# Can a Machine Translate Without Knowing What a Verb Is?

Kevin Knight

USC/Information Sciences Institute


Joint work with ISI/LW folks…

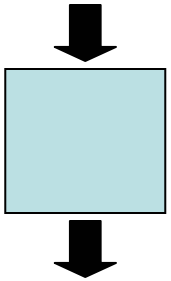
*Daniel Marcu, Wei Wang, Jonathan Graehl, Michael Pust, Jens Voeckler, Ignacio Thayer, Radu Soricut, Dragos Munteanu, Alex Fraser, Steven DeNeefe, Jonathan May*

… and exceptional summer visitors!

*Michel Galley, Mark Hopkins, Liang Huang, Hao Zhang Victoria Fossum, David Kauchak*

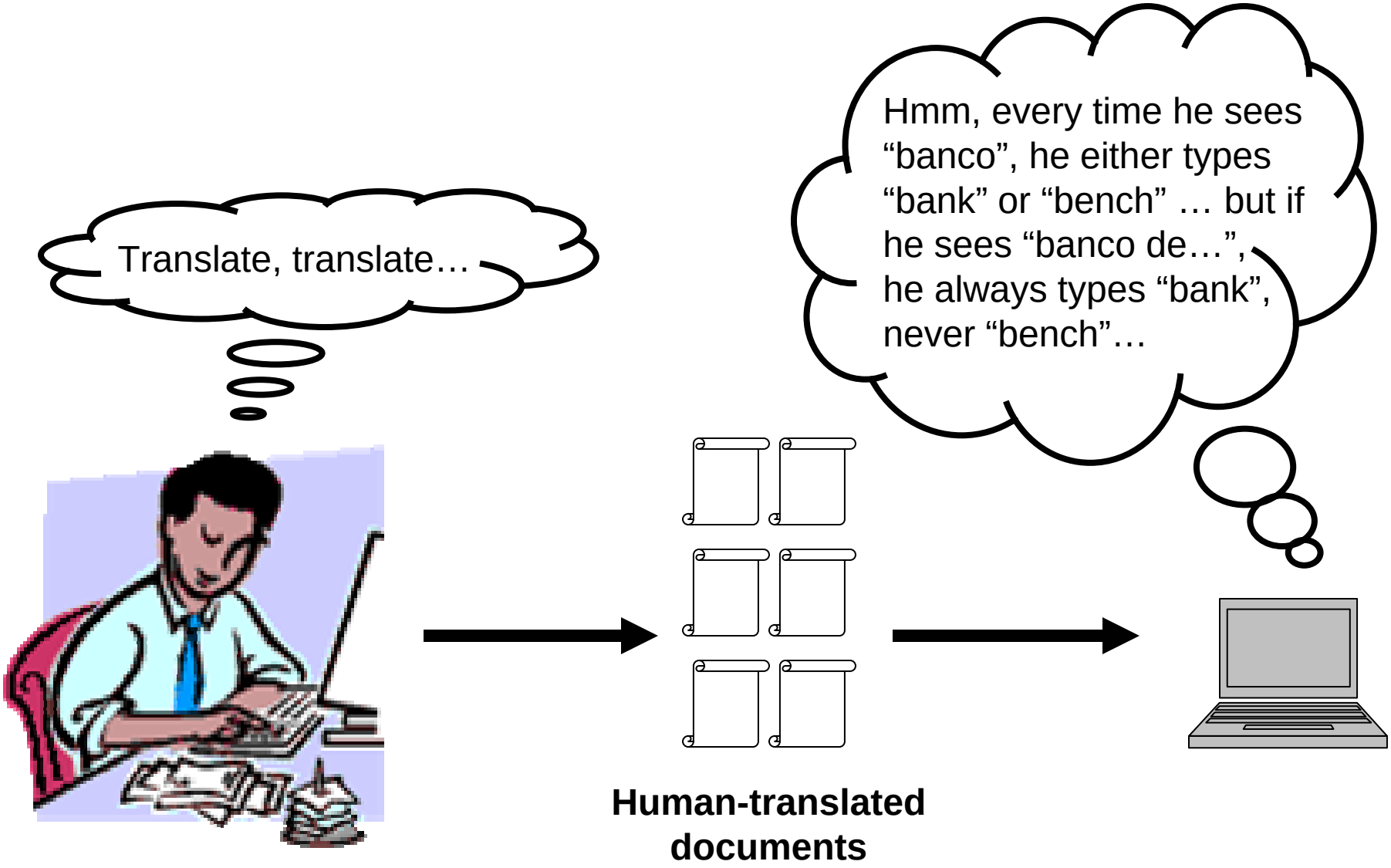
# Machine Translation

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。

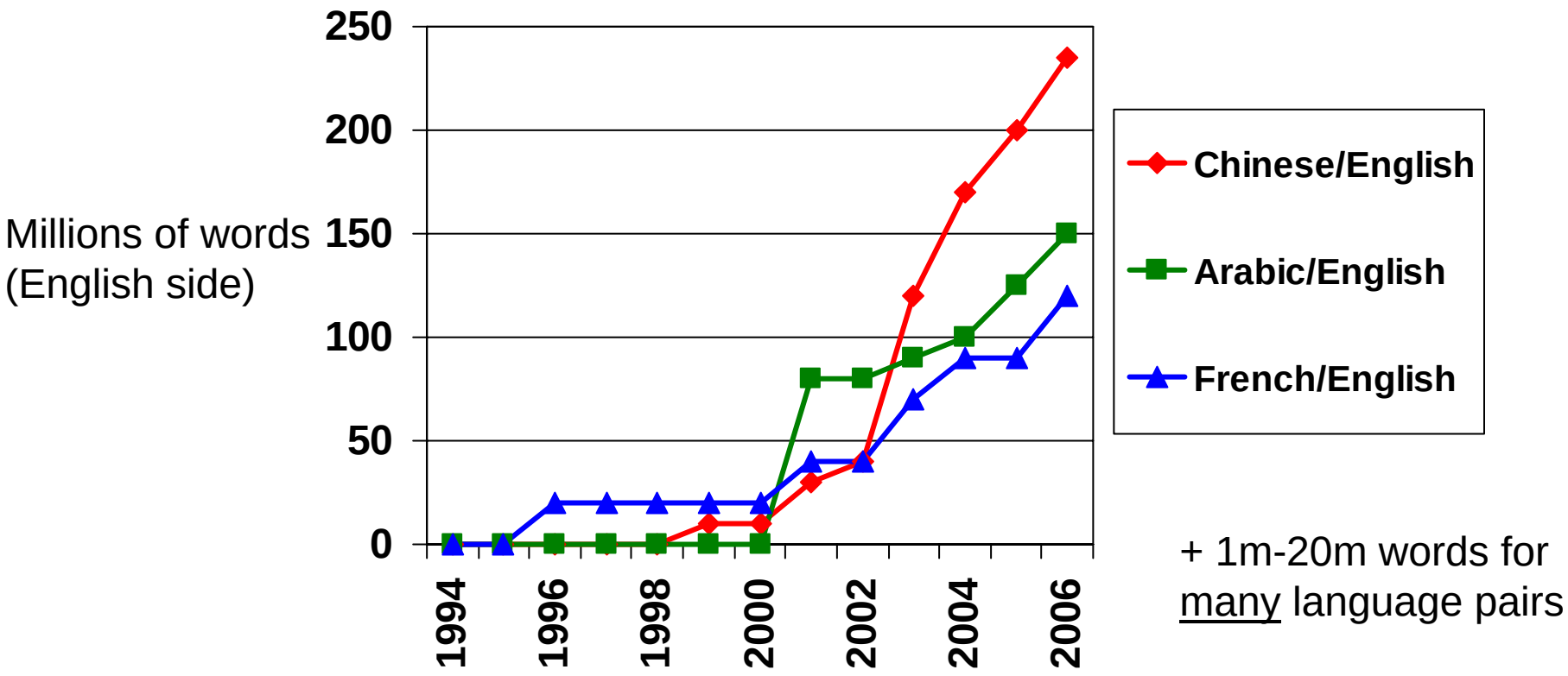


The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

# Statistical Machine Translation



# Ready-to-Use Online Bilingual Data




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

# Bilingual Text (200m words)

English strings - - - - - - - - - - - - - - - - - - … - - - - - -

Bilingual text

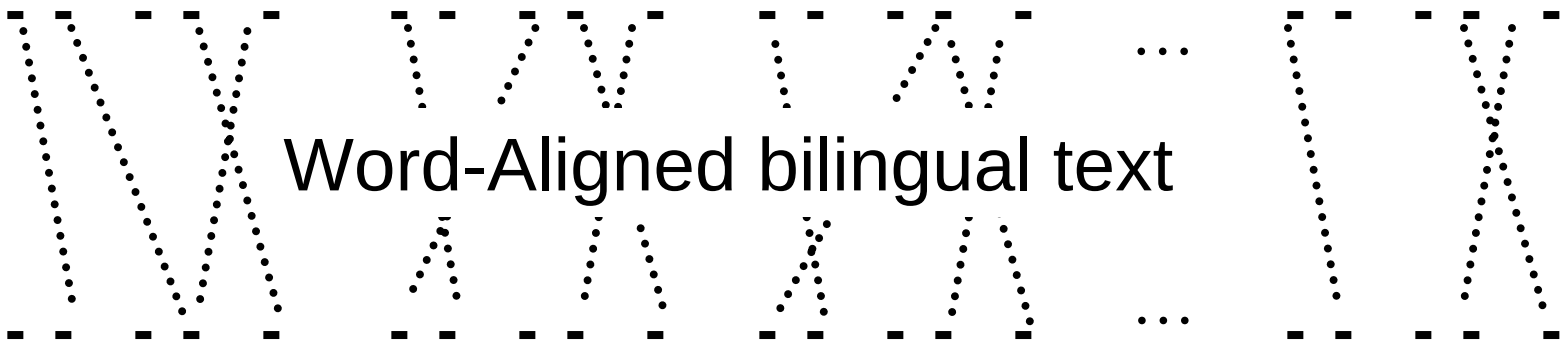
Chinese strings - - - - - - - - - - - - - - - - - - … - - - - - -

# Bilingual Text (200m words)

English strings

Word alignments

Chinese strings



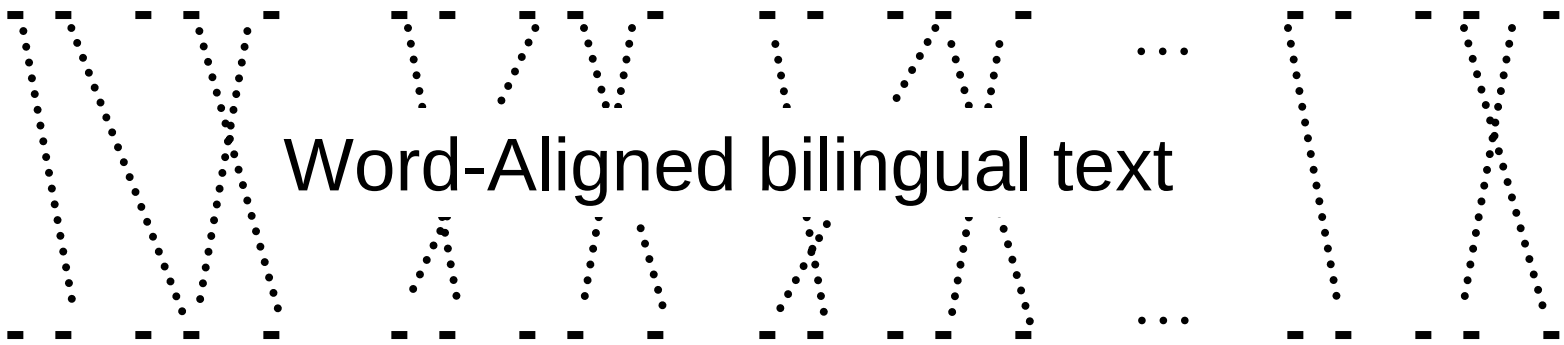
Word-Aligned bilingual text

# Bilingual Text (200m words)

English strings

Word alignments

Chinese strings




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

Vast Database of Phrase Pairs

# 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.

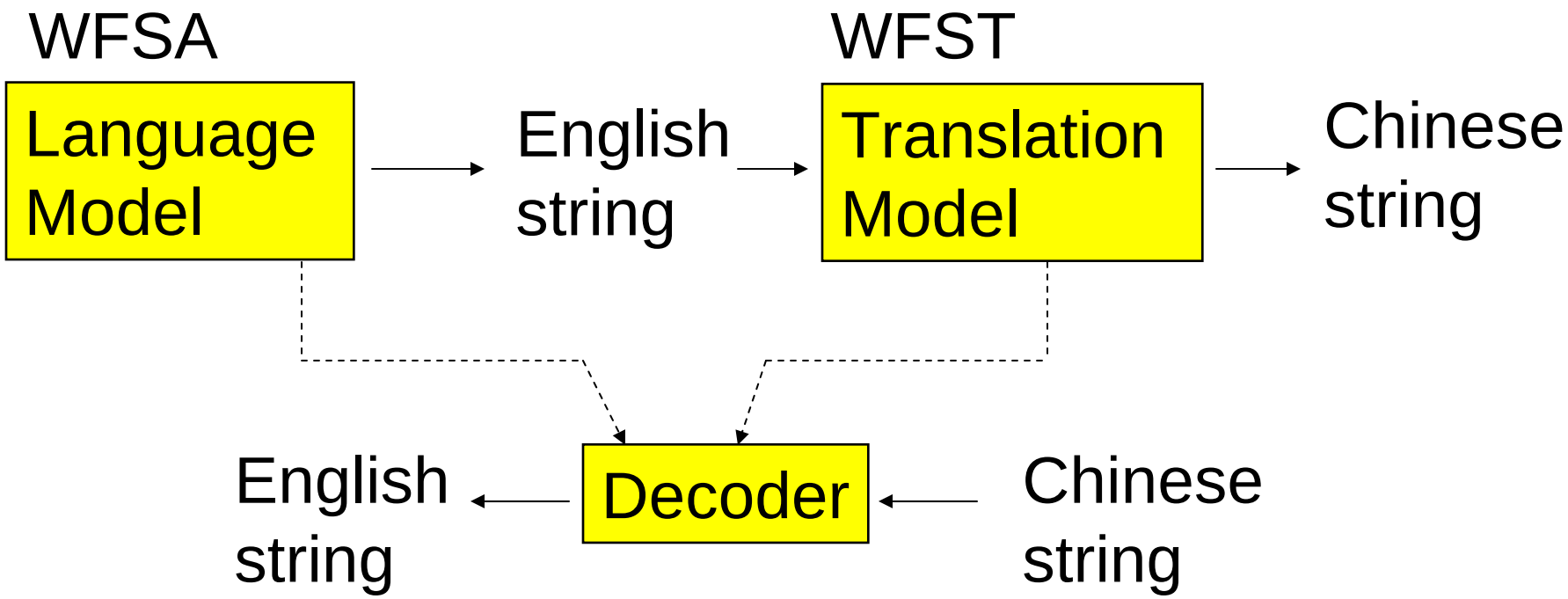
# Phrase-Based Translation

| 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 .

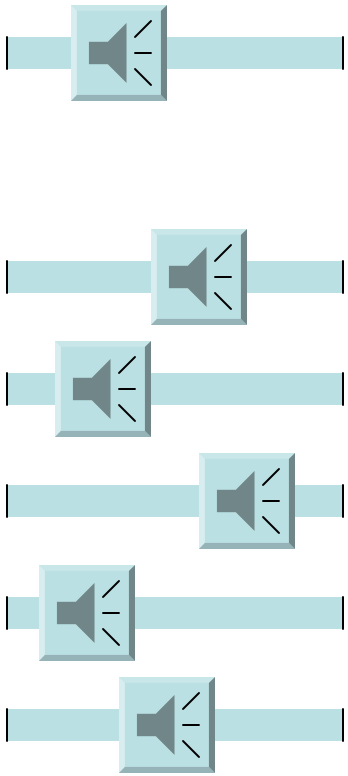
Scoring: Try to use phrase pairs that have been frequently observed.
Try to output a sentence with frequent English word sequences.

# Phrase-Based Noisy Channel



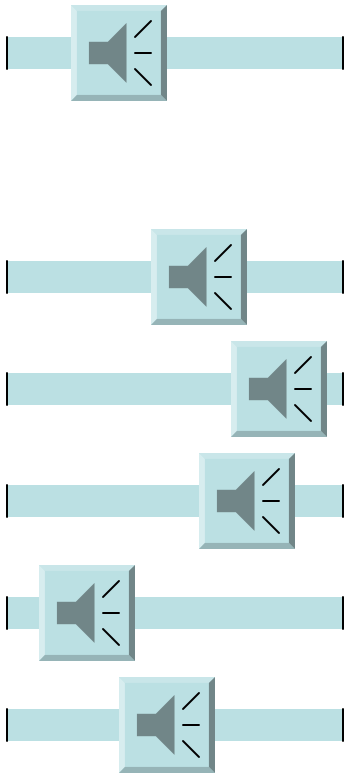
# Features and Tuning

- English trigram language model
- Phrase pairs
  - Conditional probability
  - Bad-phrase spotter
  - Word-drop spotter
  - “Move Me” preference
- English output length



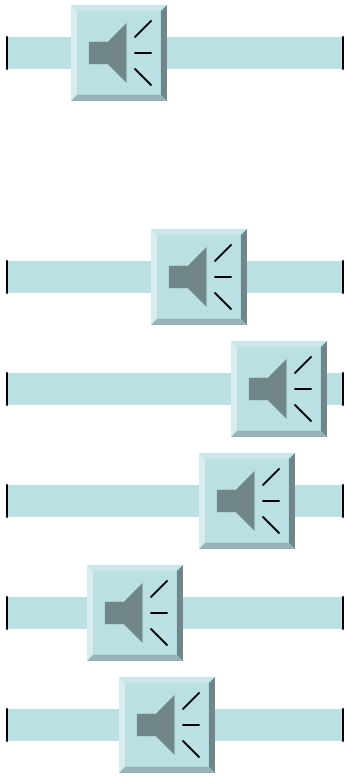
# Features and Tuning

- English trigram language model
- Phrase pairs
  - Conditional probability
  - Bad-phrase spotter
  - Word-drop spotter
  - “Move Me” preference
- English output length



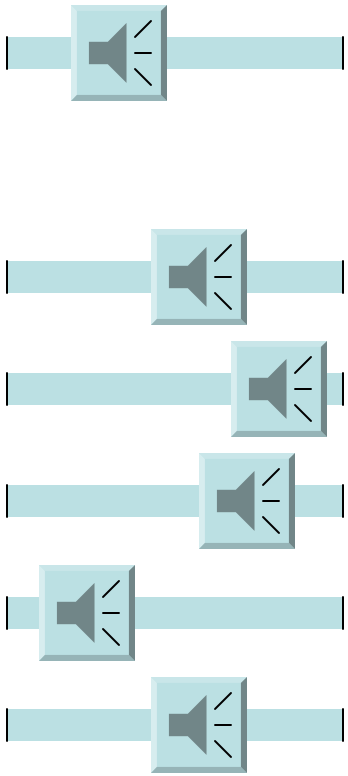
# Features and Tuning

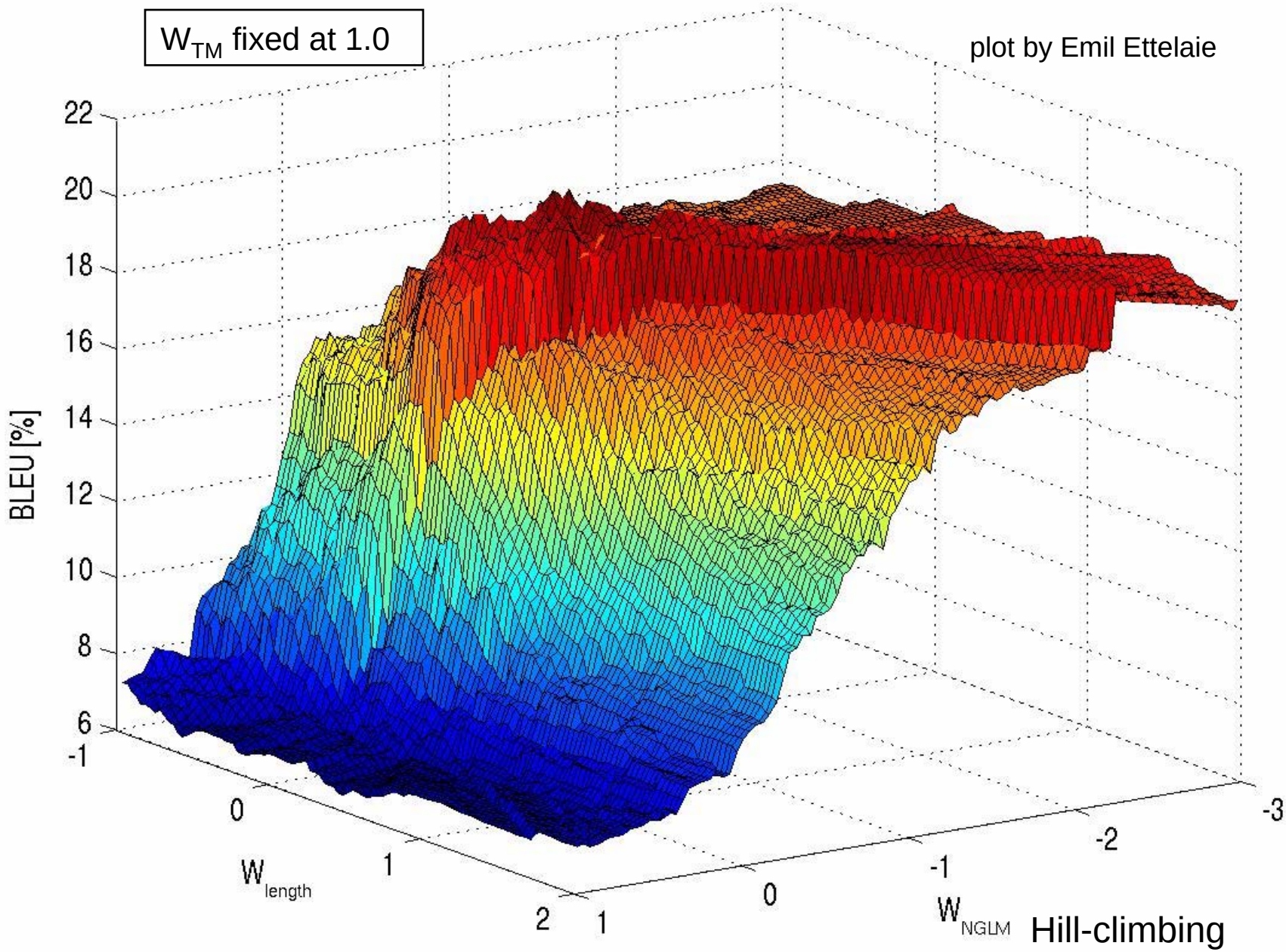
- English trigram language model
- Phrase pairs
  - Conditional probability
  - Bad-phrase spotter
  - Word-drop spotter
  - “Move Me” preference
- English output length



