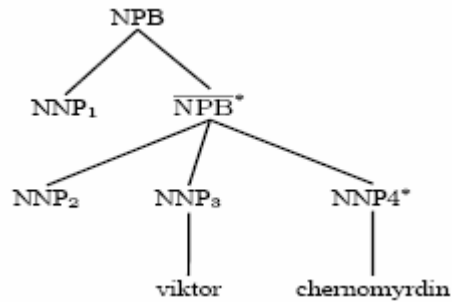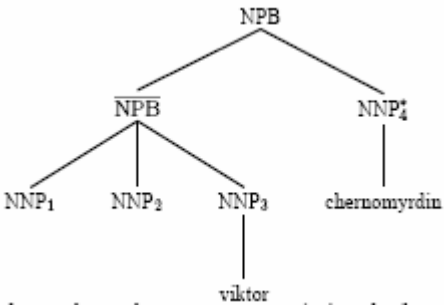
(6) left-binarization   (7) right-/head-binarization

Parallel Binarization

Parallel Binarization
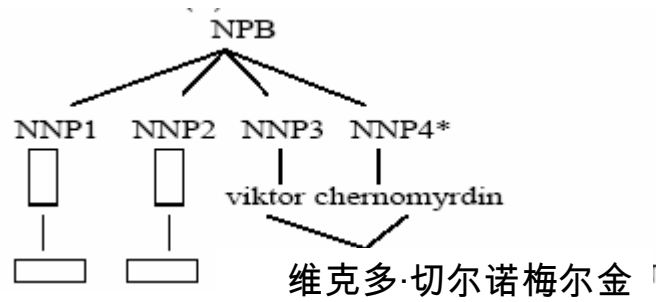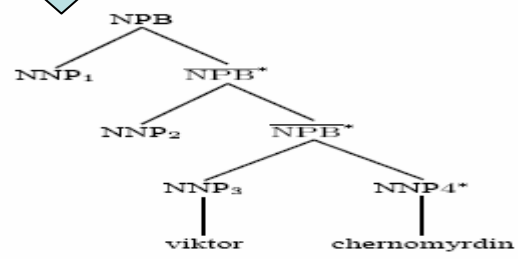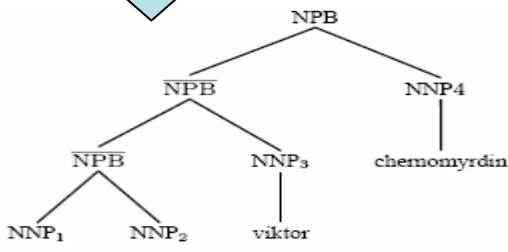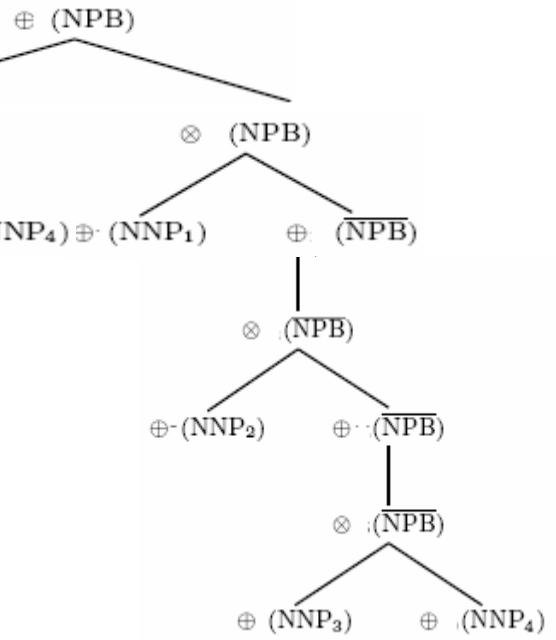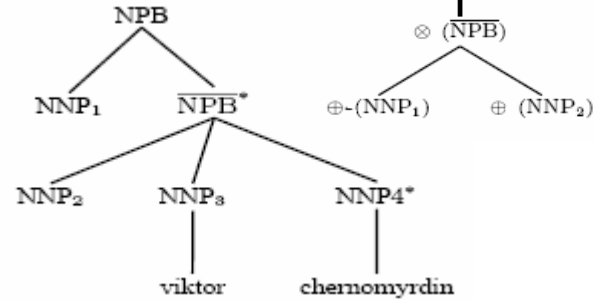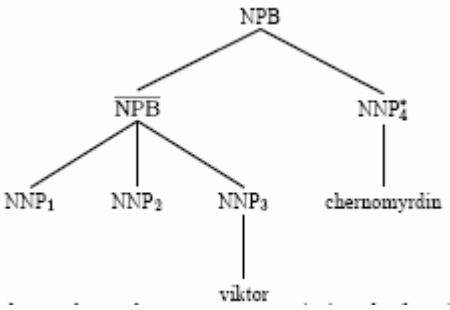
Forest-Based Rule Extraction

- Gets **all** minimal rules consistent with word alignment and **some** binarization

- Run EM algorithm to determine best binarization of each node in each tree

Binarization Using EM
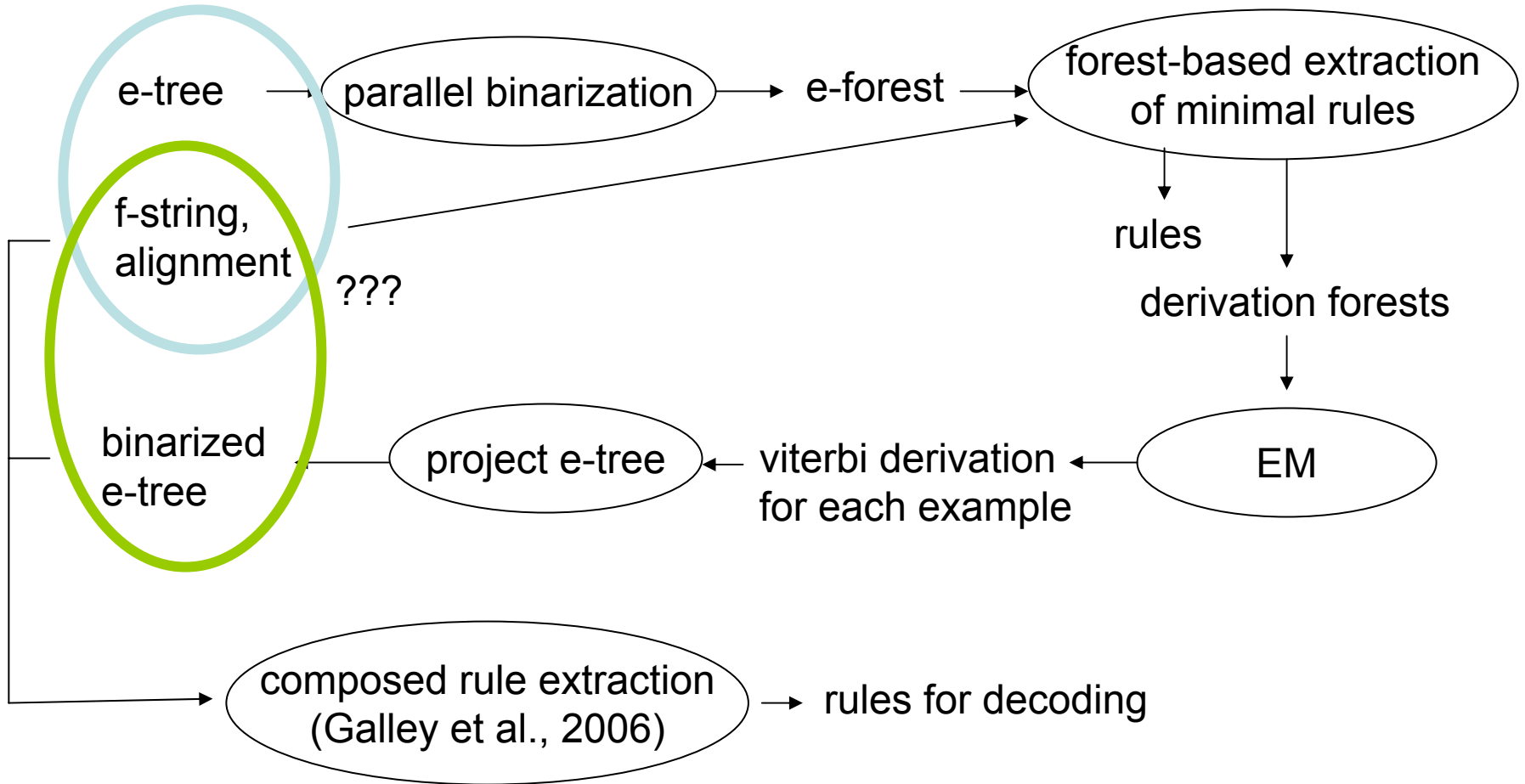
Experimental Results
(Wang, Knight, Marcu 2007)

Type of Binarization	# of Rules Learned	Test Bleu (NIST-03)
None	63.4m	36.94
Left	114.0m	37.47 (p=0.047)
Right	113.0m	37.49 (p=0.044)
Head	113.8m	37.54 (p=0.086)
EM	115.6m	**37.94** (p=0.0047)

Tree binarized by EM training
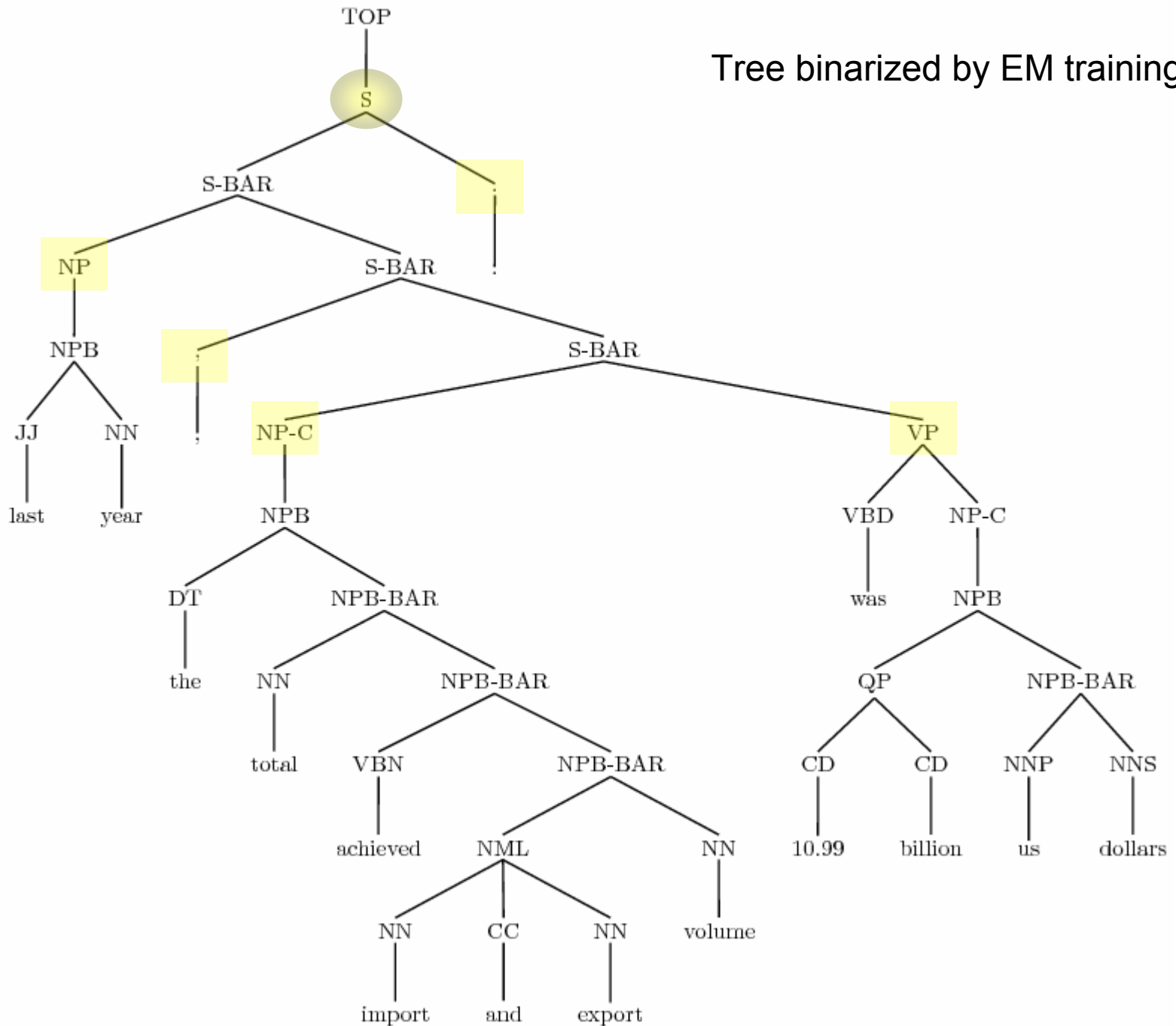
Syntax-Based Word Alignment

- GIZA++ string-based alignments
  - are errorful
  - don't match our syntax-based MT system

- We would like to use our tree-based translation model to align data

Syntax-Based Word Alignment

English trees
Foreign strings

throw away GIZA alignments

EM alignment ("Training Tree Transducers", Graehl & Knight'04)