# Features and Tuning

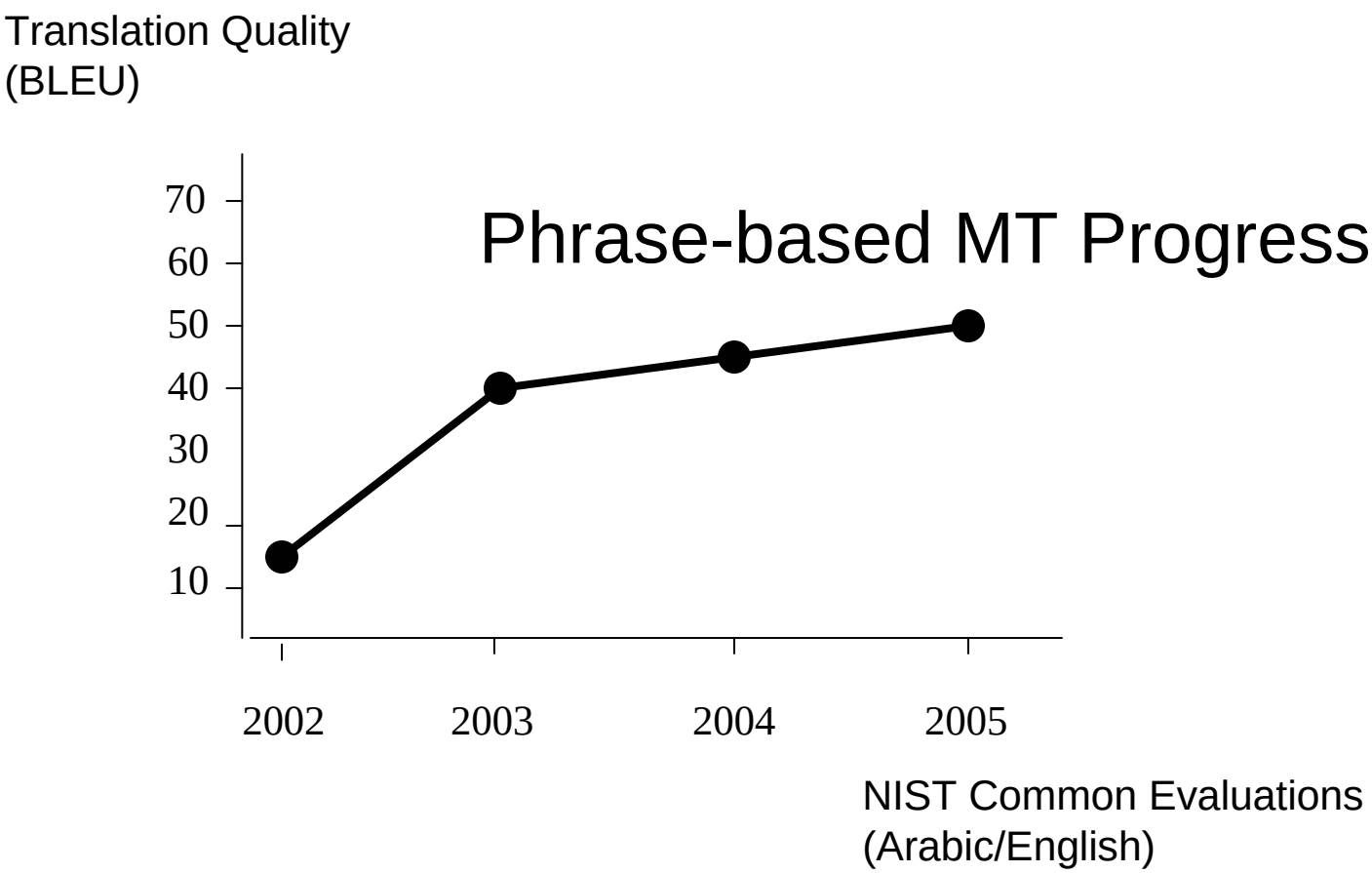
- English trigram language model
- Phrase pairs
  - Conditional probability
  - Bad-phrase spotter
  - Word-drop spotter
  - “Move Me” preference
- English output length






Hill-climbing
MaxBleu [Och 2003]

# These Ideas Work!



# Can a machine translate between Chinese and English without knowing what a verb is?

- Of course
- But the output is frequently bad

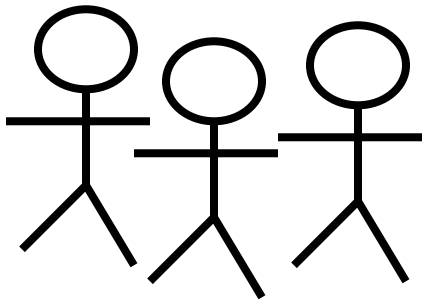
"Frequent high-tech exports are bright spots for foreign trade growth of Guangdong has made important contributions."

- This phrase-based story is a little bit crazy

# Syntax

Maybe we need some grammar?

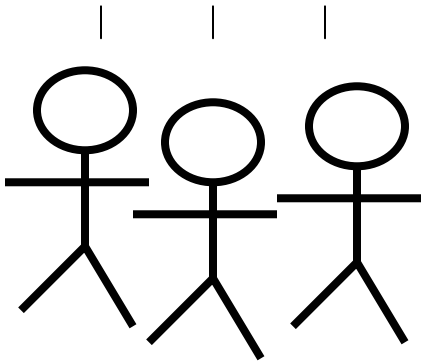
# MT Research Landscape



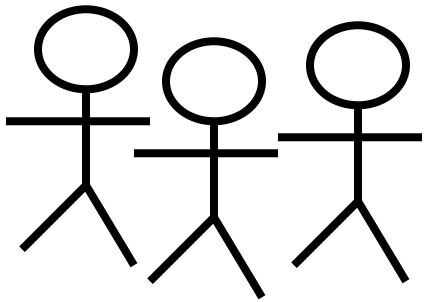
Working on syntax-based approach to translation (nouns, verbs, prepositional phrases…)

# 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!



Language Engineers

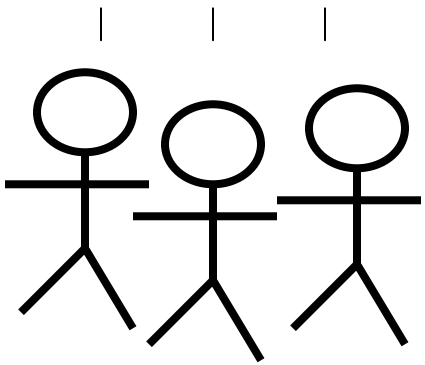


Working on syntax-based approach to translation (nouns, verbs, prepositional phrases…)

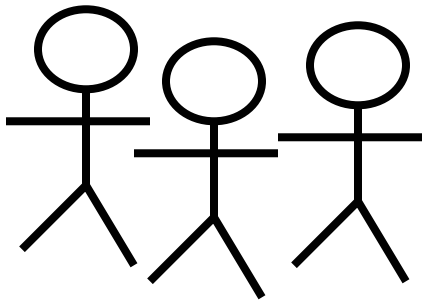
# MT Research Landscape

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!

[Koehn et al 03]



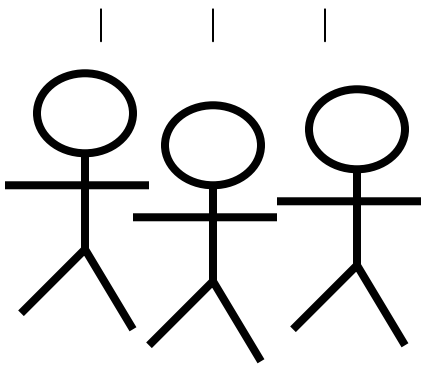
Language Engineers



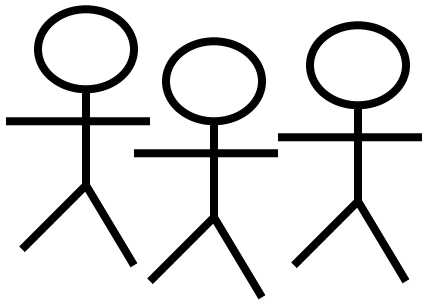
Working on syntax-based approach to translation (nouns, verbs, prepositional phrases…)

# 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!

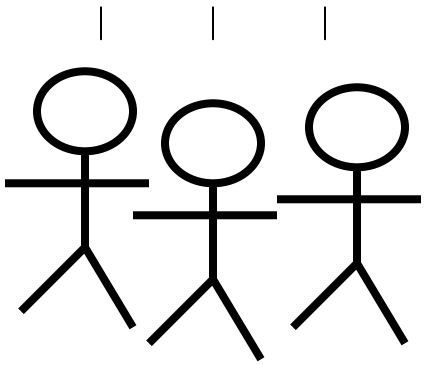


Language Engineers



Working on syntax-based approach to translation (nouns, verbs, prepositional phrases…)

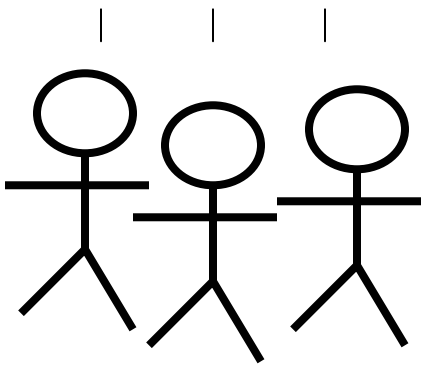
Syntax will never work!
You need *semantics*!
Language is about the world!
You are crazy!



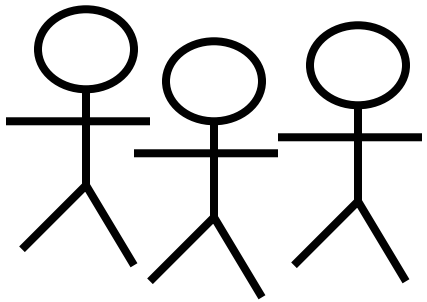
AI Fellows

# 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!**

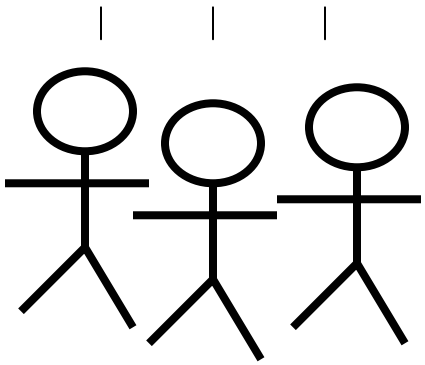


Language Engineers



Working on syntax-based approach to translation (nouns, verbs, prepositional phrases…)

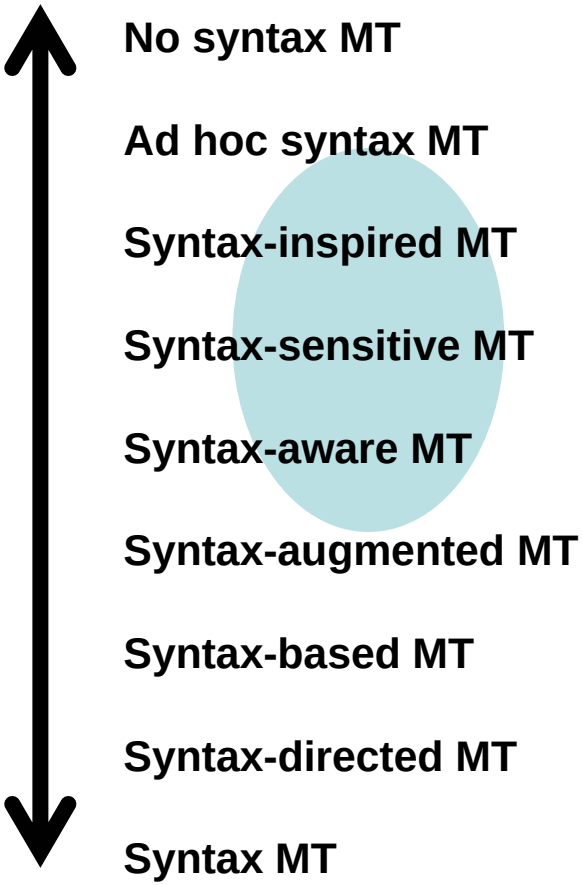
**Syntax will never work!**
You need *semantics*!
Language is about the world!
**You are crazy!**



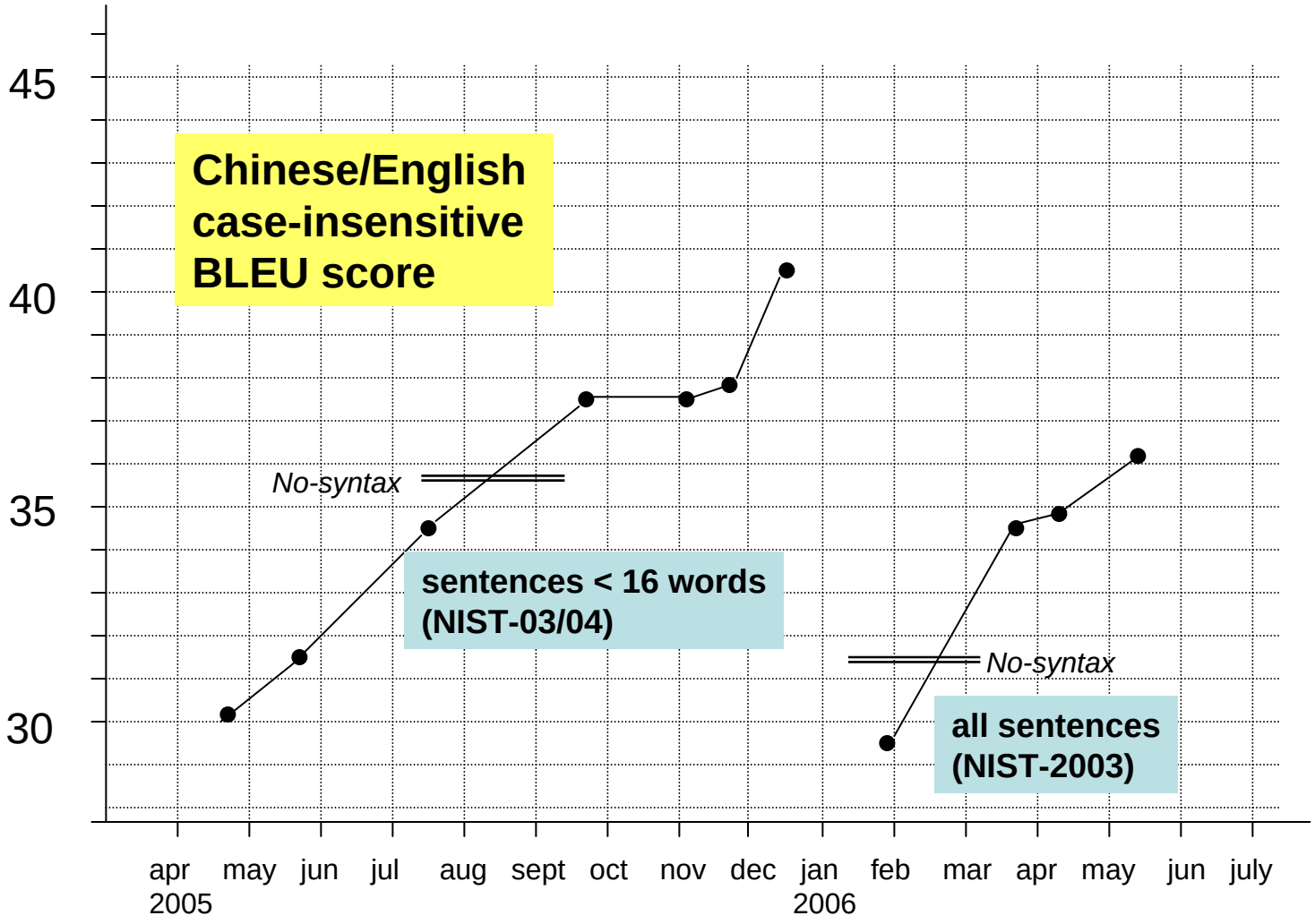
AI Fellows

# NIST-2005 Common Evaluation of Machine Translation Systems

- Chinese/English
  - ISI No-Syntax system: 30.7
  - ISI Syntax system: 24.3
  - Google system: ~35
- Higher is better (not like golf)



# Syntax Started to Work in 2006…



# **NIST-2006** Common Evaluation of Machine Translation Systems

- Chinese/English
  - ISI No-Syntax system: ~30
  - ISI Syntax system: 33.9

- Similar results for BN genre
- No-syntax = Syntax for BC genre and for Arabic/English
- Detailed testing with ASR just beginning