GIZA++ → initial word alignments → GHKM syntax rule extraction → minimal rules

Viterbi derivations → Improved word alignments

composed rule extraction

better rules for decoding

May & Knight, 07

Result: +0.5-1.0 Bleu
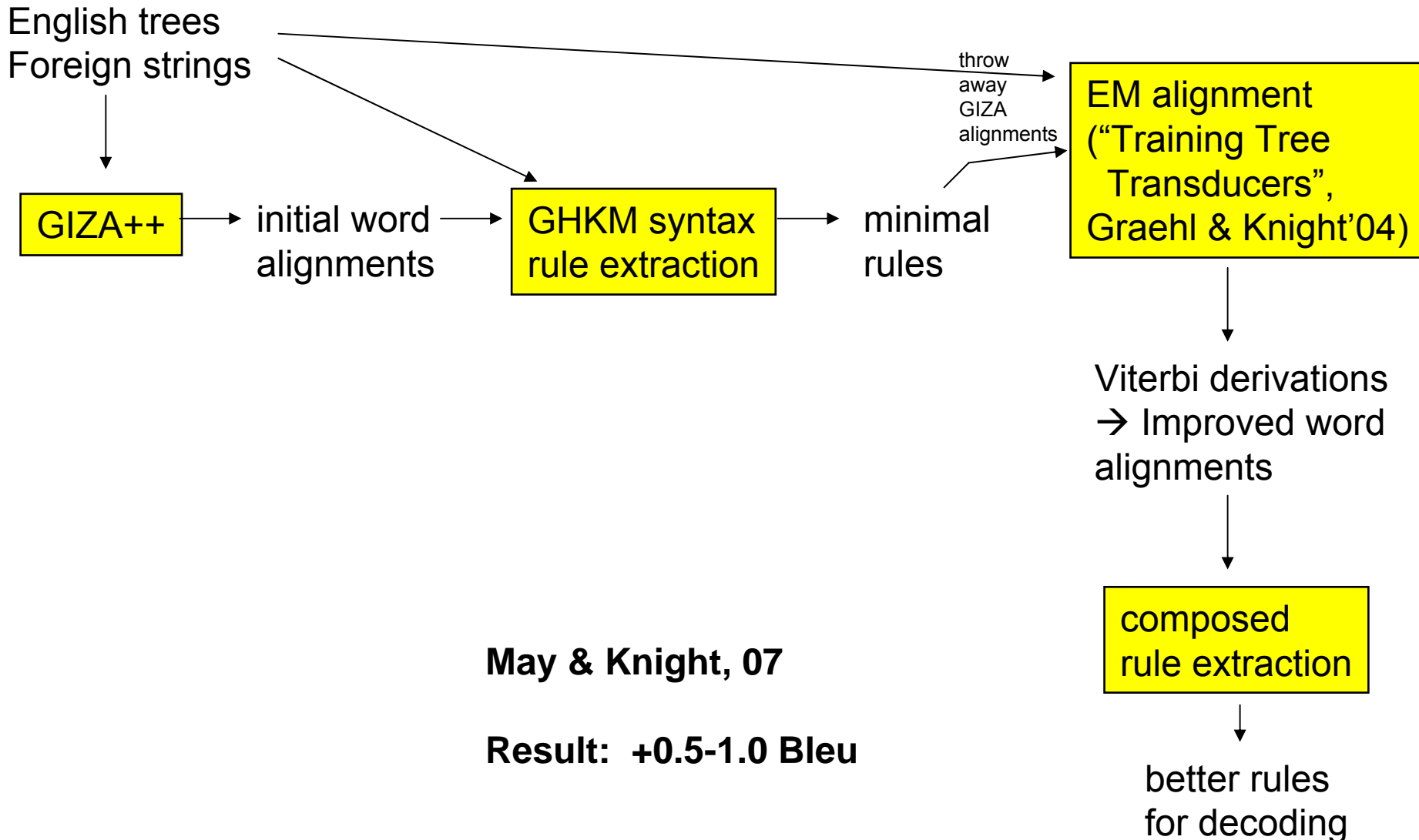
Remarks

- Phrase-based and syntax-based extraction algorithms have different coverage.

- Syntax-based coverage can be improved:
  - composed rules
  - phrasal rules
  - binarizing English trees with EM
  - re-aligning tree/string pairs with EM

- Improvements lead to better translation accuracy.

Some Sample Outputs

input

dev-little (line 47) - dev-little

Input: 中 资 已 成为 澳门 最 大的 外来 投资者 。
Reference: the chinese enterprises have become the biggest outside investors in macao .
AlTemp-e: investment $_0$ | in $_1$ | macao $_2$ | has $_3$ | become $_4$ | the largest $_5$ | foreign $_6$ | investors $_7$ | . $_8$
AlTemp-f: 中 $_1$ | 资 $_0$ | 已 $_3$ | 成为 $_4$ | 澳门 $_2$ | 最 大的 $_5$ | 外来 $_6$ | 投资者 $_7$ | 。 $_8$
[dev-little] 1-Best: the chinese - funded enterprises have become the largest external investor in macao .

[dev-little] 1-Best Tree
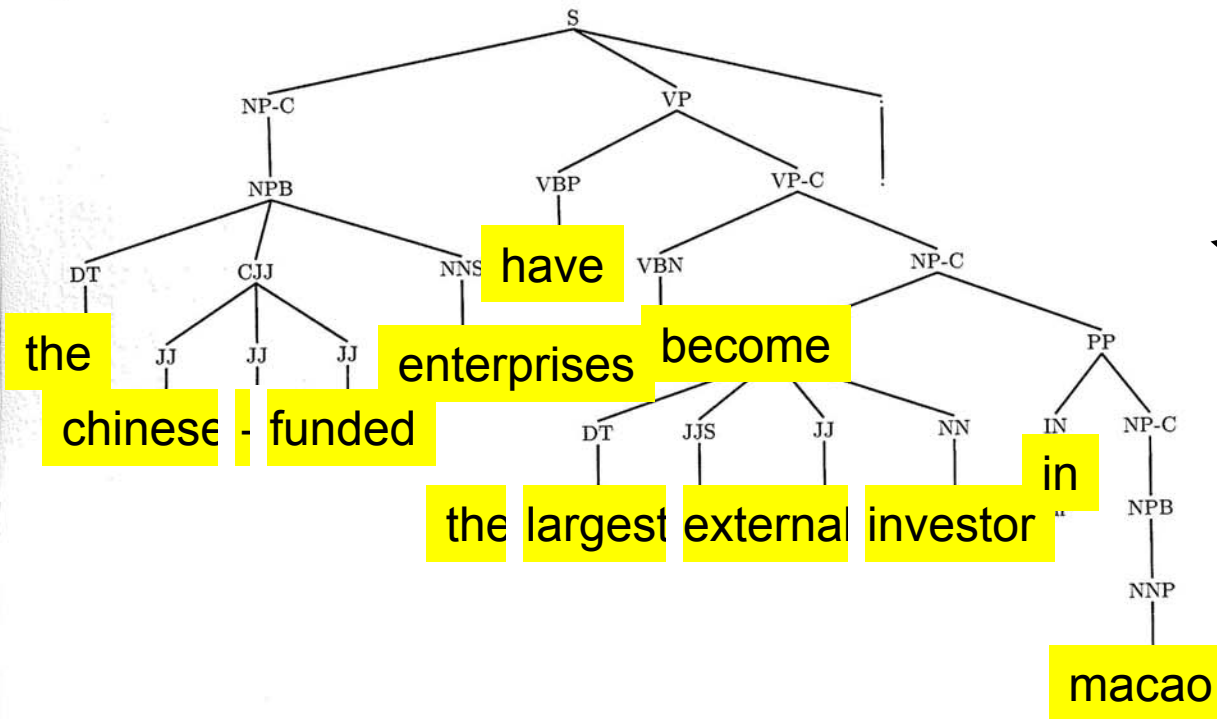
phrase-based
system output

syntax-based
system output

input

dev-little (line 59) - dev-little

Input: 基 纳 纳 对 中国 过去 向 坦 桑 提供 的 大量 援助 表示 感谢 。

Reference: keenana expressed gratitude to china for its great assistance to tanzania in the past .

AlTemp-e: kinana $_0$ | to $_1$ | china $_2$ | in the past to $_3$ | tanzania $_4$ | expressed appreciation $_5$ | for the substantial $_6$ | assistance $_7$ | . $_8$

AlTemp-f: 基 纳 纳 $_0$ | 对 $_1$ | 中国 $_2$ | 过去 向 $_3$ | 坦 桑 $_4$ | 提供 的 大量 $_6$ | 援助 $_7$ | 表示 感谢 $_5$ | 。 $_8$

[dev-little] 1-Best: kinana thanked china for substantial assistance to tanzania in the past .

phrase-based system output

[dev-little] 1-Best Tree
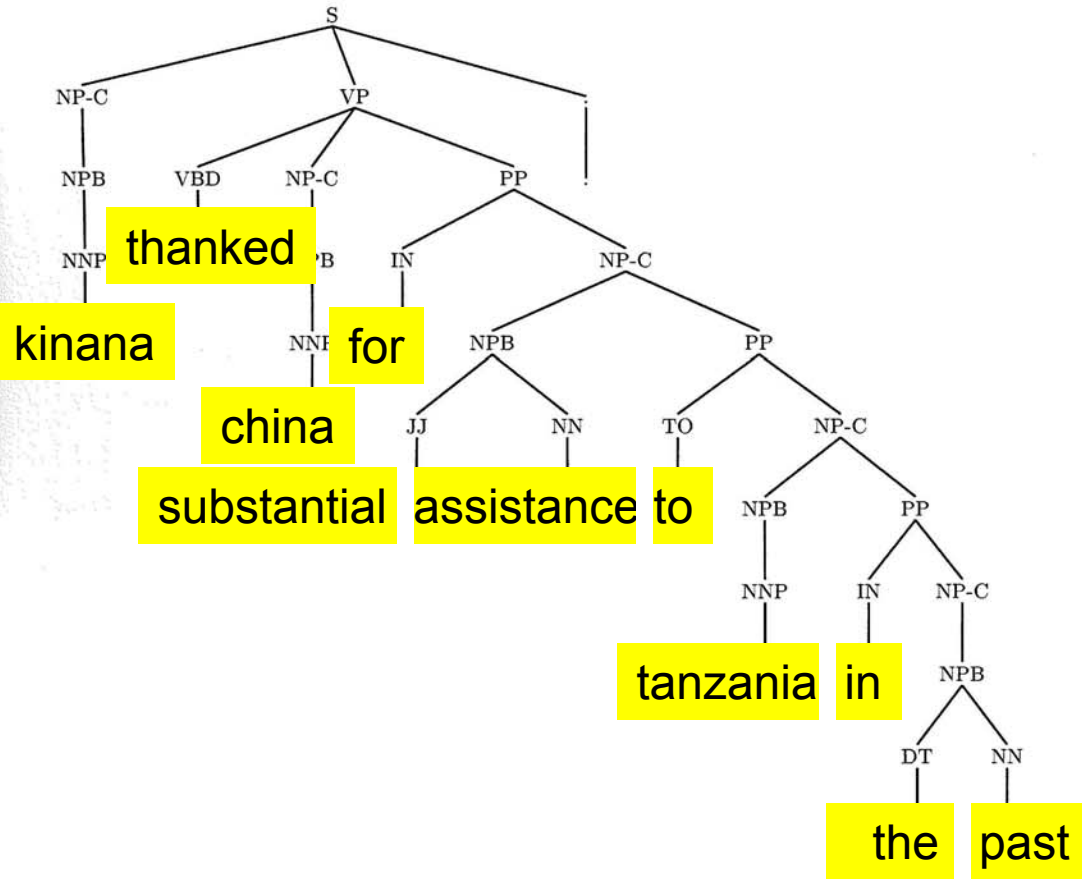
syntax-based system output

dev-little (line 38) - dev-little

Input: 此次 为期 两天 的 研讨会 ， 由 世界贸易组织 上海 研究中心 与 上海市 对外 服务 有限公司 联合 举办 。

Reference: the two - day seminar is jointly sponsored by the wto shanghai research center and shanghai foreign service company limited .

AlTemp-e: the $_0$ | two - day $_1$ | seminar $_2$ | by the world trade organization $_3$ | , $_4$ | shanghai research center $_5$ | and $_6$ | shanghai foreign service $_7$ | co . , ltd . $_8$ | jointly $_9$ | . $_{10}$

AlTemp-f: 此次 $_0$ | 为期 两天 的 $_1$ | 研讨会 $_2$ | , $_4$ | 由 世界贸易组织 $_3$ | 上海 研究中心 $_5$ | 与 $_6$ | 上海市 对外 服务 $_7$ | 有限公司 $_8$ | 联合 举办 $_9$ | 。 $_{10}$

[dev-little] 1-Best: the two - day seminar , organized jointly by shanghai foreign services ltd . and shanghai research center of the world trade organization .

[dev-little] 1-Best Tree

Input: 丁 豪 在 儿童 福利 院 读 完 小学 ， 随后 进入 附近 乡里 一所 学校 上 初中 。

Reference: ding hao completed his primary school at the children welfare school , and then went to a nearby township middle school .

AlTemp-e: ding hao $_0$ | in $_1$ | children 's welfare institute $_2$ | complete primary school $_3$ | , $_4$ | then entered $_5$ | commune $_6$ | near $_7$ | a school $_8$ | , $_9$ | junior high school level . $_{10}$

AlTemp-f: 丁 豪 $_0$ | 在 $_1$ | 儿童 福利 院 $_2$ | 读 完 小学 $_3$ | ， $_4$ | 随后 进入 $_5$ | 附近 $_7$ | 乡里 $_6$ | 一所 学校 $_8$ | 上 $_9$ | 初中 。 $_{10}$

dev-little] 1-Best: ding hao in children 's homes finishing elementary school and was subsequently entered junior high school at a school near their home towns .