**No syntax MT**

**Ad hoc syntax MT**

**Syntax-inspired MT**

**Syntax-sensitive MT**

**Syntax-aware MT**

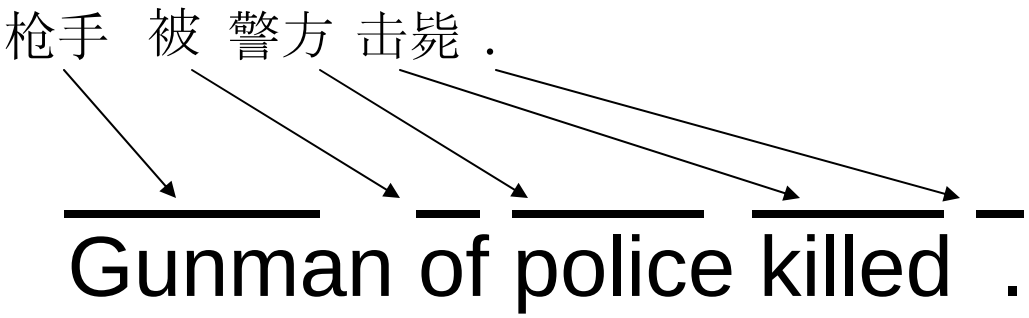
**Syntax-augmented MT**

**Syntax-based MT**

**Syntax-directed MT**

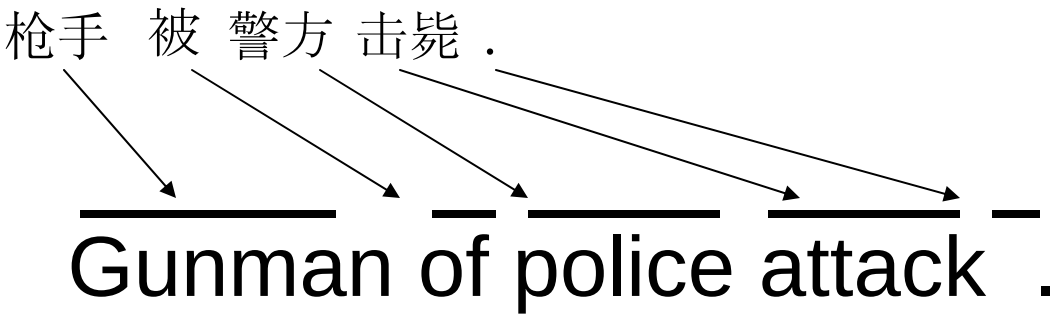
**Syntax MT**

# Phrase-Based Output



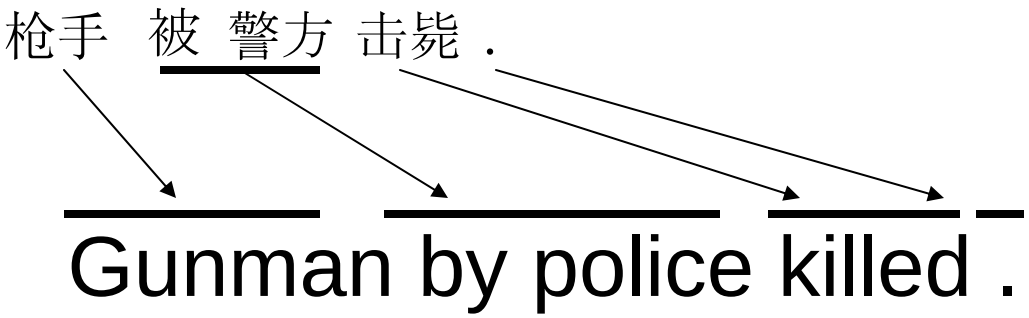

*Decoder Hypothesis #1*

# Phrase-Based Output



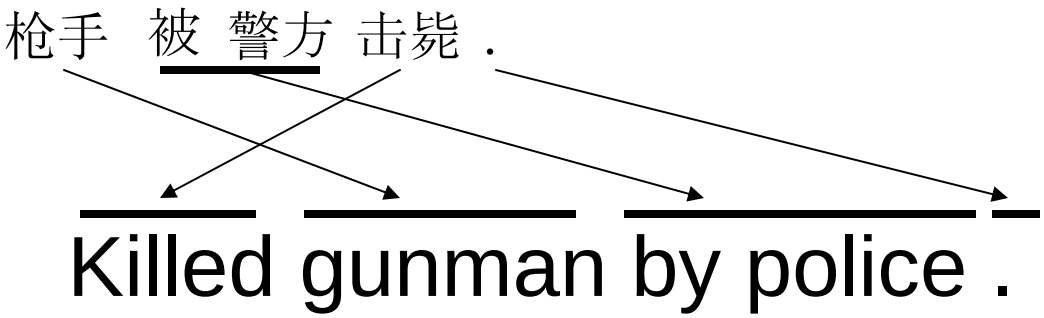

*Decoder Hypothesis #7*

# Phrase-Based Output



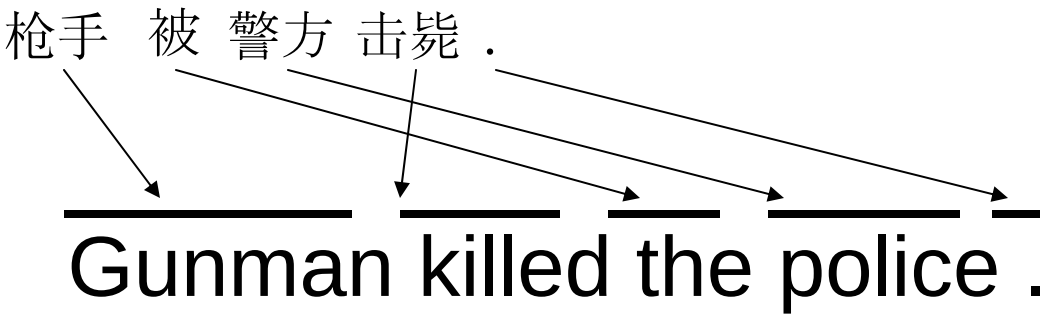

*Decoder*
*Hypothesis #12*

# Phrase-Based Output



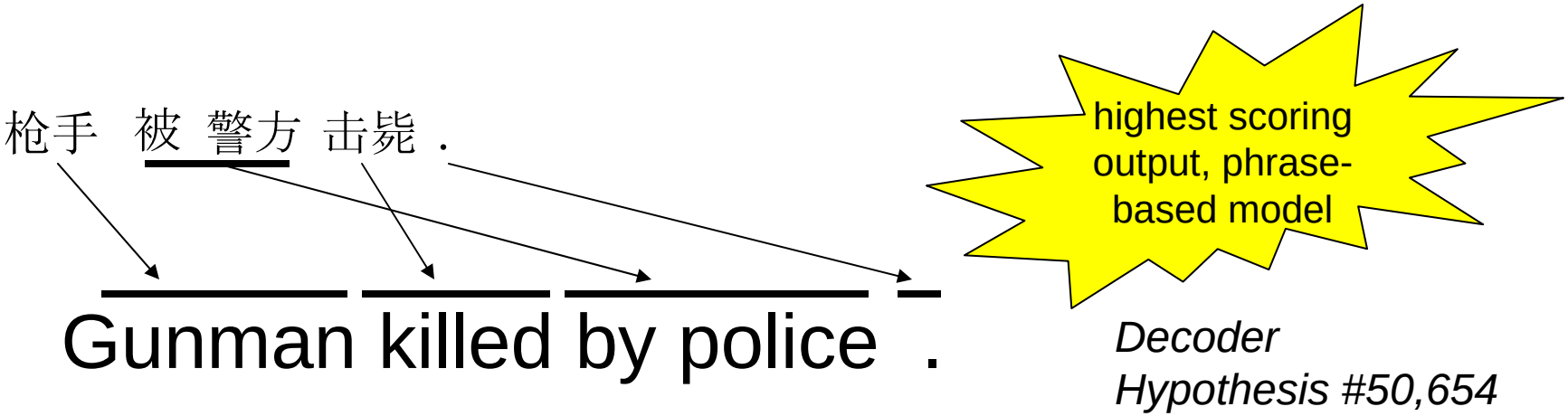

*Decoder*
*Hypothesis #134*

# Phrase-Based Output




*Decoder*
*Hypothesis #9,329*

# Phrase-Based Output




**Problematic:**

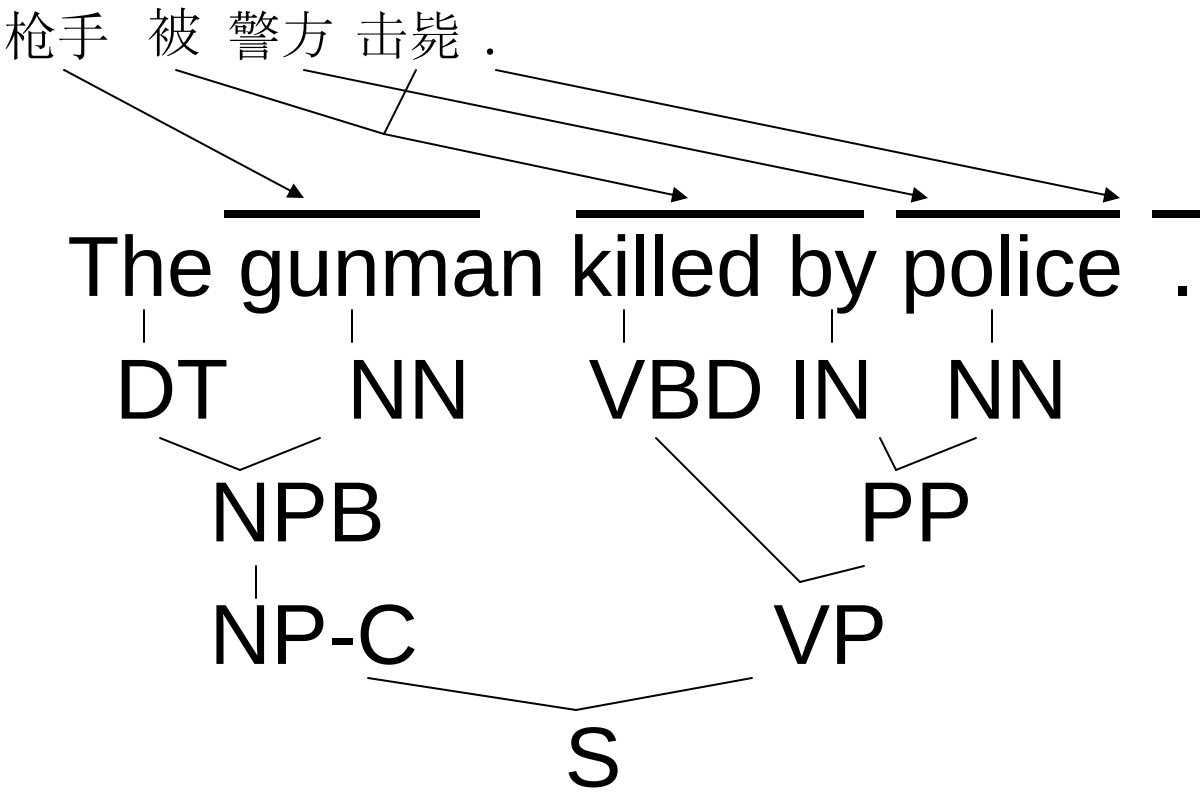
- VBD “killed” needs a direct object
- VBN “killed” needs an auxiliary verb (“was”)
- countable “gunman” needs an article (“a”, “the”)
- “passive marker” in Chinese controls re-ordering

**Can’t enforce/encourage any of this!**

# How to Get Grammar into the Statistical MT Picture?

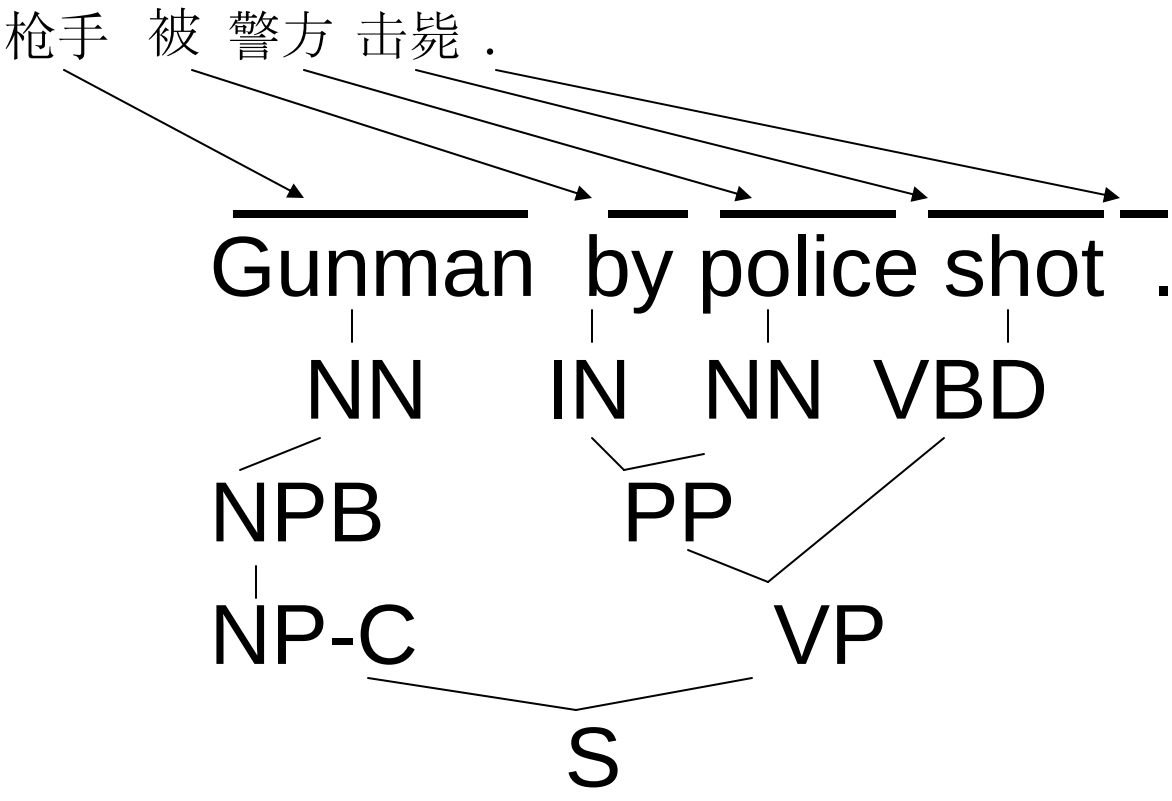
- Original work by Dekai Wu
- Yamada & Knight (2001, 2002)
- Galley, Hopkins, Marcu & Knight (2004)

# Syntax-Based Output



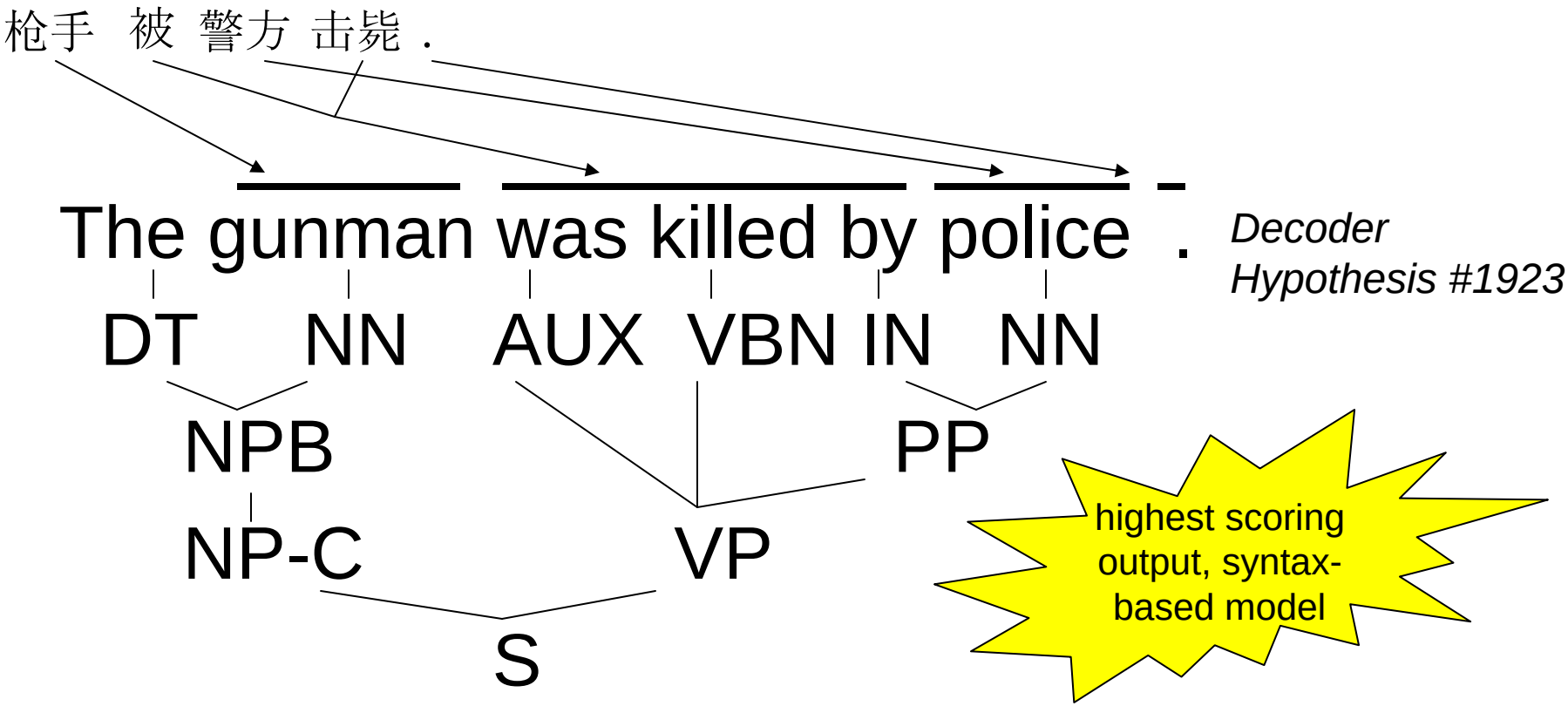

*Decoder Hypothesis #1*

# Syntax-Based Output




*Decoder Hypothesis #16*

# Syntax-Based Output




*Decoder Hypothesis #1923*

# 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
  - They control translation
- Better generalization in translation patterns

# Why Target Trees Instead of Source Trees?

- Human translators need to know a lot more about the target language.
- MT system seems to know Chinese just fine.
  - Any evidence to the contrary?
- But the system does not know English!
  - Lots of evidence
- Speech input to MT
  - We don’t have to parse source speech recognition
  - We can generalize to source lattices instead of strings

这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

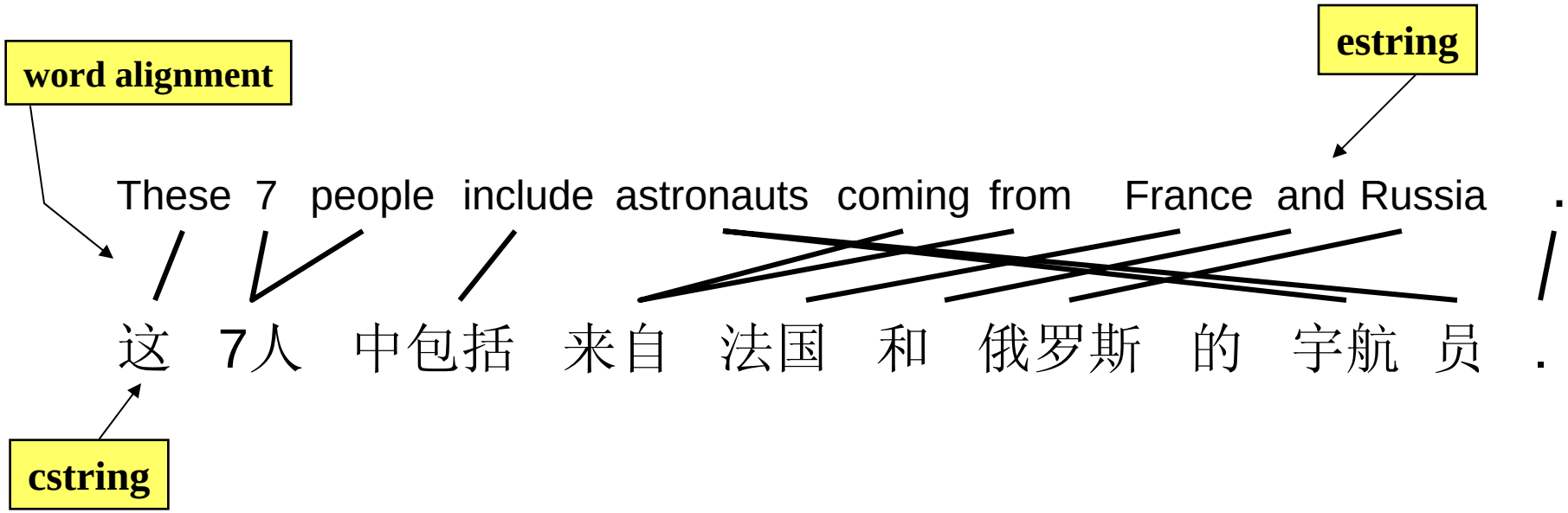
cstring

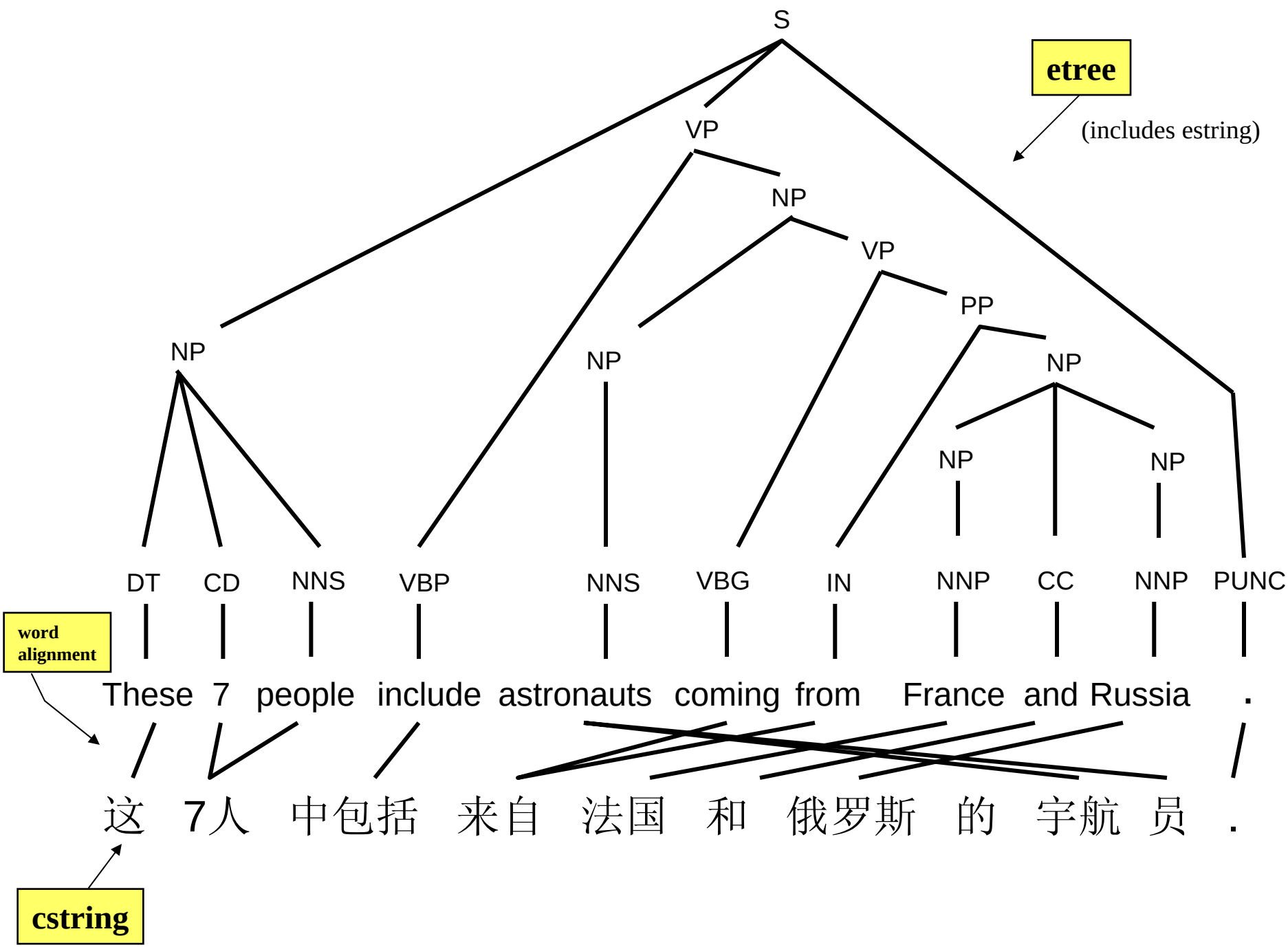
**estring**

These 7 people include astronauts coming from France and Russia .