lev-little] 1-Best Tree

dev-little (line 64) - dev-little

con became very good partner

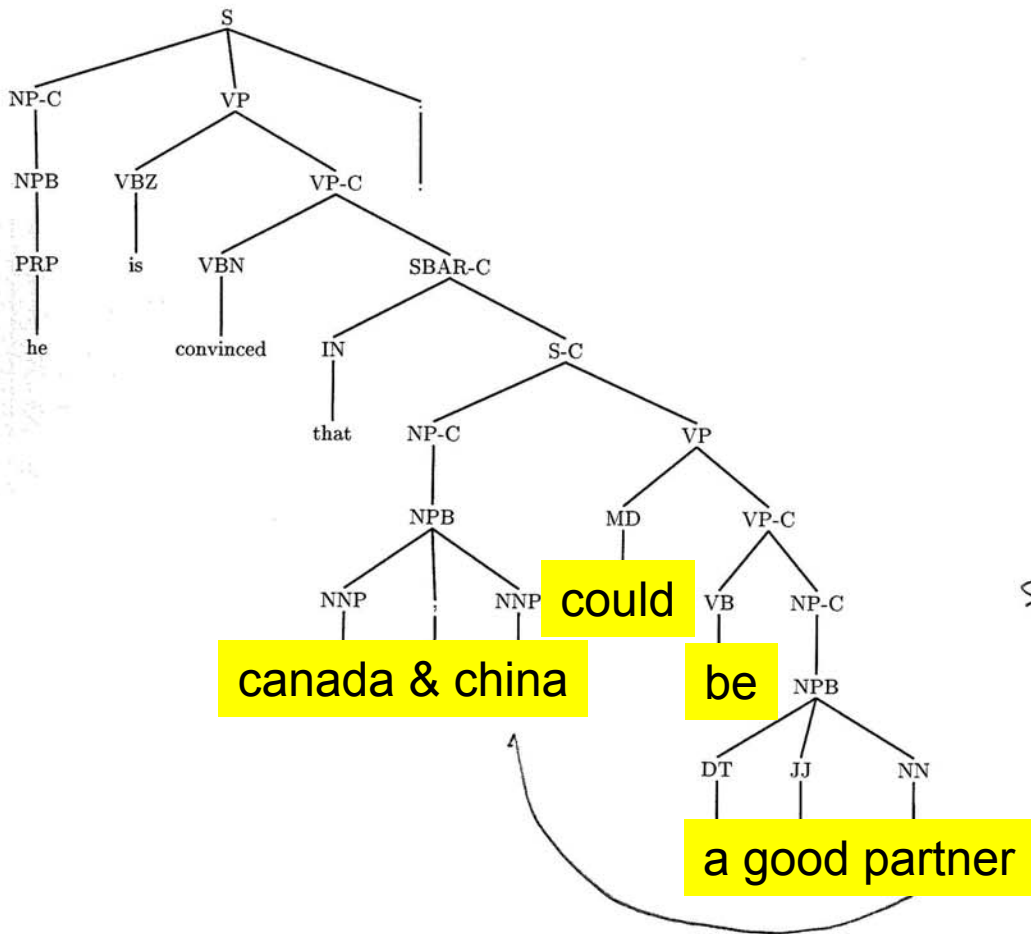|	
Input:	他 确信 ， 加 、 中 两国 可以 成为 很 好的 合作 伙伴 。
Reference:	he assured that canada and china can become very good partners .
AlTemp-e:	he was convinced that $_0$	the $_1$	two countries $_2$	, $_3$	can $_4$	become good $_5$	partners $_6$	. $_7$
AlTemp-f:	他 确信 ， $_0$	加 $_1$	、 中 $_3$	两国 $_2$	可以 $_4$	成为 很 好的 $_5$	合作 伙伴 $_6$	。 $_7$
[dev-little] 1-Best:	he is convinced that canada , china could be a good partner .

[dev-little] 1-Best Tree

Subj - obj number agreement.

dev-little (line 51) - dev-little

**Input:** 法国 外长 昨天 是 在 法国 国民议会 外事 委员会 会议 上 发表 上述 声明 的 。

Reference: the french foreign minister made the above statement in a meeting of the foreign affairs commission of the french national congress .

AlTemp-e: french $_0$ | foreign minister $_1$ | in the french national assembly $_2$ | yesterday , $_3$ | the statement delivered by $_4$ | foreign affairs $_5$ | committee meeting $_6$ | . $_7$

AlTemp-f: 法国 $_0$ | 外长 $_1$ | 昨天 是 $_3$ | 在 法国 国民议会 $_2$ | 外事 $_5$ | 委员会 会议 上 $_6$ | 发表 上述 声明 $_4$ | 的 。 $_7$

[dev-little] 1-Best: the french foreign minister was the statement issued at the french national assembly foreign affairs committee meeting yesterday .

[dev-little] 1-Best Tree
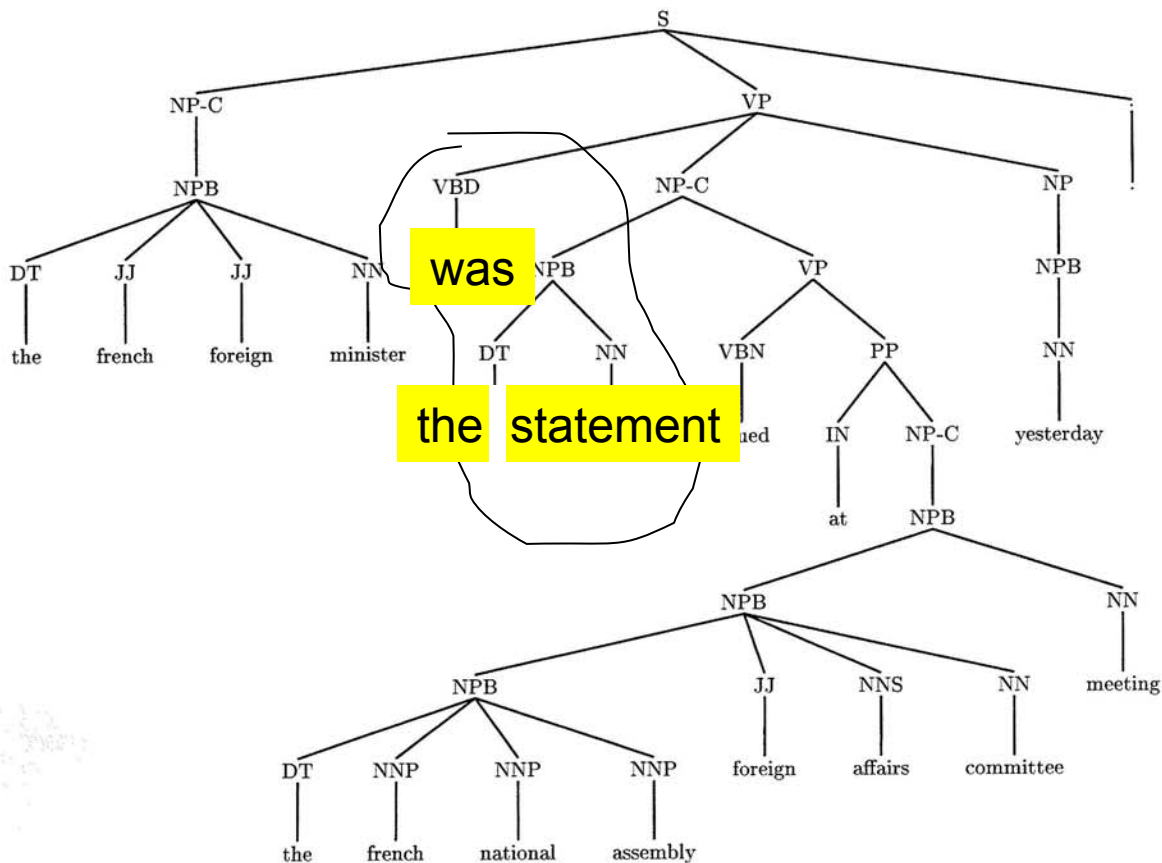
dev-little (line 125) - dev-little