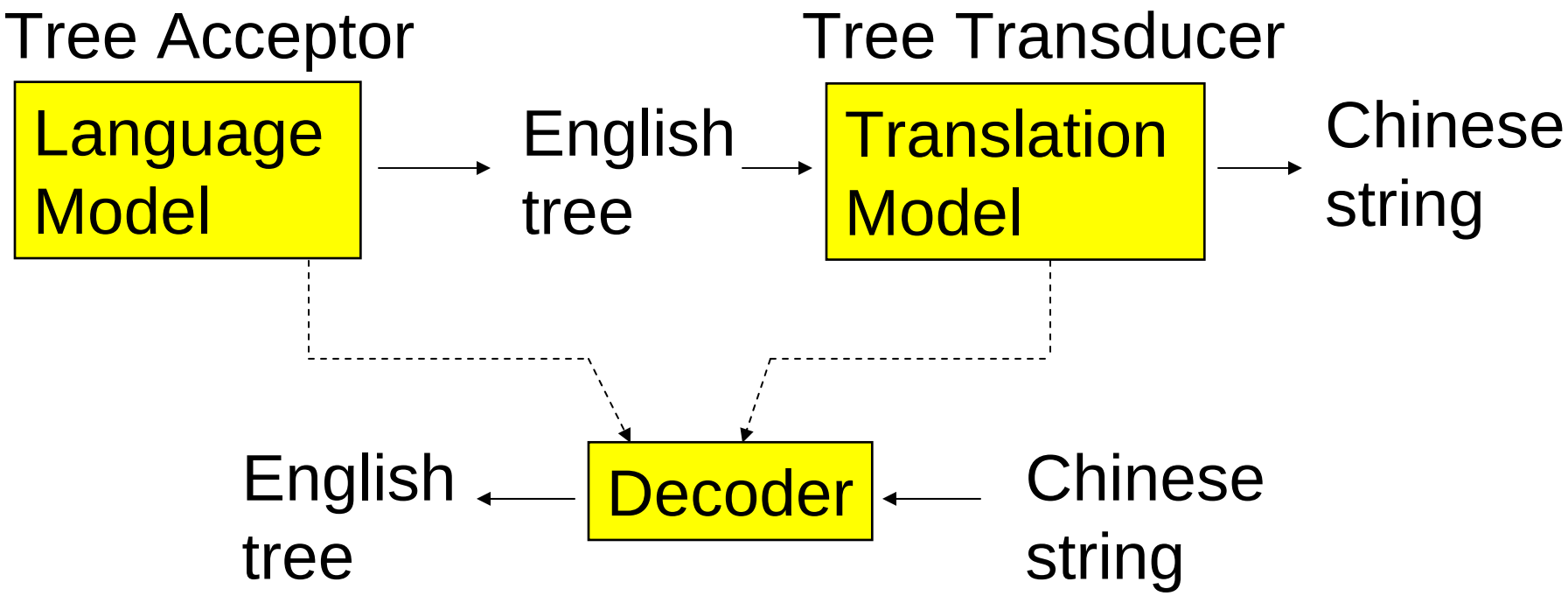
这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

**cstring**


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

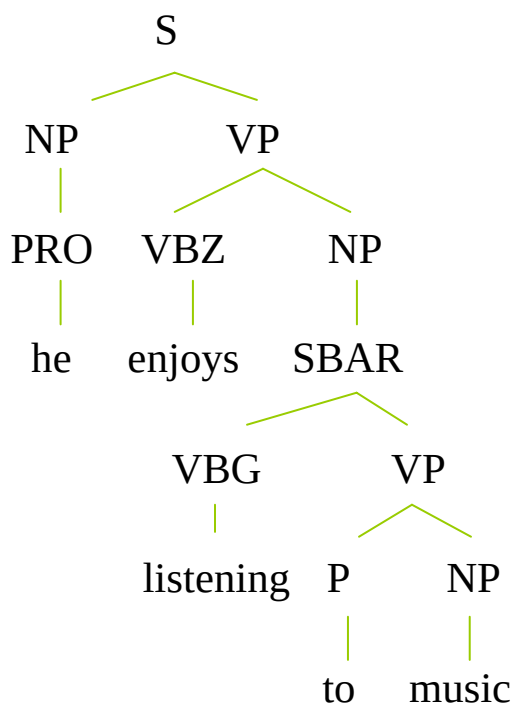

S
etree
(includes estring)
VP
NP
VP
PP
NP
NP
NP
NP
NP
DT
CD
NNS
VBP
NNS
VBG
IN
NNP
CC
NNP
PUNC
word alignment
These 7 people include astronauts coming from France and Russia .
这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .
cstring

# Syntax-Based Noisy Channel

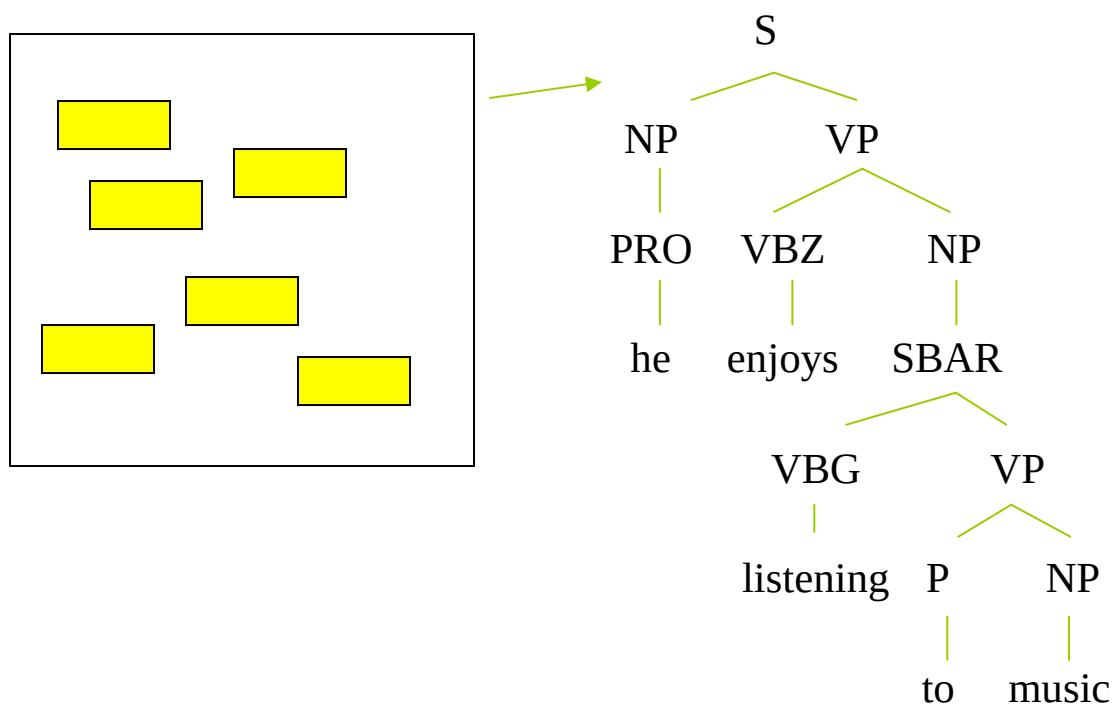


# Top-Down Tree-to-String Transducer

Original input:

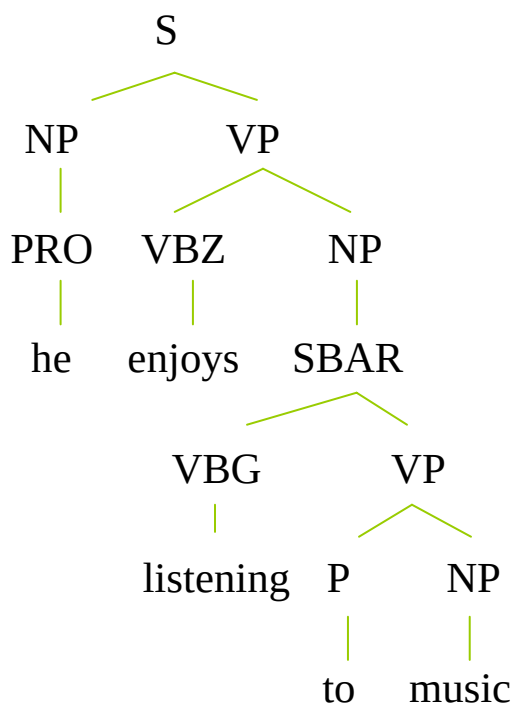



Transformation:

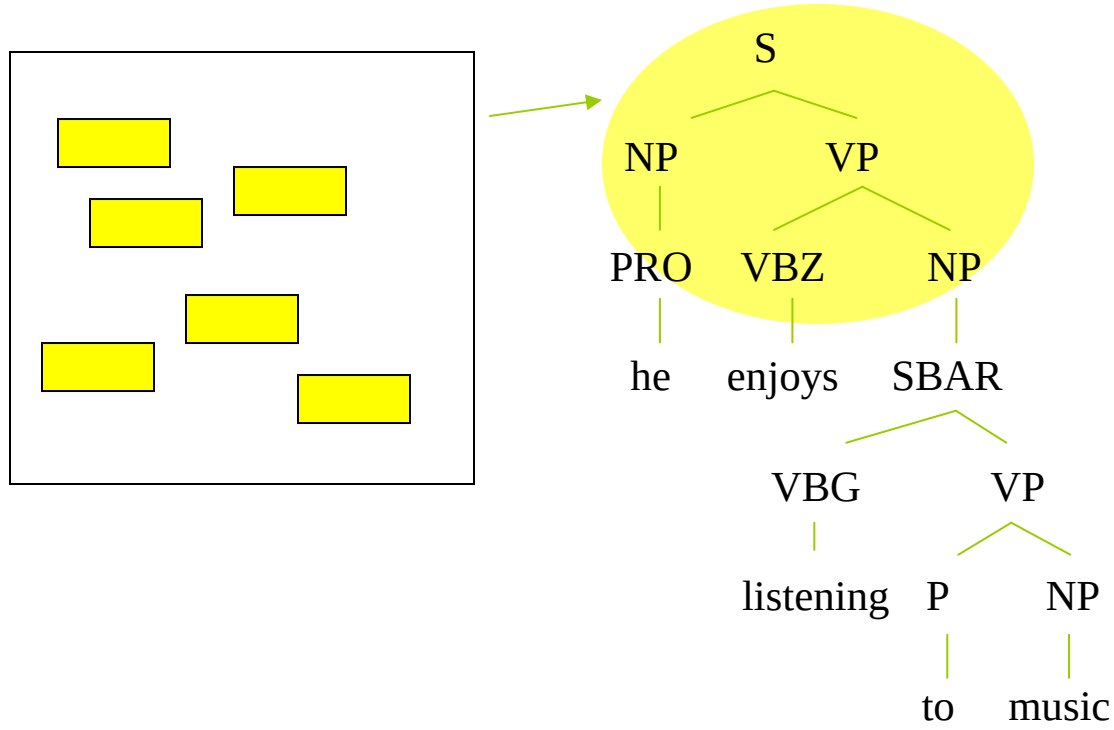


# Top-Down Tree-to-String Transducer

Original input:

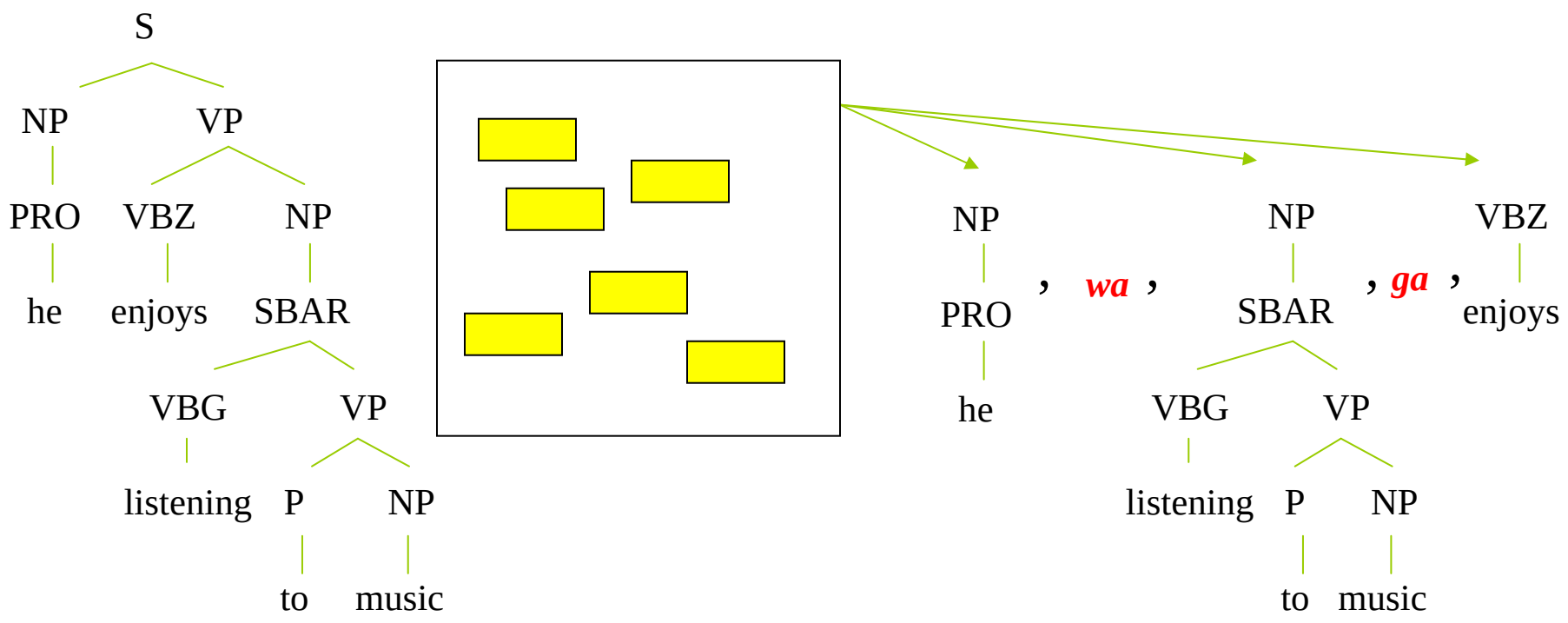



Transformation:



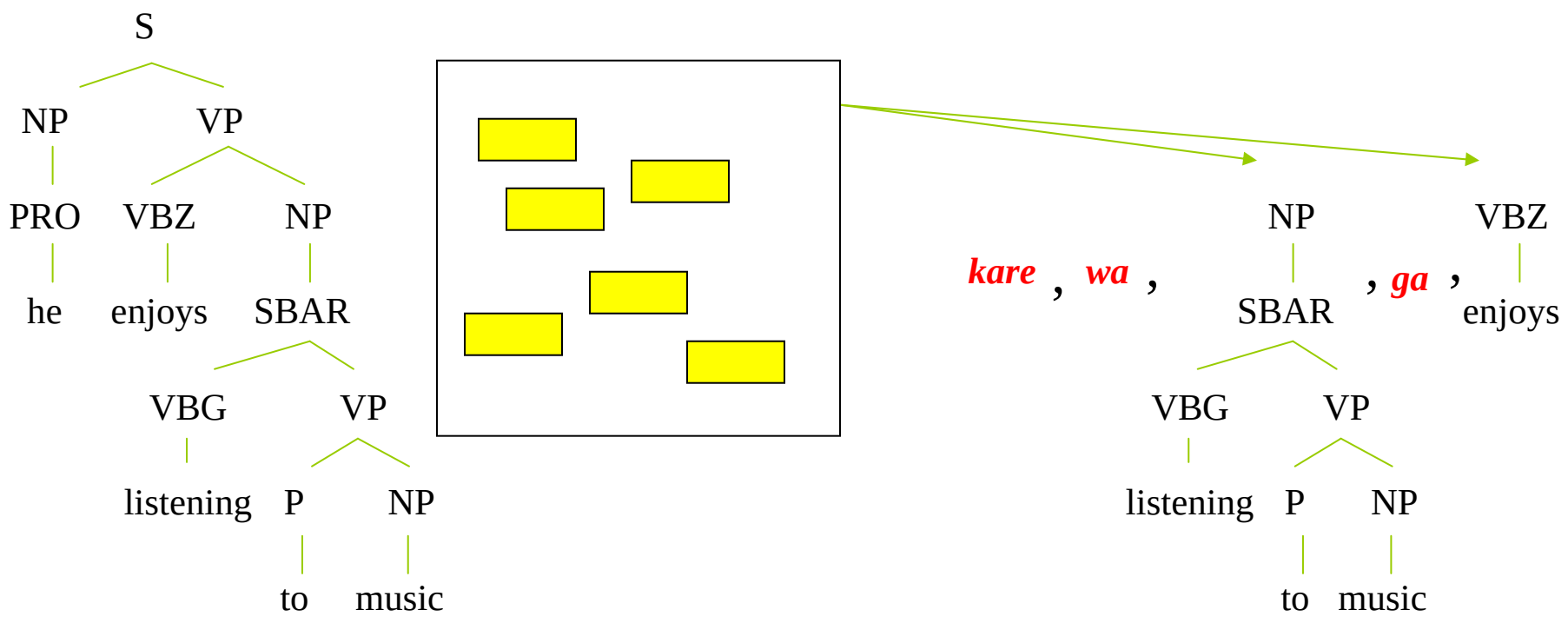
# Top-Down Tree-to-String Transducer

Original input: Transformation:



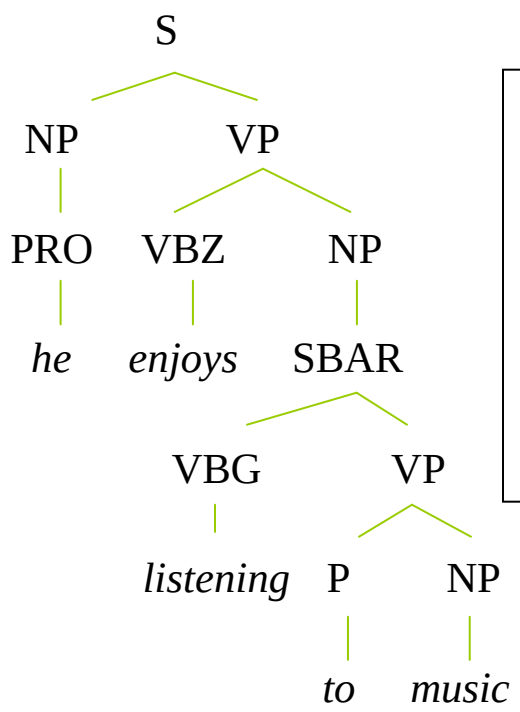
# Top-Down Tree-to-String Transducer

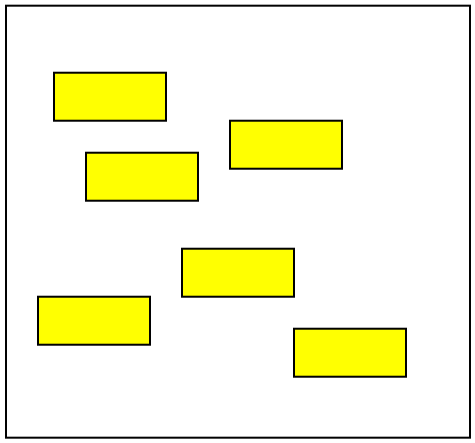
Original input: Transformation:



# Top-Down Tree-to-String Transducer

Original input:



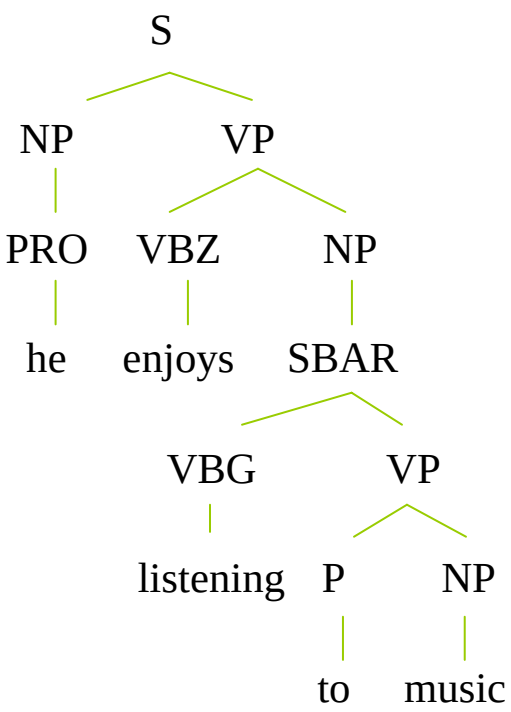



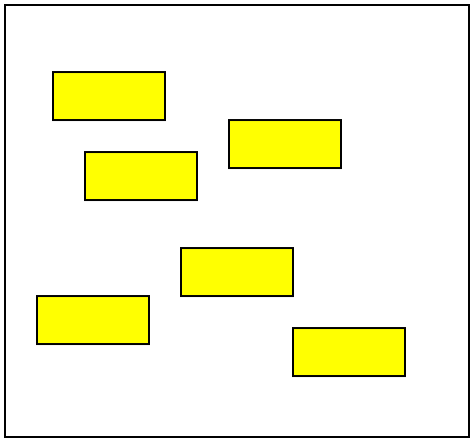
Final output:

***kare*, *wa* , *ongaku*, *o* , *kiku* , *no*, *ga*, *daisuki*, *desu***

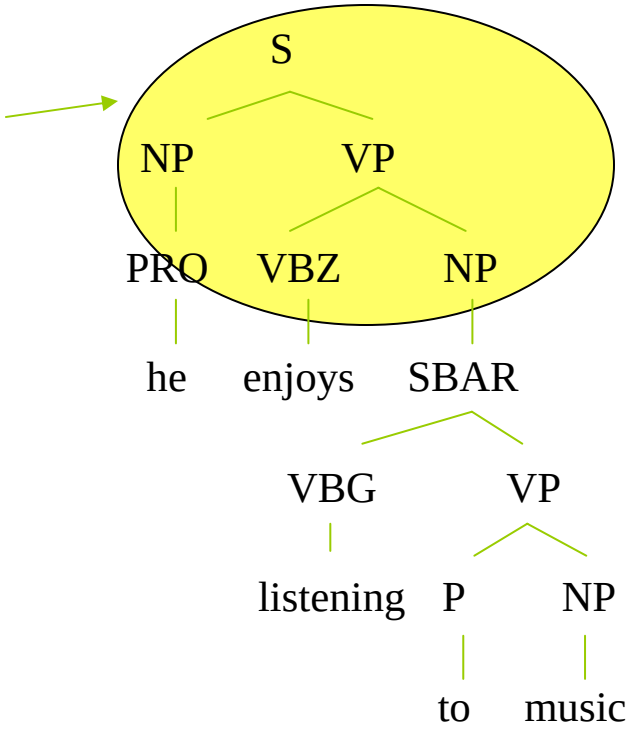
# Top-Down Tree-to-String Transducer

Original input:



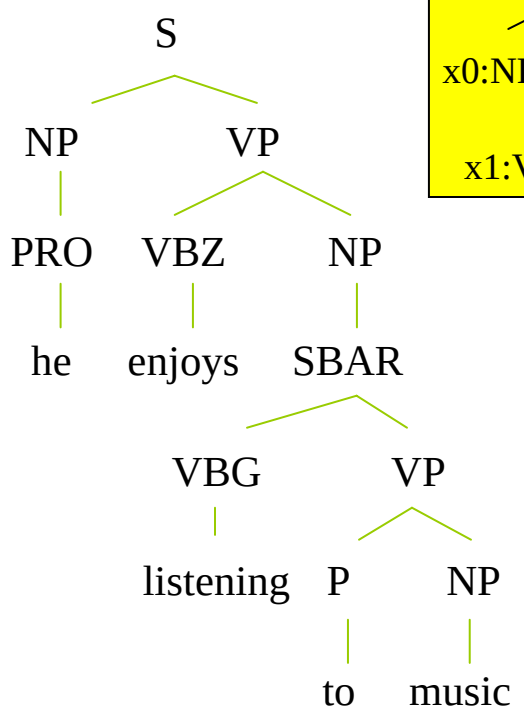



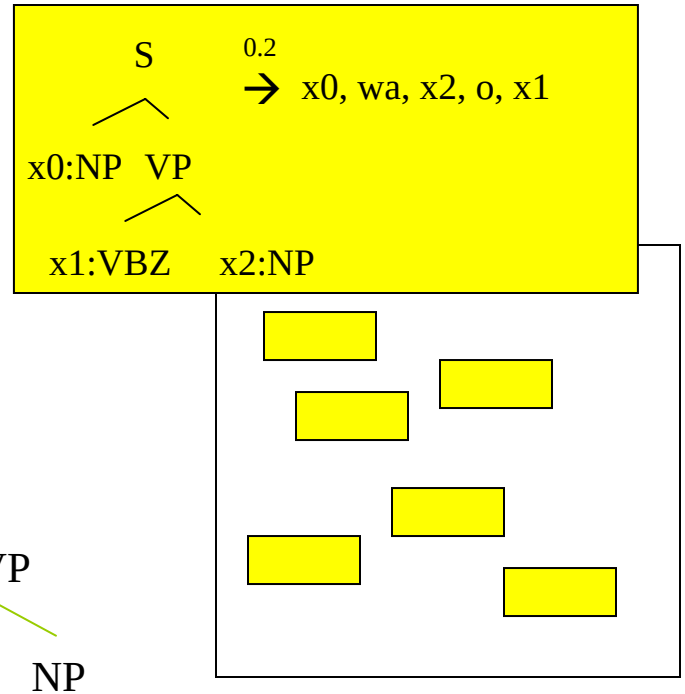
Transformation:



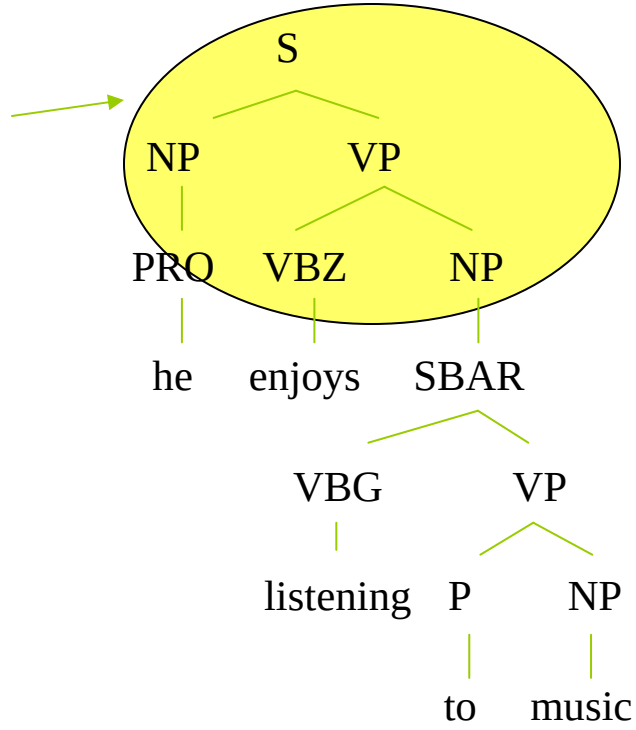
# Top-Down Tree-to-String Transducer

Original input:



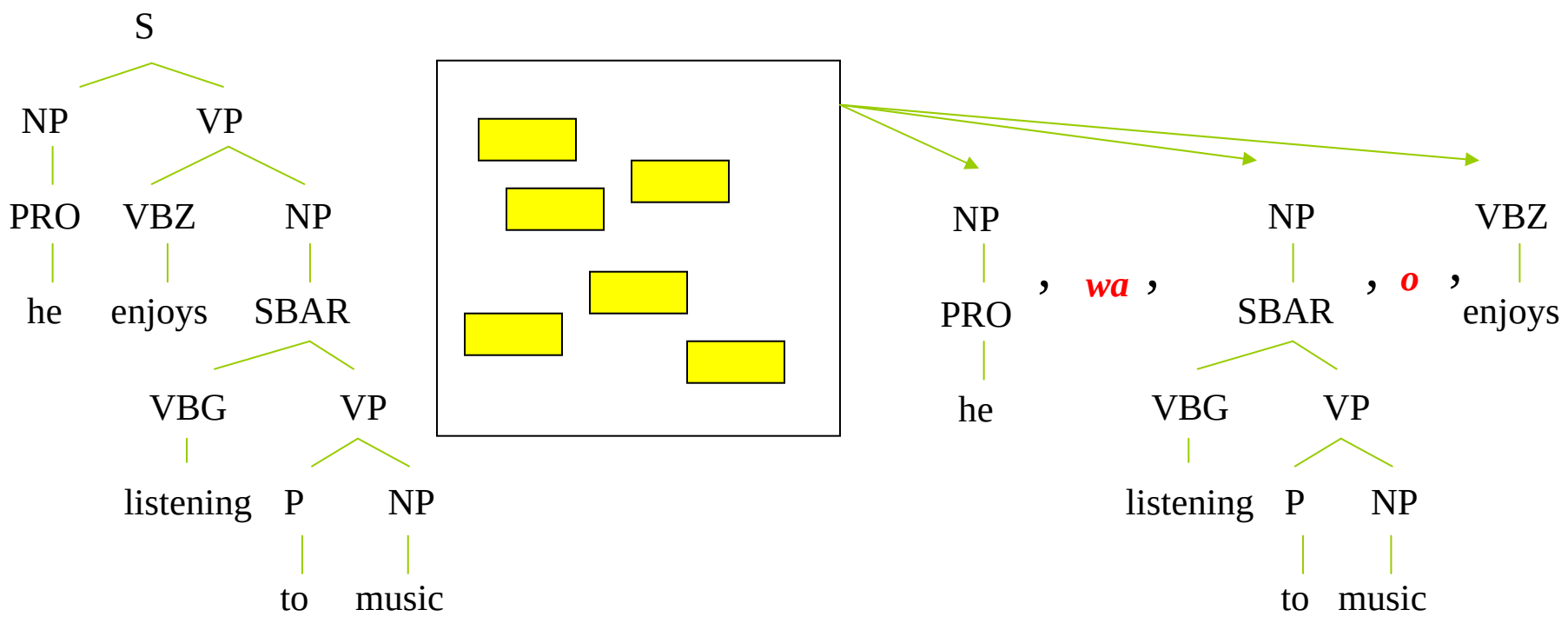




Transformation:

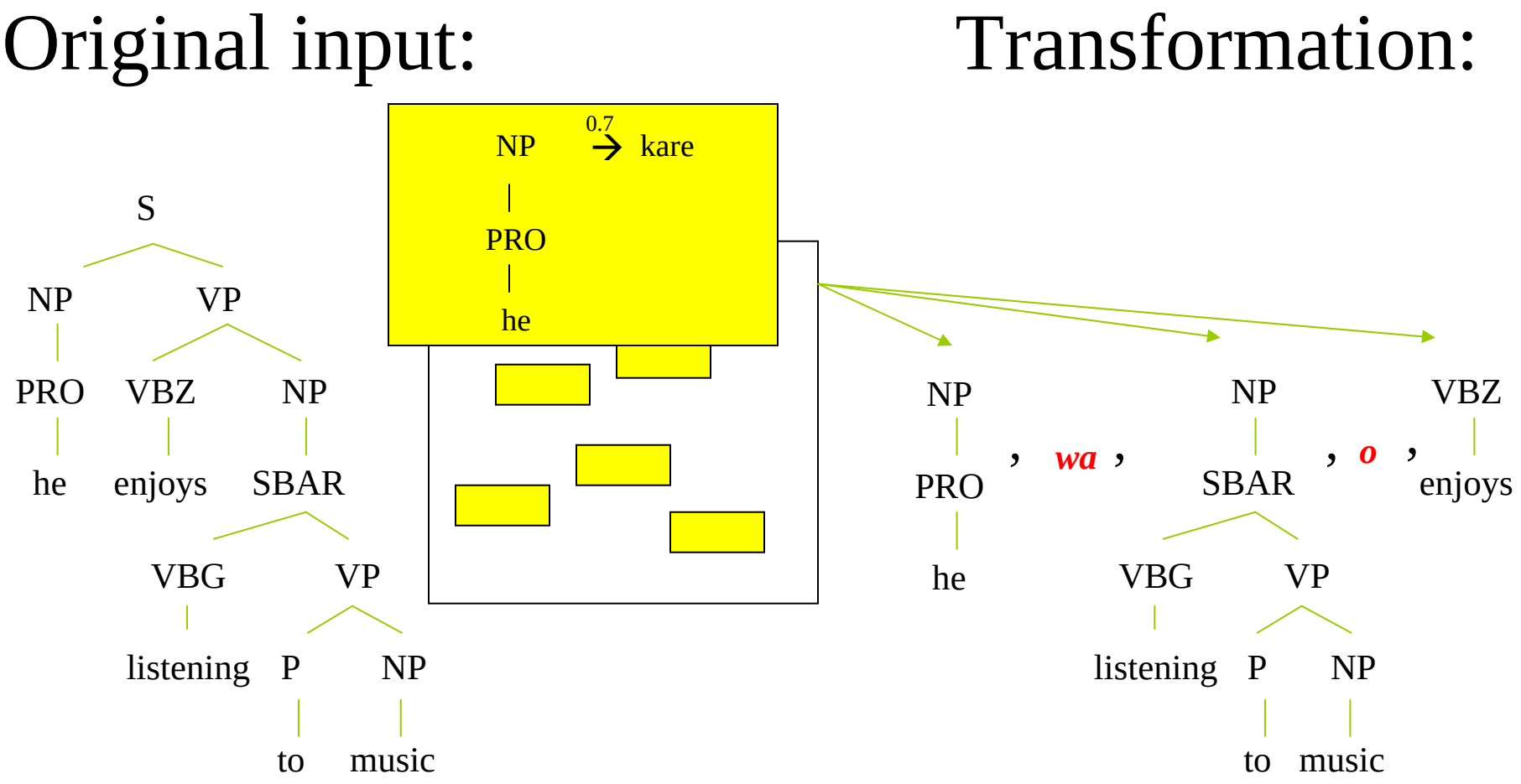


# Top-Down Tree-to-String Transducer

Original input: Transformation:

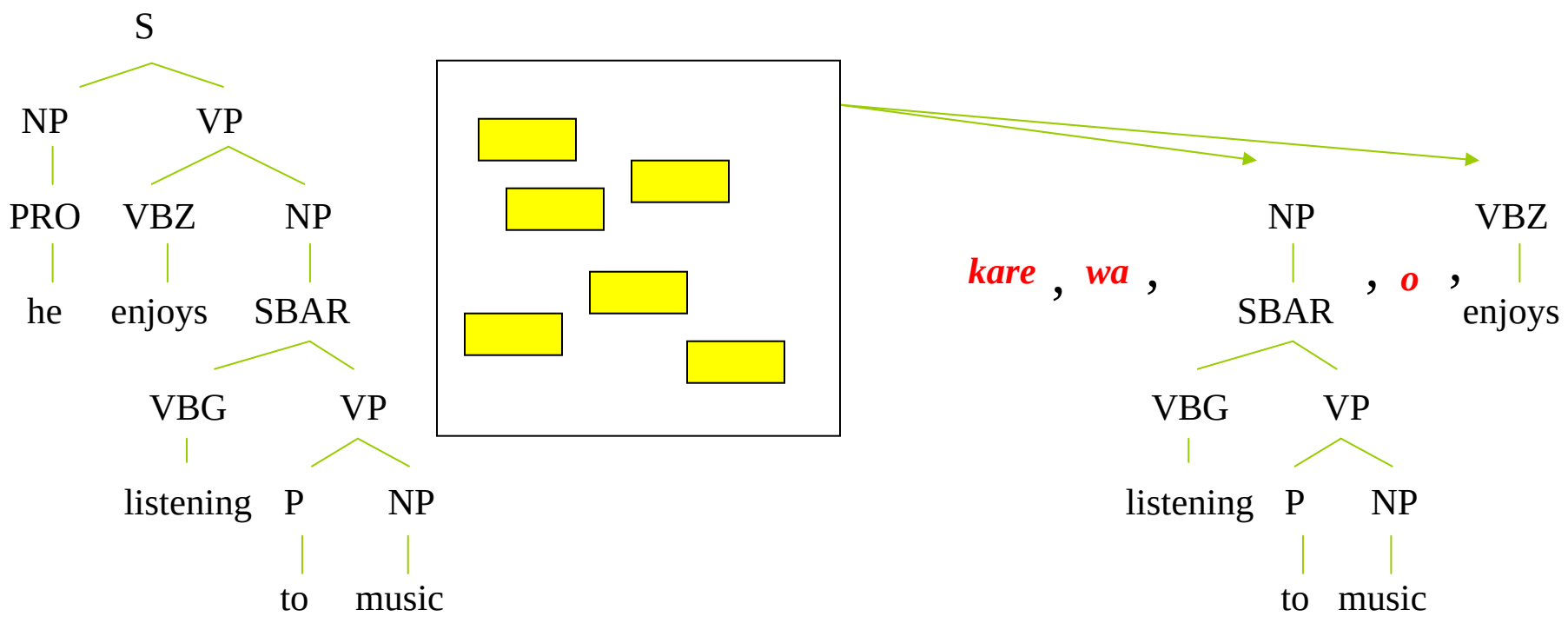


# Top-Down Tree-to-String Transducer



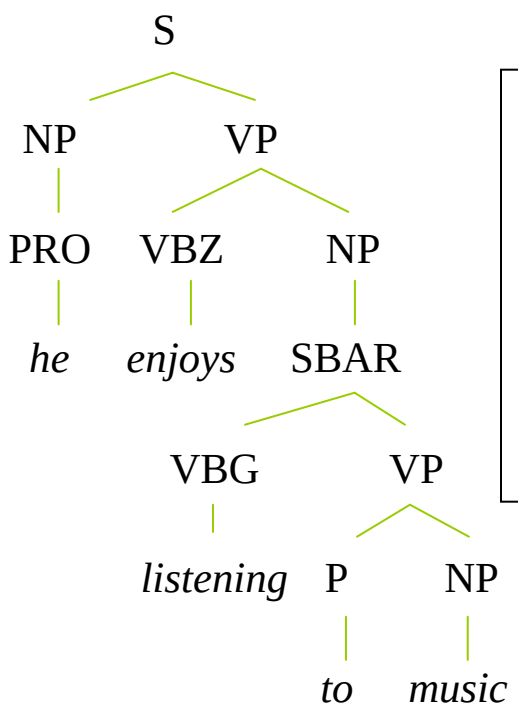
# Top-Down Tree-to-String Transducer

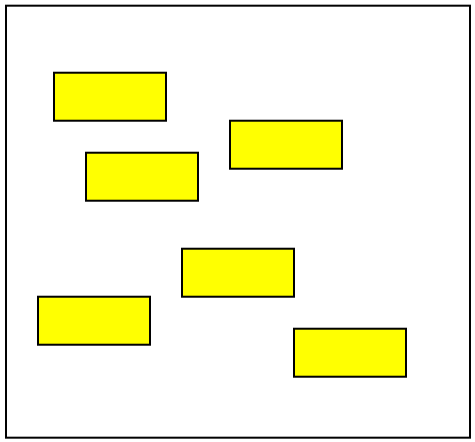
Original input: Transformation:



# Top-Down Tree-to-String Transducer

Original input:



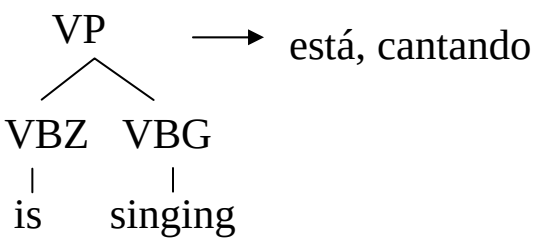
Final output:

***kare*** , ***wa*** , ***ongaku*** , ***o*** , ***kiku*** , ***no*** , ***ga*** , ***daisuki*** , ***desu***

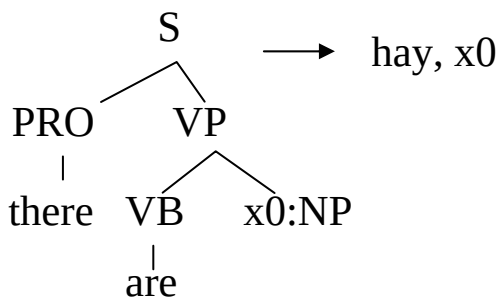
To get total probability, multiply probabilities of the individual steps.