rain

Input: 今年 在 加利福尼亚 州 和 南部 地区 (的) 豪 雨 都 归咎 于 厄尔尼诺 作 宠 。
Reference: the torrential rain this year in california and its southern part is attributed to the el nino .
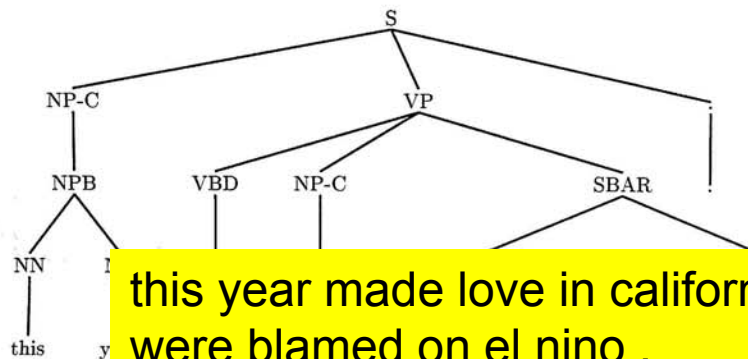AlTemp-e: this year $_0$ | in california $_1$ | and southern $_2$ | areas $_3$ | of heavy rains $_4$ | attributed to $_5$ | a $_6$ | favorite $_7$ | el nio $_8$ | . $_9$
AlTemp-f: 今年 $_0$ | 在 加利福尼亚 州 $_1$ | 和 南部 $_2$ | 地区 $_3$ | 的 豪 雨 $_4$ | 都 归咎 于 $_5$ | 厄尔尼诺 $_8$ | 作 $_6$ | 宠 $_7$ | 。 $_9$
[dev-little] 1-Best: this year made love in california and southern areas of torrential rains were blamed on el nino .
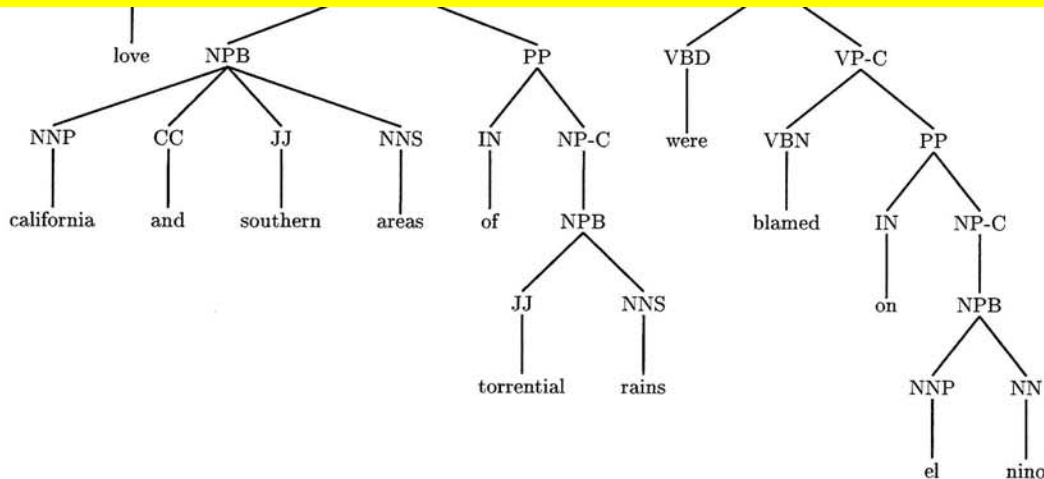
[dev-little] 1-Best Tree

Scope.

funny.

this year really NP-C

this year made love in california and southern areas of torrential rains were blamed on el nino .
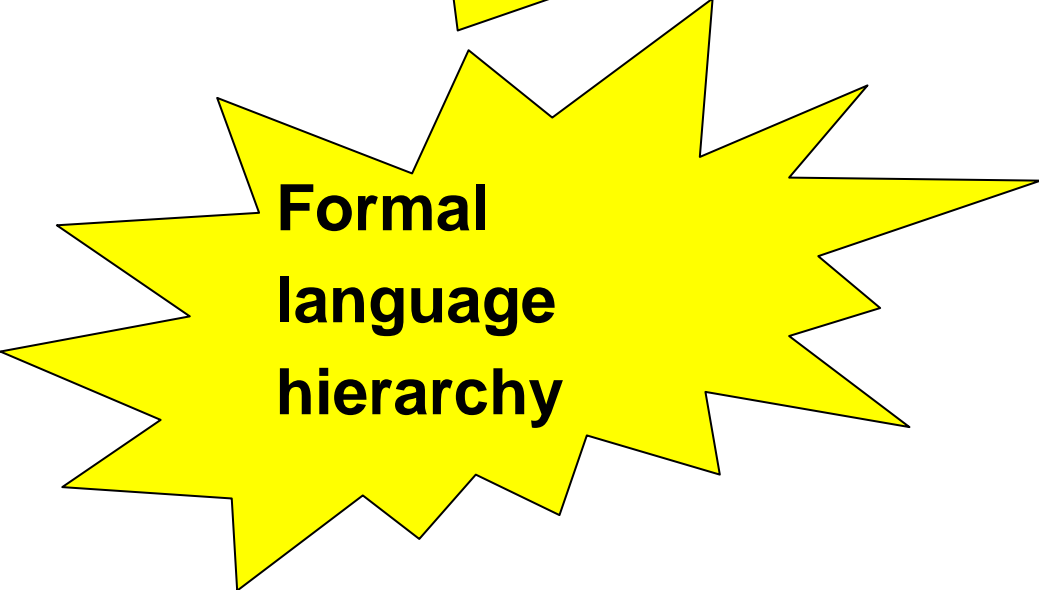
Lots of Open Problems

Chomsky's Program [1957]

- **Algorithmically distinguish between grammatical and ungrammatical sentences:**
  - John thinks Sara hit the boy
  - * The hit thinks Sara John boy
  - John thinks the boy was hit by Sara
  - Who does John think Sara hit?
  - John thinks Sara hit the boy and the girl
  - * Who does John think Sara hit the boy and?
  - John thinks Sara hit the boy with the bat
  - What does John think Sara hit the boy with?
  - Colorless green ideas sleep furiously.
  - * Green sleep furiously ideas colorless.

This Research Program has Contributed Powerful Ideas

Context-free grammar

Formal language hierarchy

Syntax, Phonology…

This Research Program is Really Unfinished

Type in your English sentence here:

Is this grammatical?

Is this sensible?