# Transducer Format is Expressive

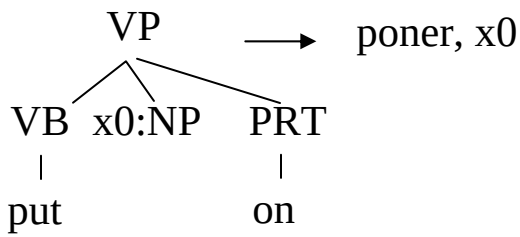
**Phrasal Translation**



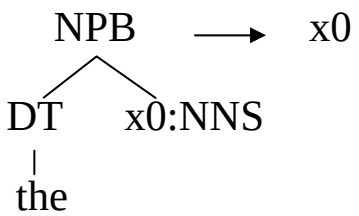

**Non-constituent Phrases**



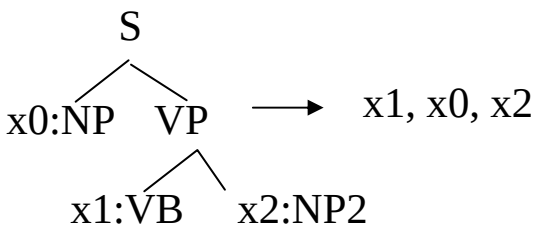

**Non-contiguous Phrases**



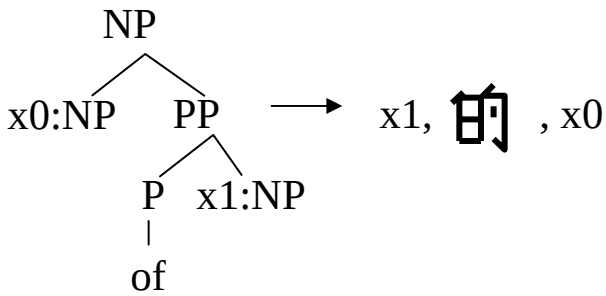

**Context-Sensitive Word Insertion**



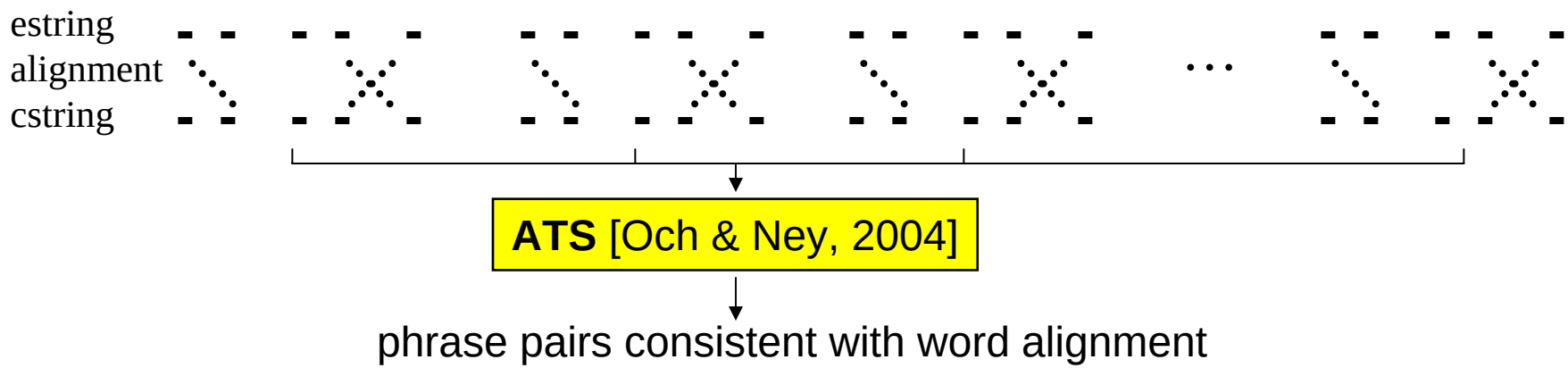

**Multilevel Re-Ordering**



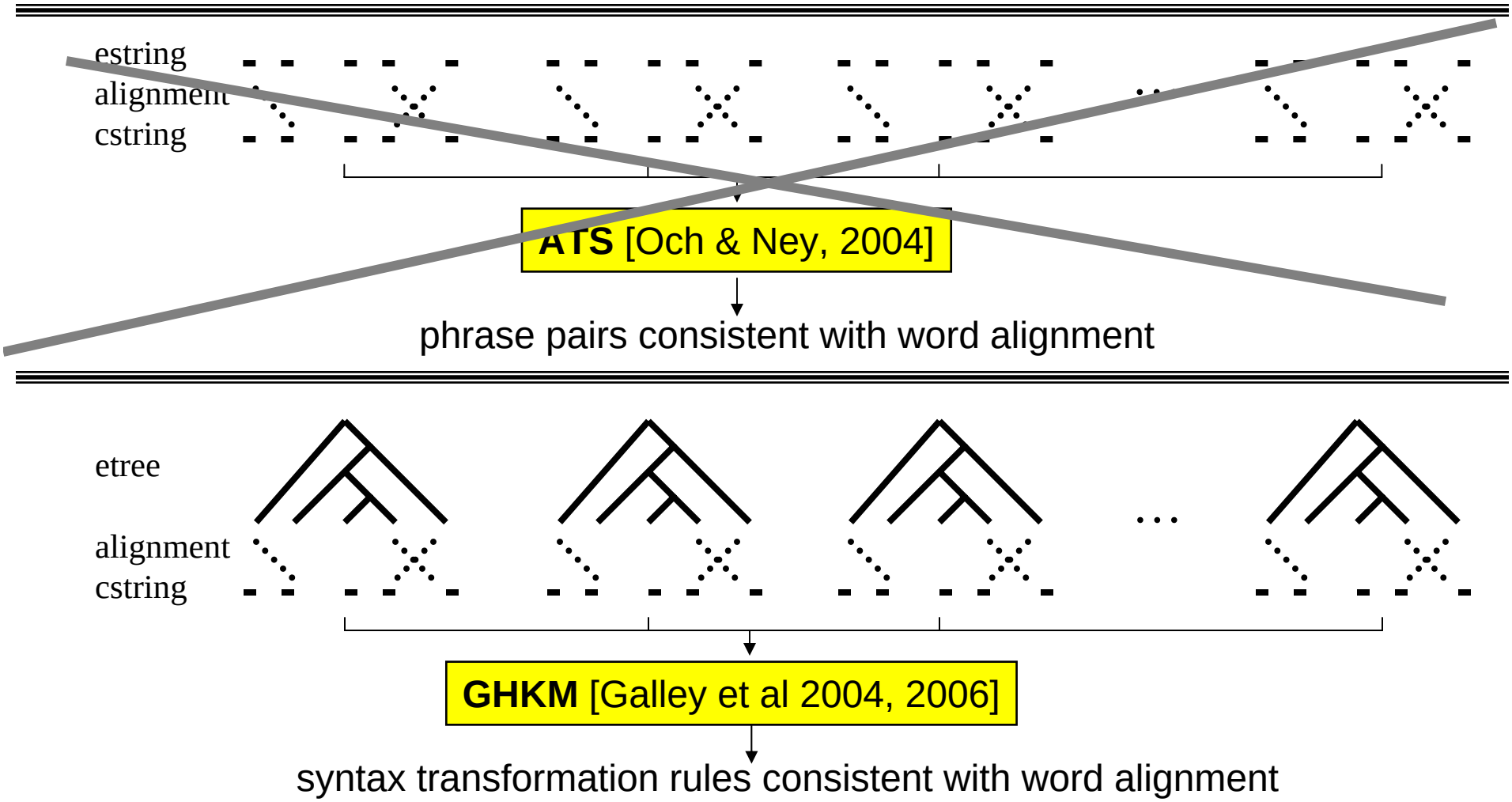

**Lexicalized Re-Ordering**



# Phrase-Based and Syntax-Based Pattern Extraction

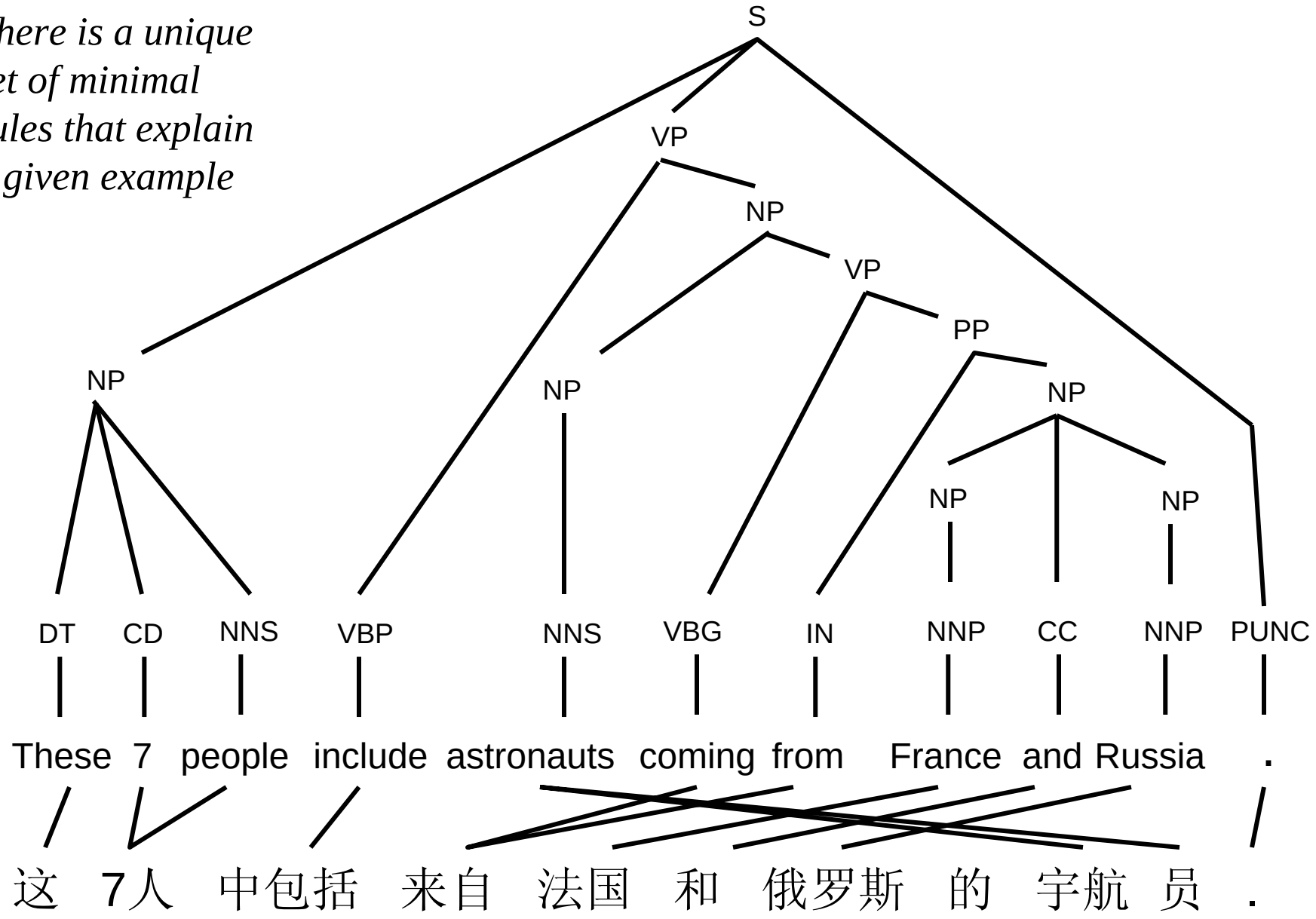


# Phrase-Based and Syntax-Based Pattern Extraction



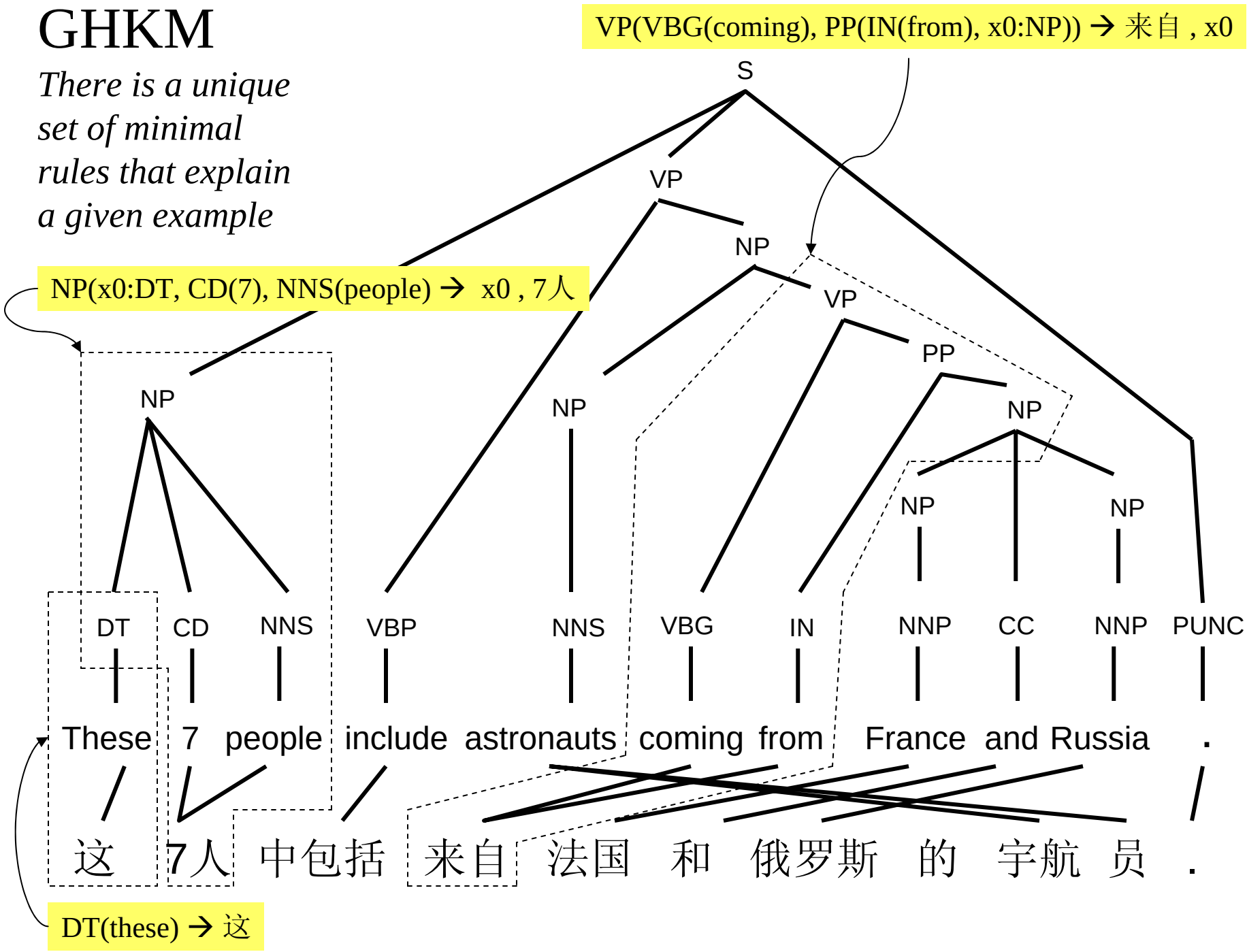
# GHKM

*There is a unique set of minimal rules that explain a given example*



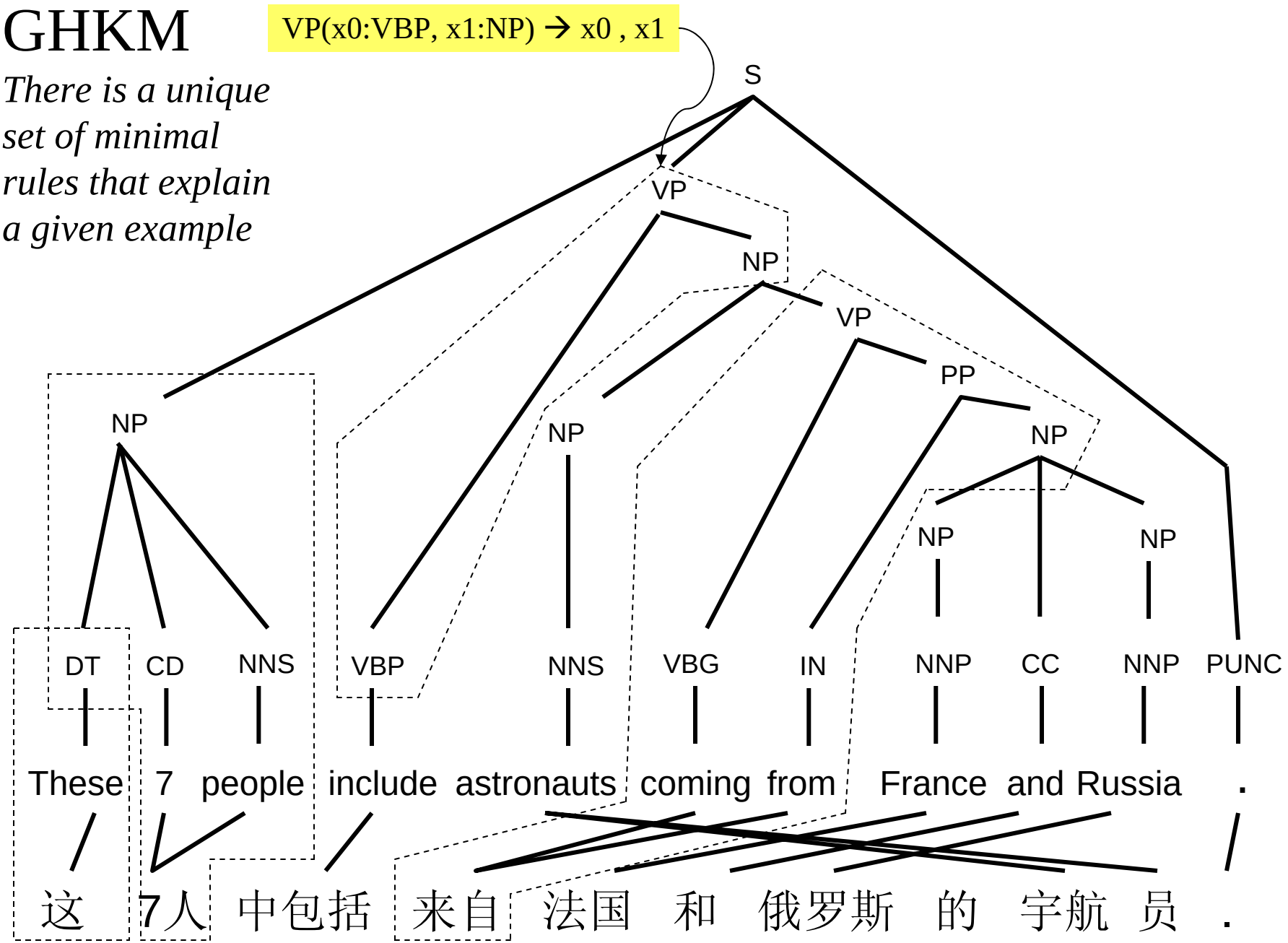
# GHKM

*There is a unique set of minimal rules that explain a given example*



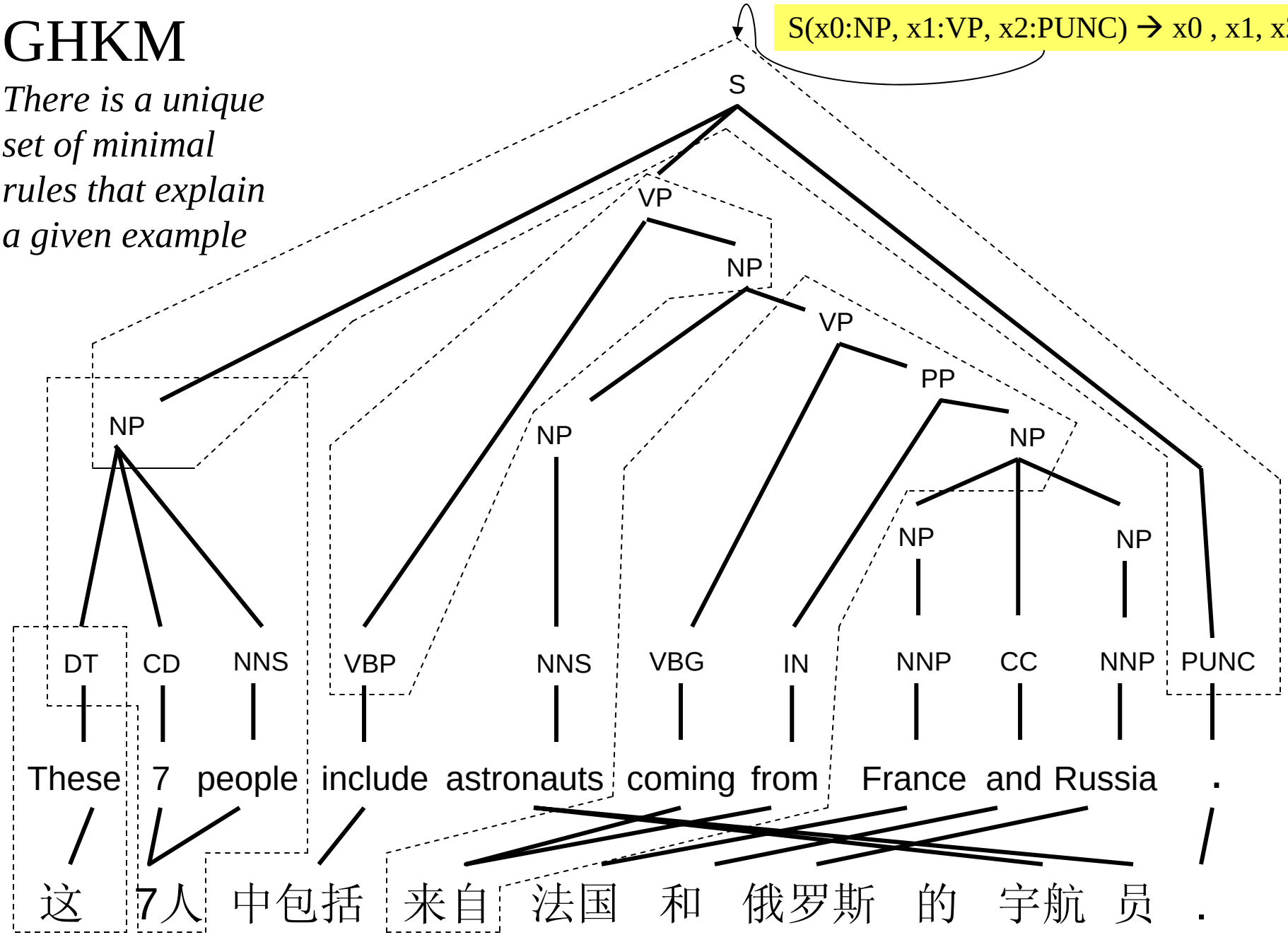
# GHKM

*There is a unique set of minimal rules that explain a given example*

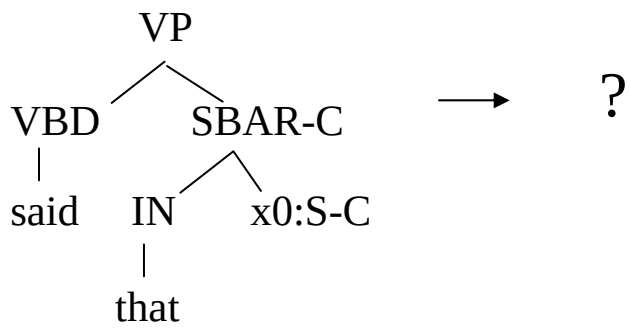


# GHKM

*There is a unique set of minimal rules that explain a given example*

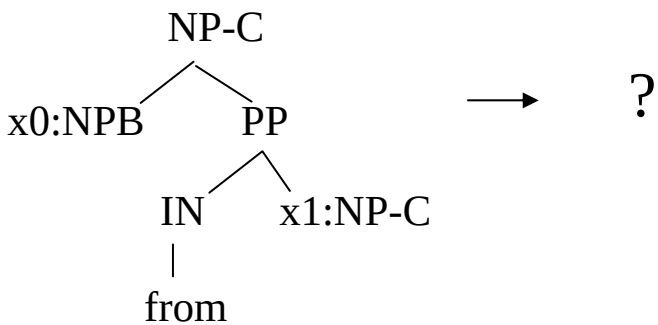


# Sample “said that” rules



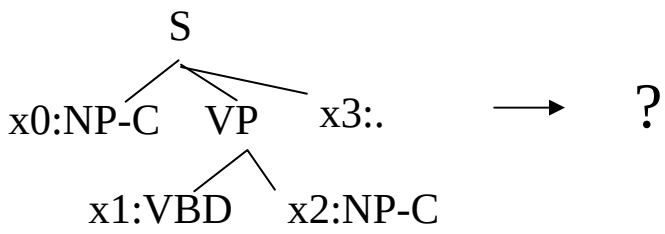

| | |
|---|---|
| 0.57 | VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> "说" "," x0 |
| 0.09 | VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> "说" x0 |
| 0.02 | VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> "他" "说" "," x0 |
| 0.02 | VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> "指出" "," x0 |
| 0.02 | VP(VBD("said")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




| | |
|---|---|
| 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




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

# Language Models

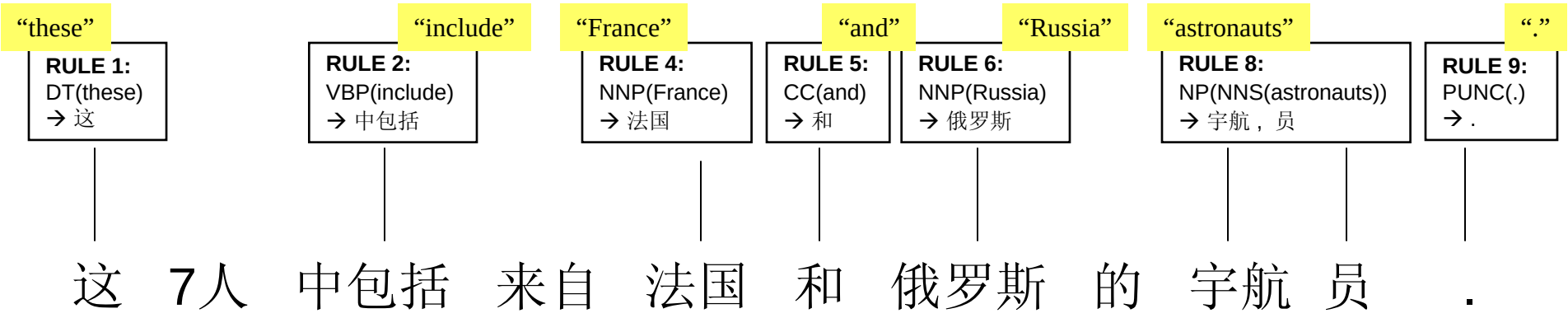
- Syntax-based Language Model
  - Assigns P(tree)  [Collins, 1997; Charniak, 2001]
  - Unlike parser, must be trained on domain data
  - Still unproven!
- N-gram Language Model
  - Standard trigram model
  - “Only judge a tree by its leaves”
  - Used in current syntax-based MT systems

# Decoder

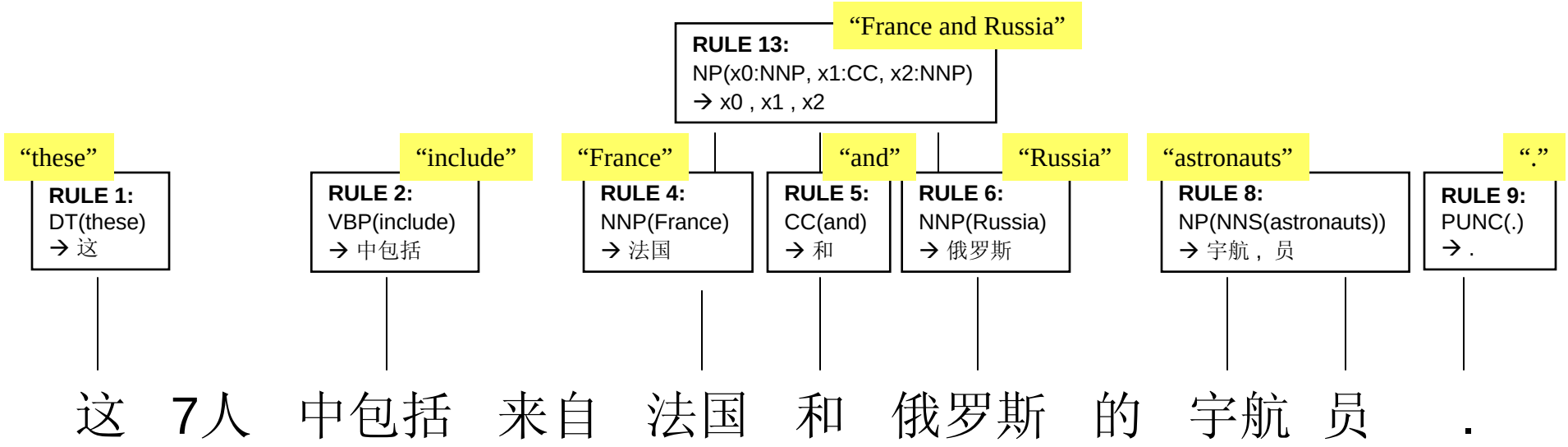
- Bottom-up CKY parser
- Builds English constituents on top of Chinese spans
- Record of rule applications (the derivation) provides information to construct English tree
- Returns k-best trees

# Syntax-Based Decoding

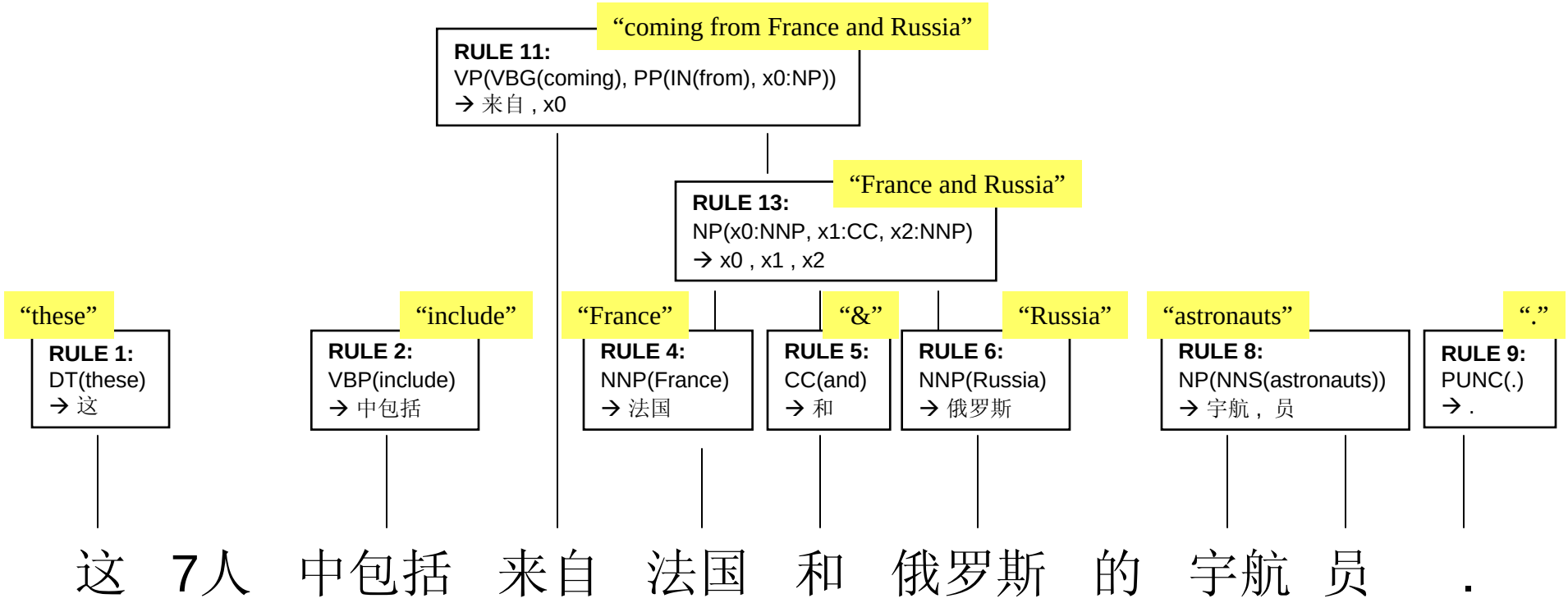
Rules apply when their right-hand sides (RHS) match some portion of the input.



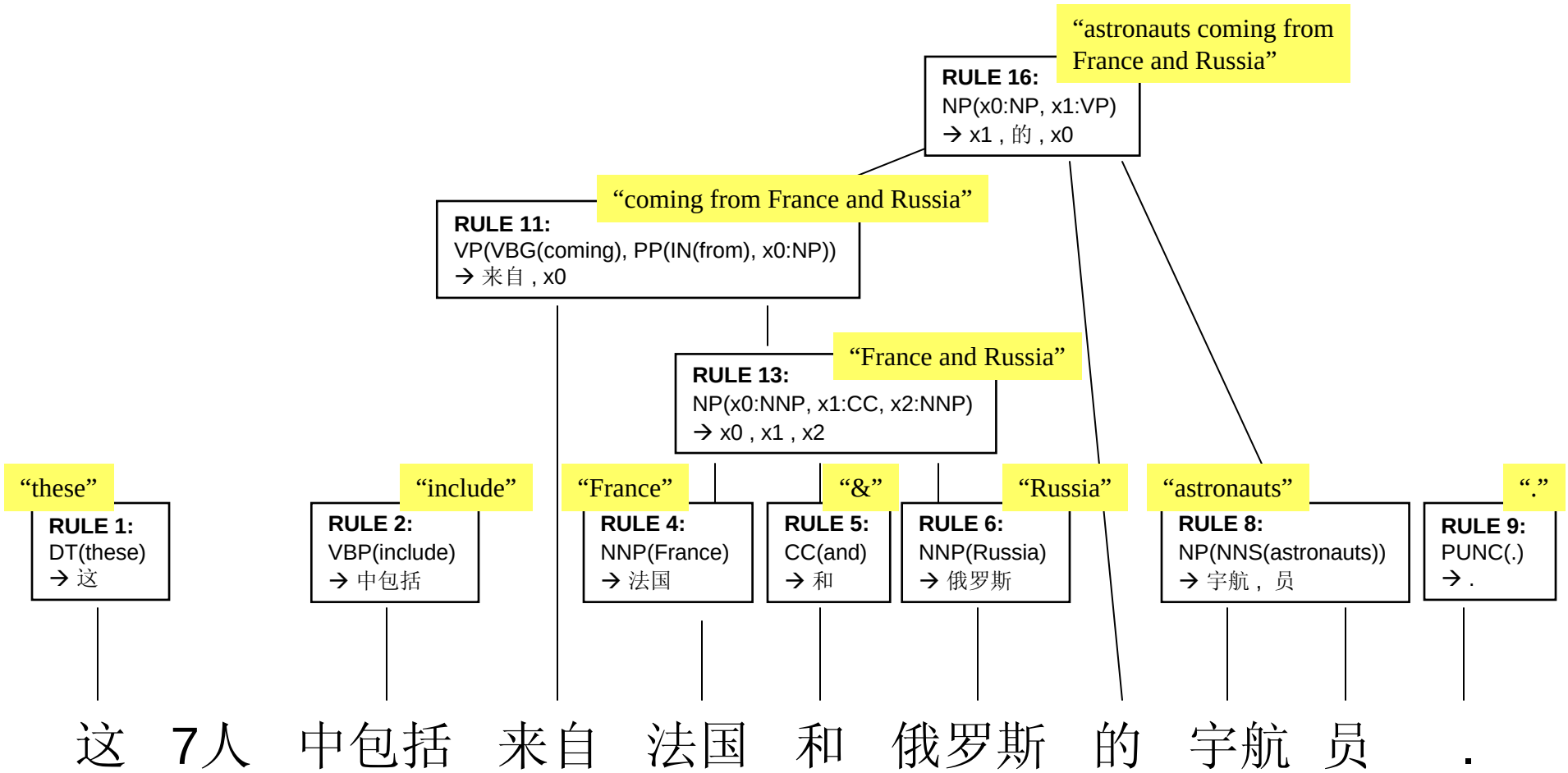
# Syntax-Based Decoding

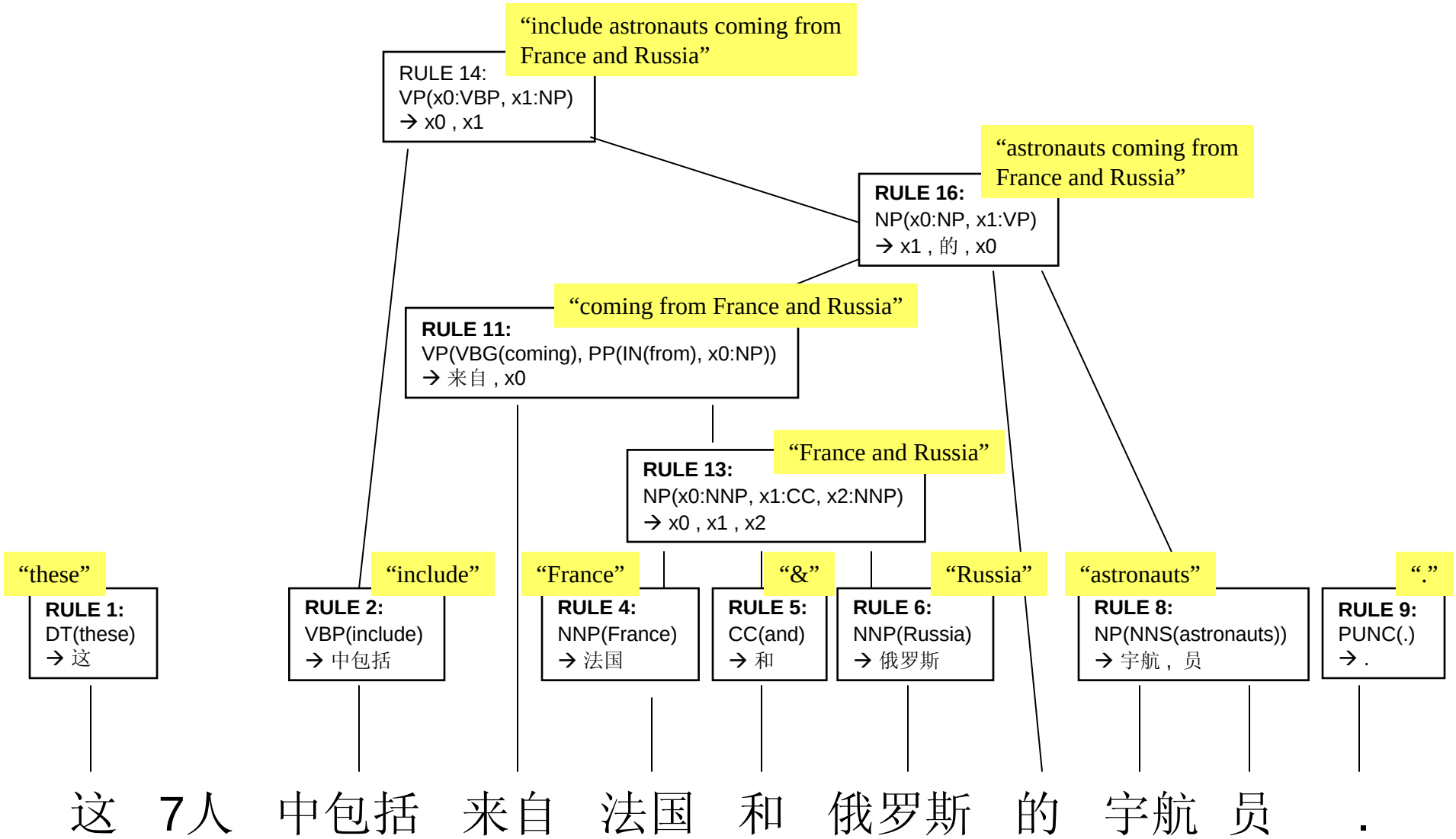


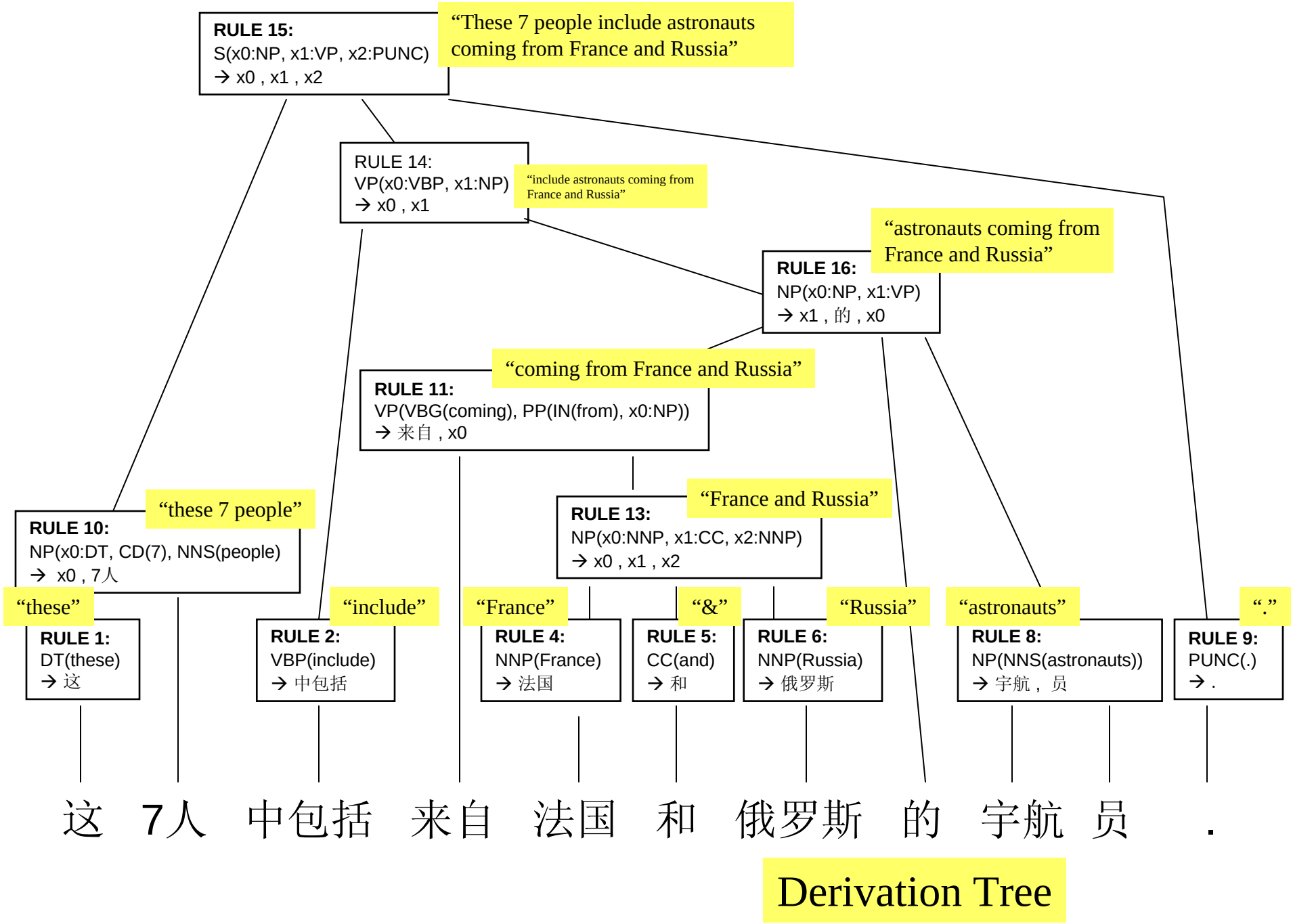
# Syntax-Based Decoding



# Syntax-Based Decoding

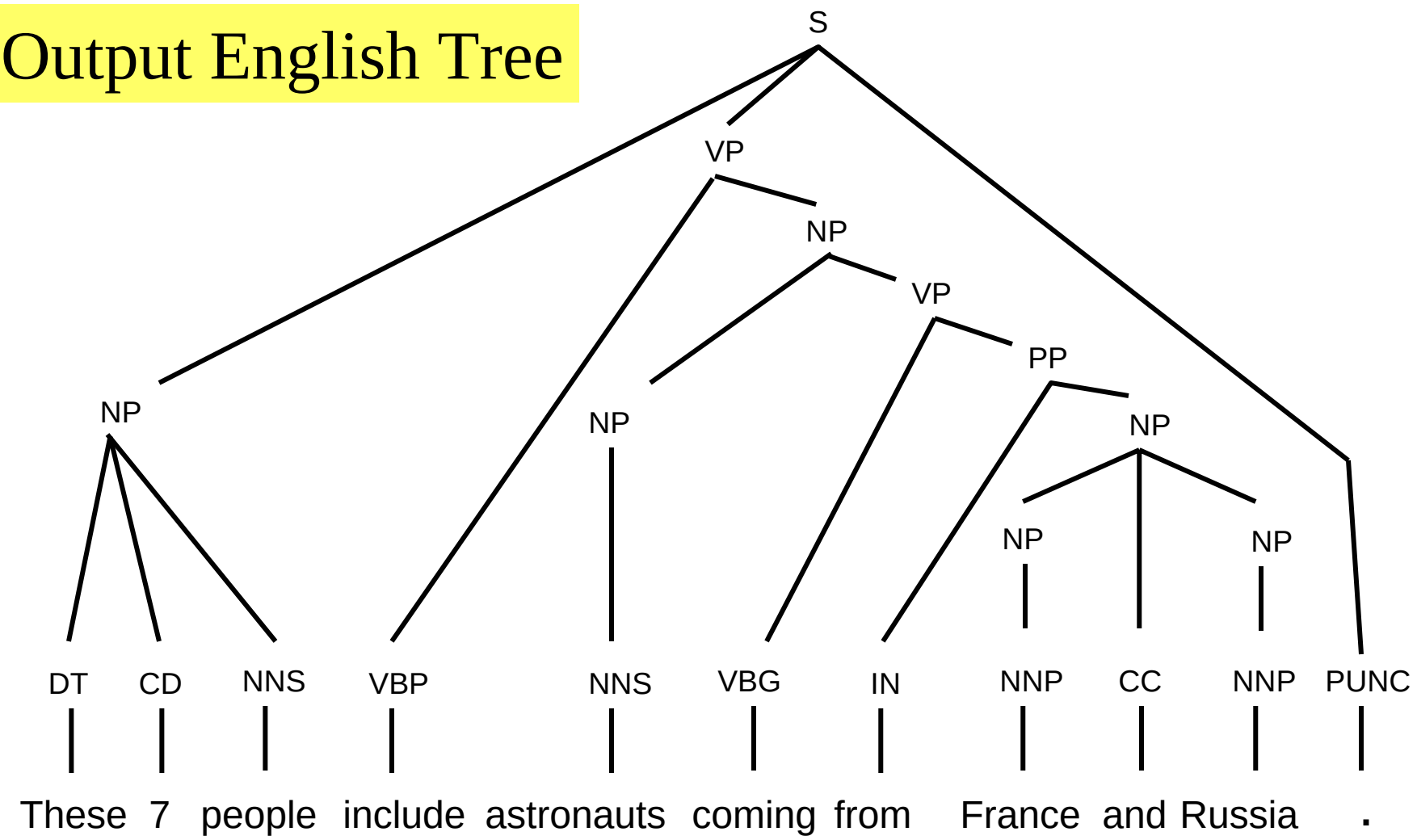



"include astronauts coming from France and Russia"
RULE 14:
VP(x0:VBP, x1:NP)
→ x0 , x1
"astronauts coming from France and Russia"
RULE 16:
NP(x0:NP, x1:VP)
→ x1 , 的 , x0
"coming from France and Russia"
RULE 11:
VP(VBG(coming), PP(IN(from), x0:NP))
→ 来自 , x0
"France and Russia"
RULE 13:
NP(x0:NNP, x1:CC, 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 .




Derivation Tree

# Output English Tree



# Empirical Questions

- With the acquired rules:
  - Can you always reach a top S for new Chinese sentences?
  - Does reaching a top S result in overall grammaticality?
  - Is it only possible to reach a top S by changing the meaning?
  - Is the overall translation accuracy good?