Lots of Open Problems

- Modeling English fluency, using trees
  - phrase-based output – need to parse it to score it
  - syntax-based output – already in scorable tree form
  - initial work: [Charniak, Knight, and Yamada, 2003]

- Choosing syntactic categories that are appropriate for translation
  - initial work: [B. Huang and K. Knight, 2006]

- Decoder search in runtime translation
  - Search errors hurt MT accuracy
  - Faster speed is needed to support experimentation
  - Some key ideas to date:
    - cube pruning [Chiang, 2007]
    - rule binarization [Zhang, Huang, Knight, Gildea, 2006]

Lots of Open Problems

- More context for rule choice
  - compare word-based SMT
    - context-sensitive word translation probabilities [Berger et al 96]
  - compare phrase-based SMT
    - bilingual n-gram translation models [de Gispert & Mariño 02]
    - context-based phrasal TM "WSD" [Chan, Ng, Chiang 07; Carpuat & Wu 07]
- Morphology in translation rules
- More generally applicable rules
  - Adjoining transducers (tree-adjoining grammar)
- Open theory problems in the underlying automata models…

Tree Automata

Doner (1968), Rounds (1970), Thatcher (1970), Engelfriet PhD thesis (1975), Gecseg & Steinby textbook (1984), …
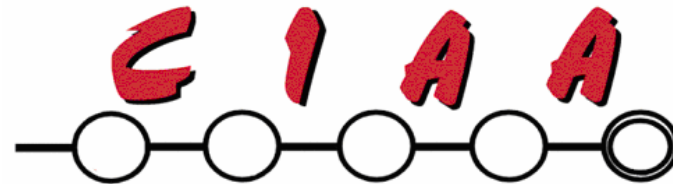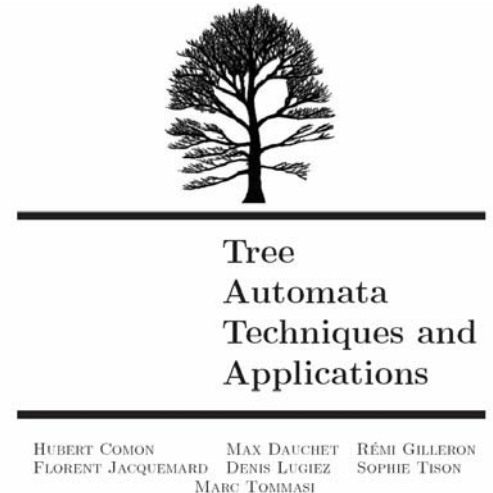
Conference on Implementation
and Application of Automata

Tiburon: A Tree Automata Toolkit

- Developed by Jonathan May, ISI
- First version distributed in April 2006, includes tutorial
- Inspired by string automata toolkits
- Prototype ideas, teach tree automata to yourself or others

- You cast your problem in terms of tree acceptors and transducers
  - doesn't have to be MT
- You get implemented algorithms for free
  - e.g., Kumar/Byrne'03 (use AT&T FSM for MT)
  - e.g., Pereira/Riley'96 (use AT&T FSM for ASR)

Tiburon: A Tree Automata Toolkit

Towards simplifying system ideas:

e = yield(best-tree(intersect(**lm.rtg**,

b-apply(cstring, **tm.tt**)))

What tree automata operations are
needed/supported?

String World & Tree World

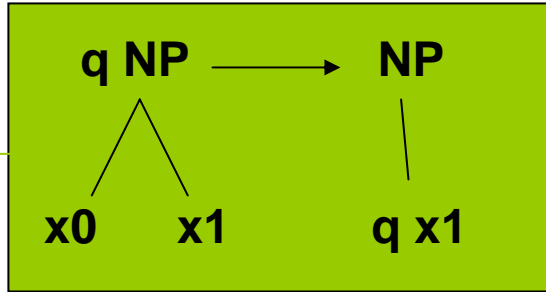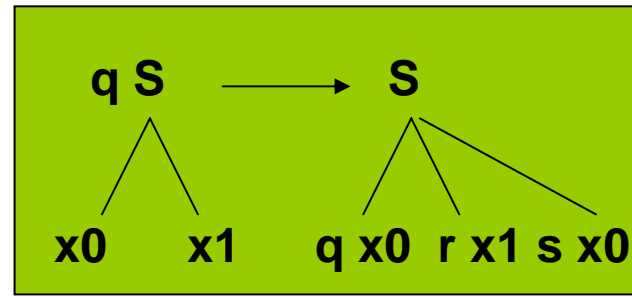|	String World	Tree World
N-best …	… paths through a lattice (Viterbi, 1967; Eppstein, 1998)	… trees in a forest (Huang & Chiang, 2005)
EM training	Forward-backward EM (Baum & Welch, 1971)	Tree transducer EM training (Graehl & Knight, 2004)
Determinization …	… of weighted string acceptors (Mohri, 1997)	… of weighted tree acceptors (May & Knight, 2005)
Intersection	WFSA intersection	Tree acceptor intersection (despite CFG not closed)
Applying transducers	string → WFST → WFSA	tree → TT → weighted tree acceptor
Transducer composition	WFST composition (Pereira & Riley, 1996)	Many tree transducers are **not closed** under composition! (Rounds 70; Engelfriet 75)

Classes of Tree Transducers
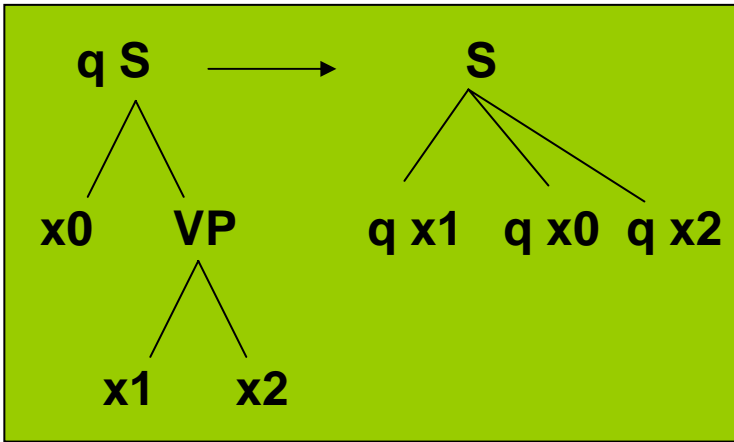
copying

non-copying

deleting

non-deleting

q S $\longrightarrow$ S

x0   x1   q x0 r x1 s x0

copying rule

q NP $\longrightarrow$ NP

x0   x1   q x1

deleting rule

T

LT

LNT

Classes of Tree Transducers

$xT_R = T_R$

xT

GK'04

T

GS'84

copying

non-copying

xLT

LT

deleting

non-deleting

$xLNT$

MGHK'08
GHKM'04

LNT

GS'84

q S $\longrightarrow$ S

x0   VP   q x1 q x0 q x2

x1   x2

x-tended rule

Expressive power theorems in Maletti, Graehl, Hopkins, Knight (submitted)

Classes of Tree Transducers

$xT_R = T_R$

T^3 B^3

T^2 B^2

bottom up transducers

xT

GK'04

T GS'84 B

copying

non-copying

xLT

L-MBOT

$LB = LT_R$

M'06

deleting

LT

non-deleting

$xLNT$

MGHK'08
GHKM'04

LNT

GS'84

Classes of Tree Transducers
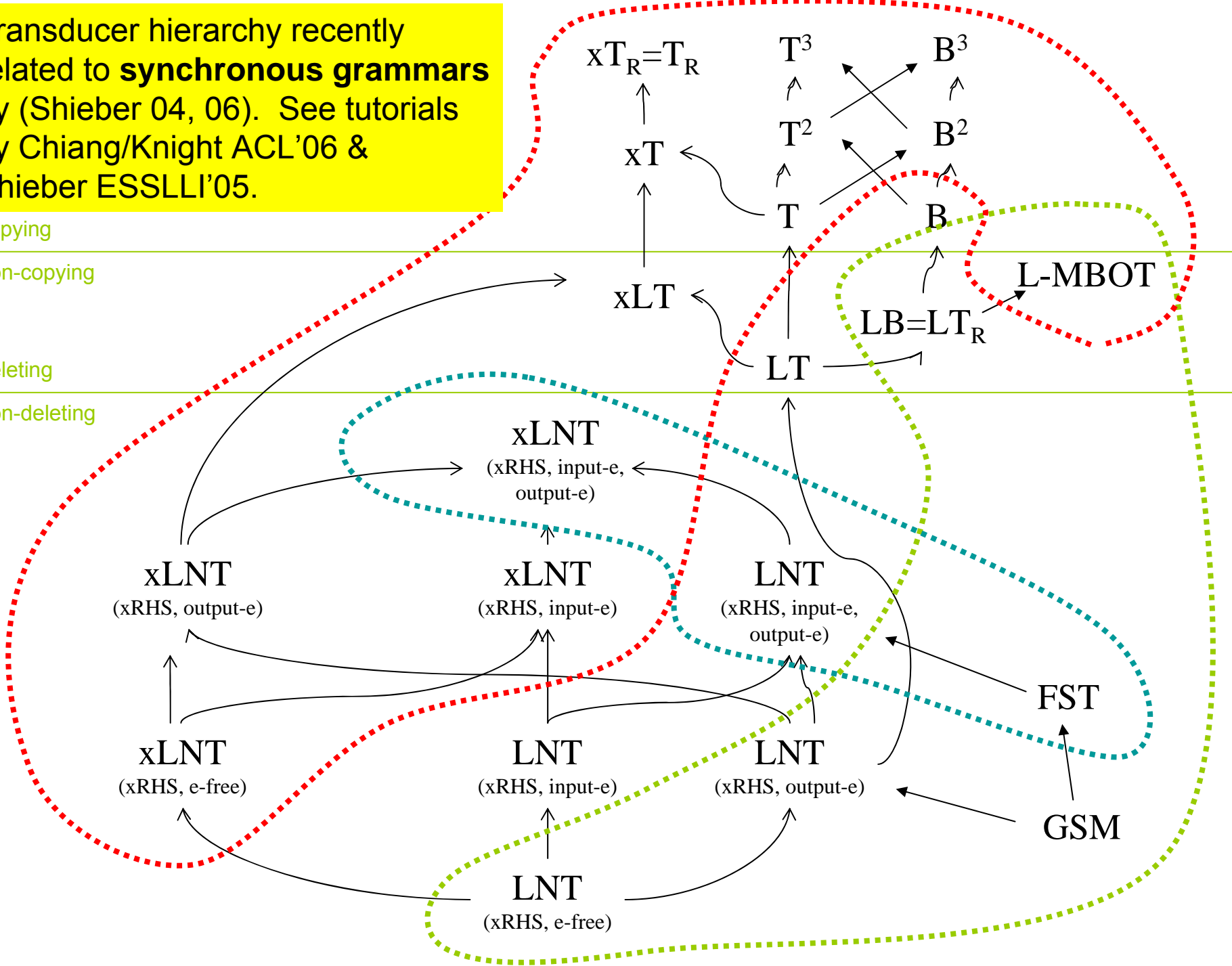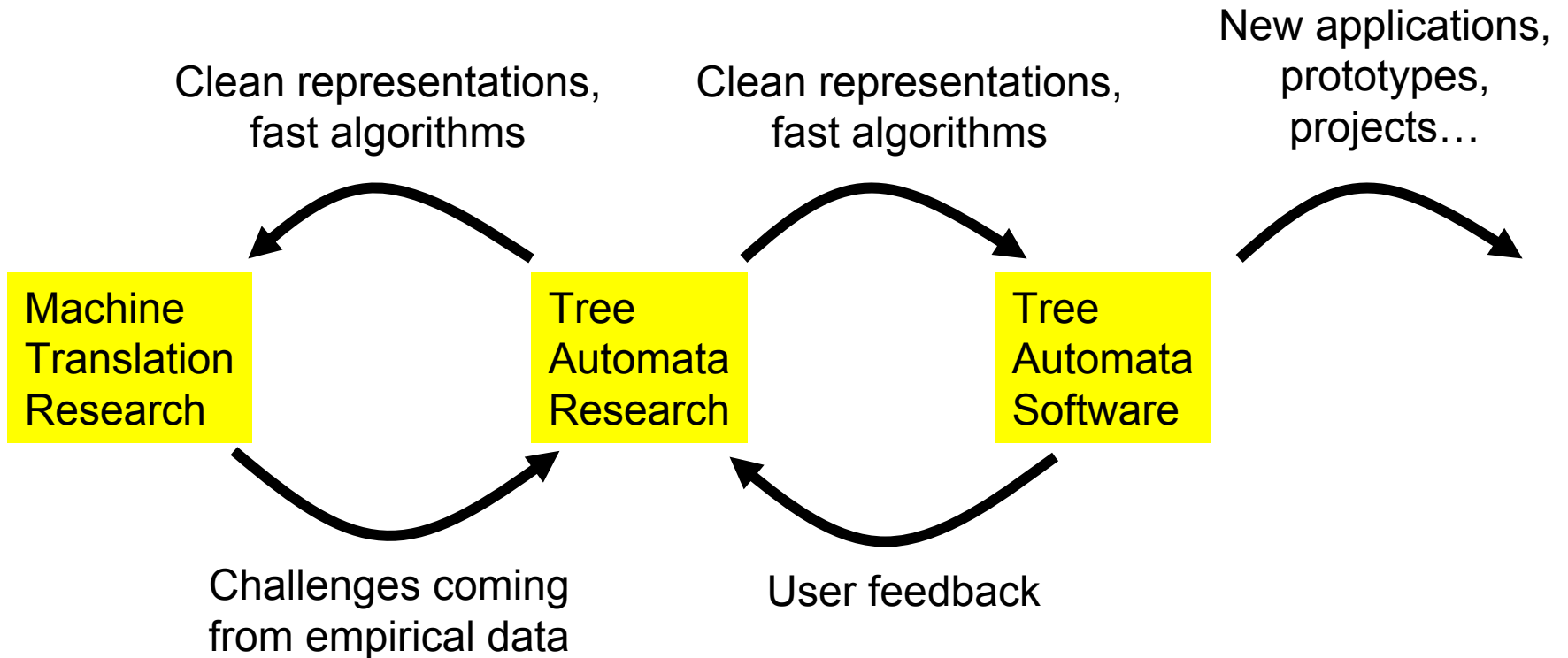
Transducer hierarchy recently related to **synchronous grammars** by (Shieber 04, 06). See tutorials by Chiang/Knight ACL'06 & Shieber ESSLLI'05.

copying

non-copying

deleting

non-deleting

$xT_R = T_R$

T^3 B^3

T^2 B^2

xT T   B

xLT $LB = LT_R$ L-MBOT

LT

xLNT
(xRHS, input-e, output-e)

xLNT
(xRHS, output-e)   xLNT
(xRHS, input-e)   LNT
(xRHS, input-e, output-e)

xLNT
(xRHS, e-free)   LNT
(xRHS, input-e)   LNT
(xRHS, output-e)   FST

GSM

LNT
(xRHS, e-free)

Research Synergy

This Is Interdisciplinary Research

- Machine Learning
- Engineering
- Linguistics
- Data
- Efficient search algorithms
- Automata theory
- Grid computing

…