## 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 $_7$ | 。 $_8$

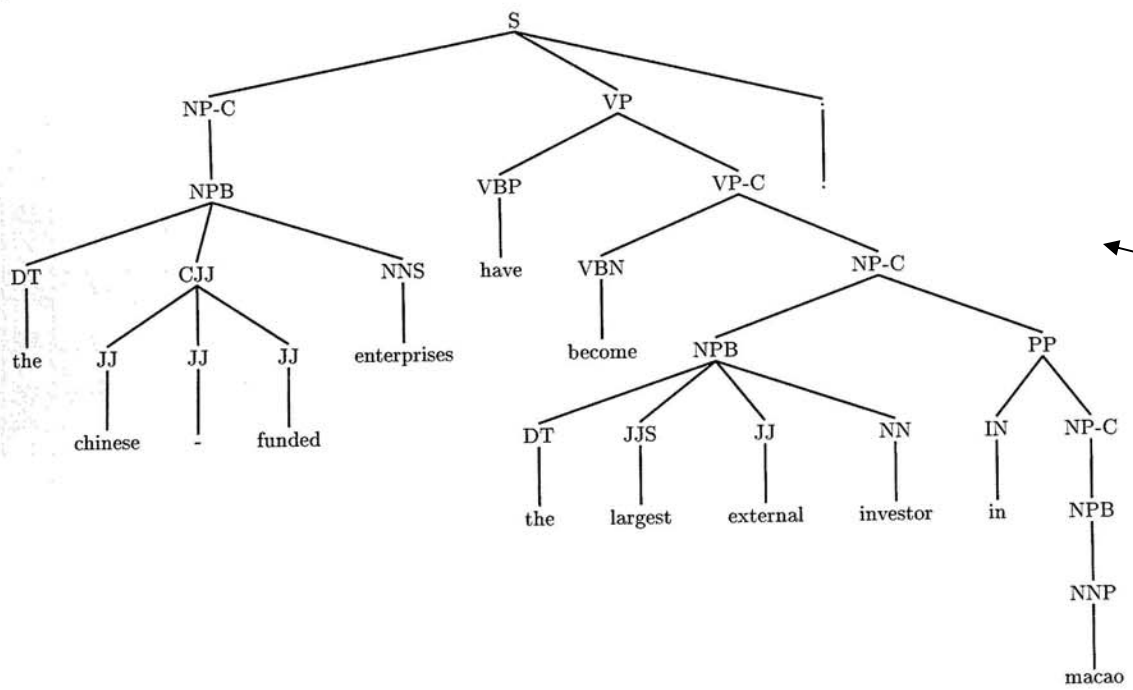
**[dev-little] 1-Best:** the chinese - funded enterprises have become the largest external investor in macao .

input

phrase-based system output

syntax-based system output

### [dev-little] 1-Best Tree



## 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 $_7$ | . $_8$

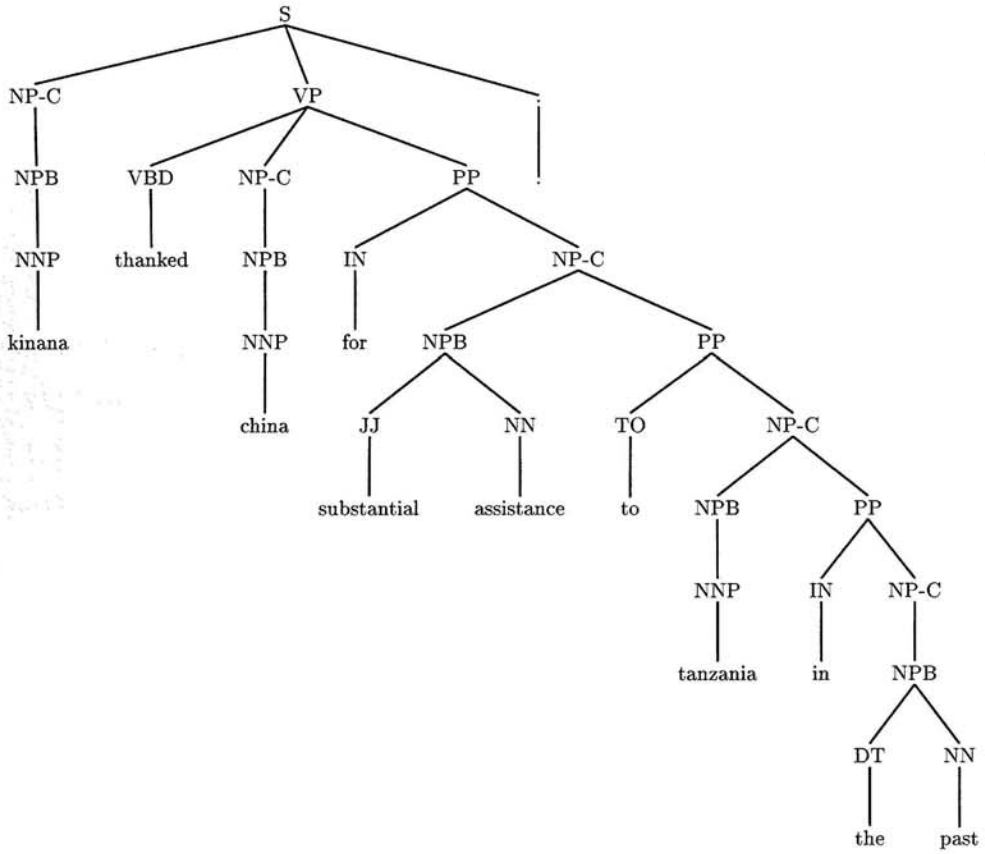
**AlTemp-f:** 基 纳 纳 $_0$ | 对 $_1$ | 中国 $_2$ | 过去 向 $_3$ | 坦 桑 $_4$ | 提供 的 大量 $_6$ | 援助 $_7$ | 表示 感谢 $_5$ | 。 $_8$

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

input

phrase-based system output

### [dev-little] 1-Best Tree

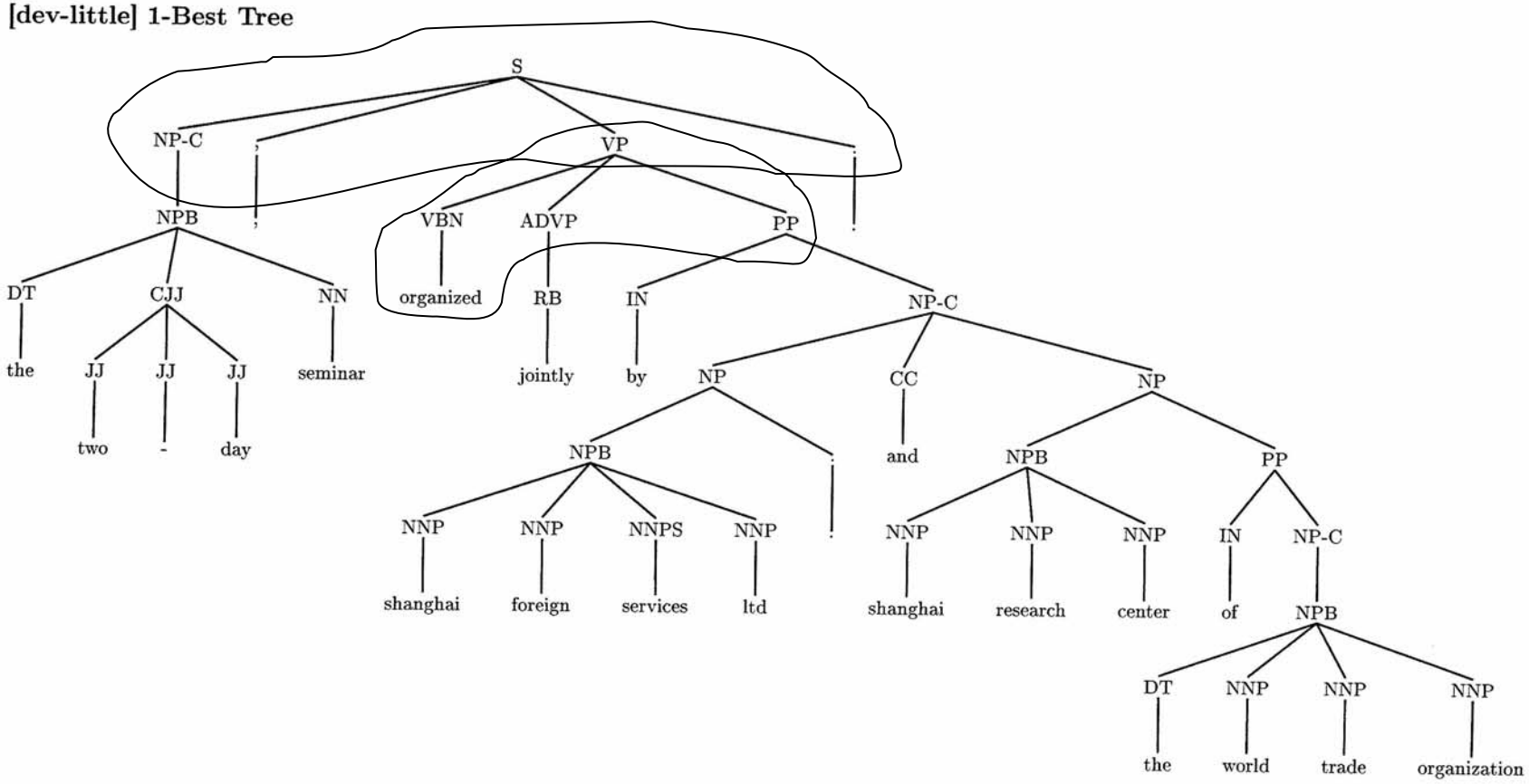



syntax-based system output

## 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 .

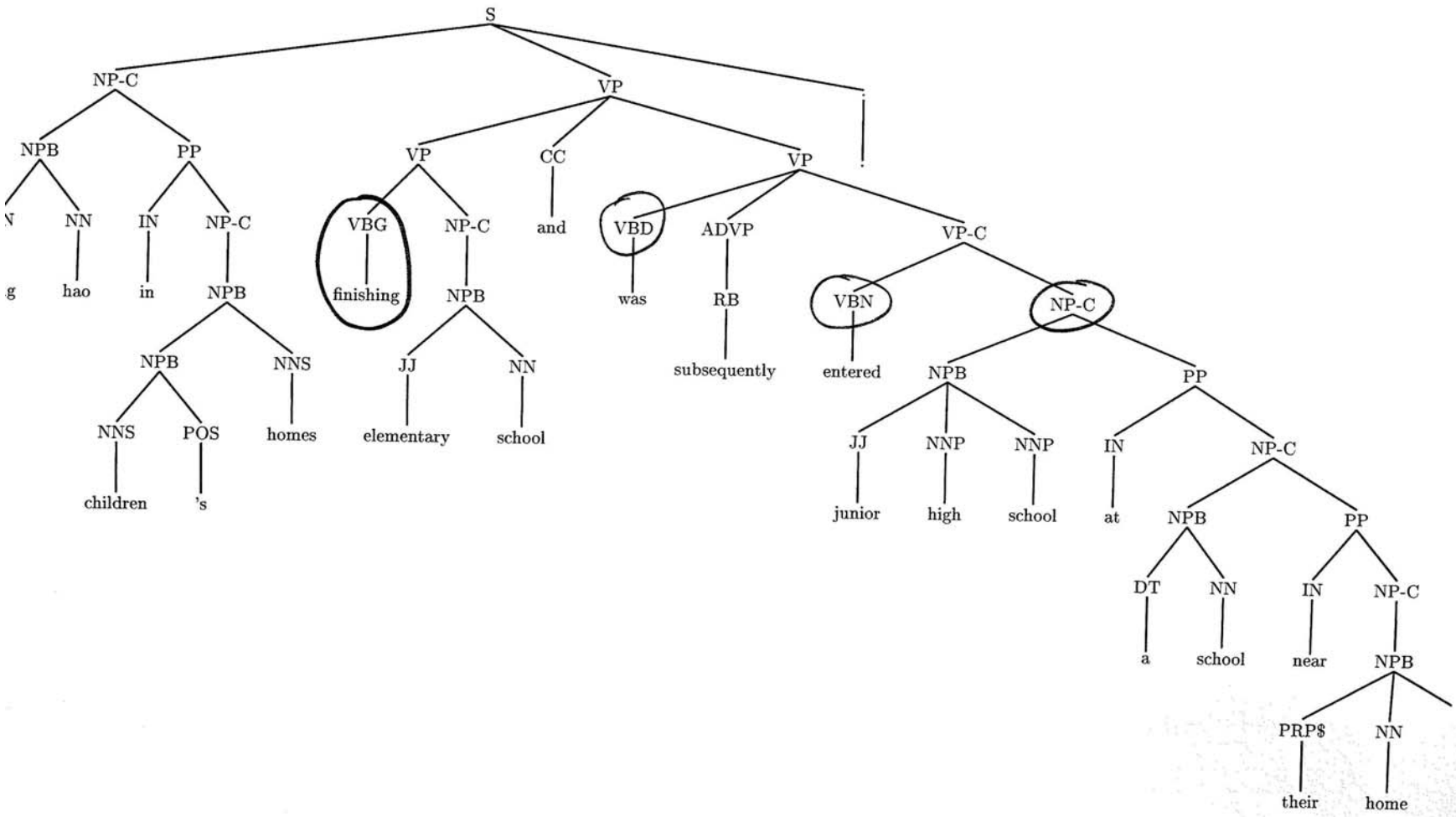
### [dev-little] 1-Best Tree



## lev-little (line 53) - dev-little

**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

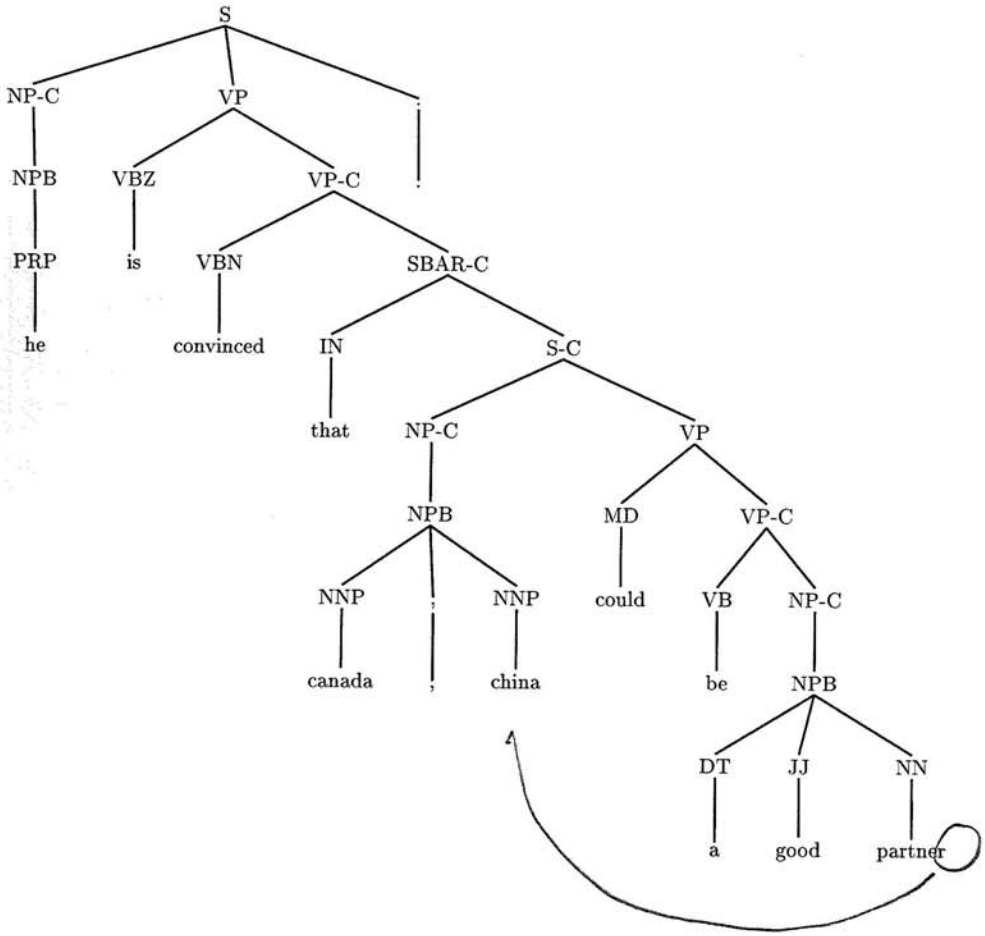


## dev-little (line 64) - dev-little

can become very good partners

**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

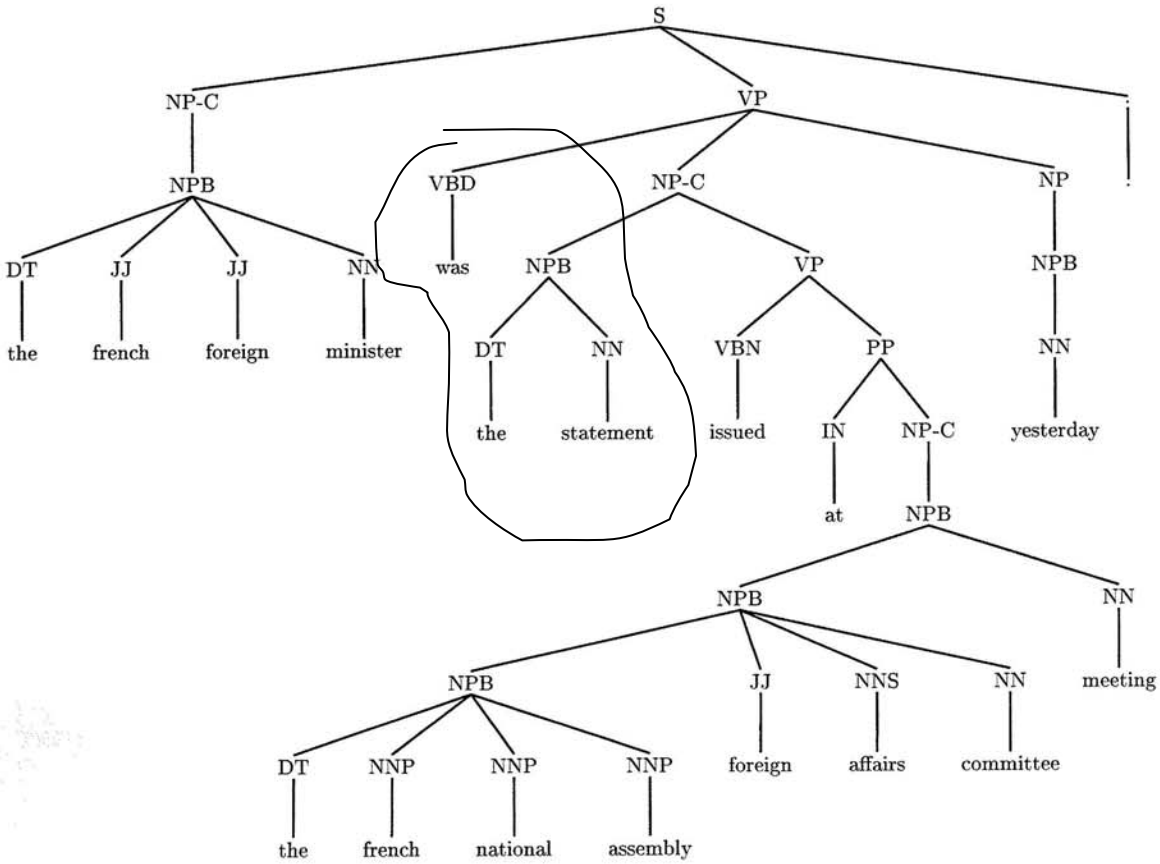



Subj - obj
number agreement.

## 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



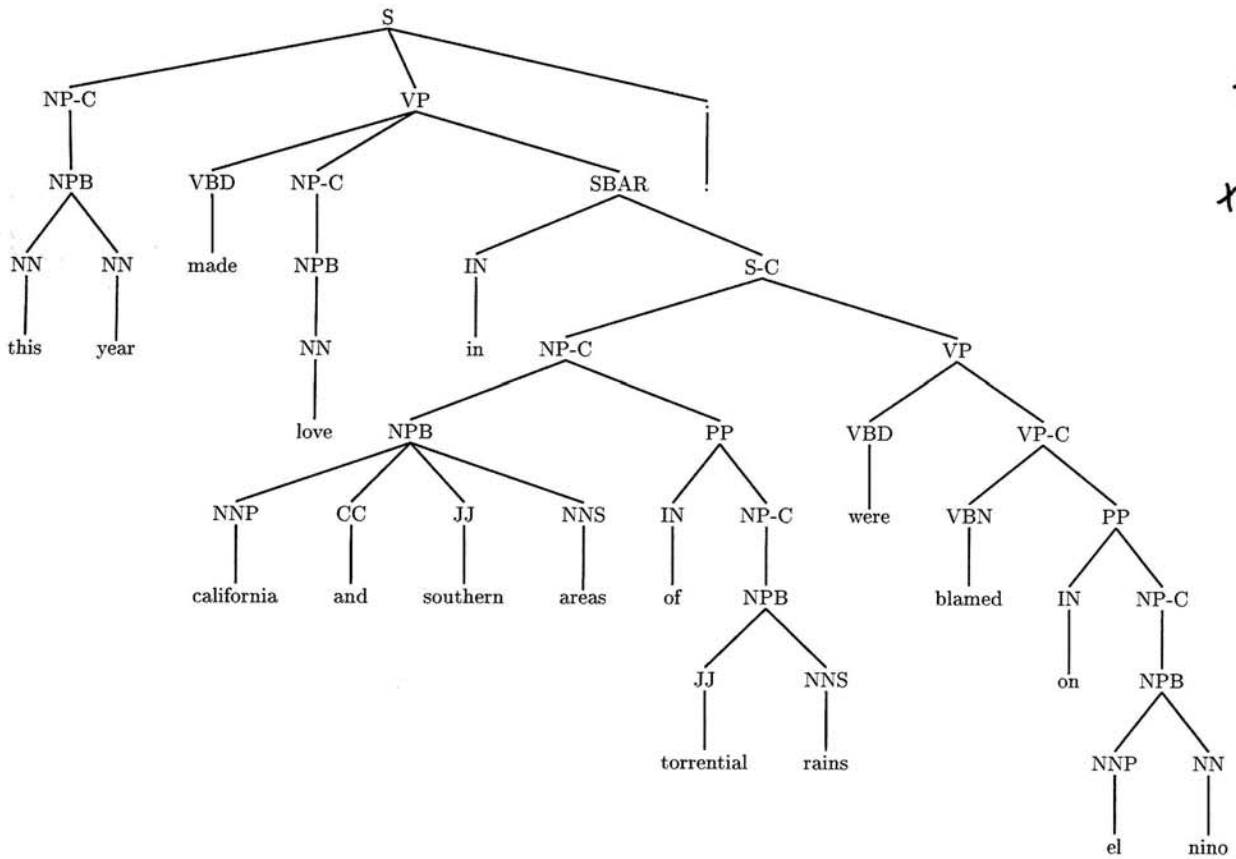
## dev-little (line 125) - dev-little

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 .

Scope.

### [dev-little] 1-Best Tree




funny.

this year rarely NP-C

# Lots of Open Problems

- Specific to MT:
  - Choosing syntactic categories that are appropriate for translation
  - Decoder search errors
  - More context for rule choice
  - Syntax-based language models
- For of NLP and beyond:
  - Modeling with tree transducers
  - Algorithms for tree transducers
  - Generic software toolkits for tree transducers

# Tiburon: A Tree Automata Toolkit

- Developed by Jonathan May, USC/ISI
- First version distributed in April (www.isi.edu...)

- You cast your problem in terms of tree acceptors and transducers
- You get implemented algorithms for free
  - Kumar/Byrne'03 do this for phrase-based MT
  - Pereira/Riley'96 do this for ASR

- Wealth of tree automata literature to drawn on
- Still lots of open problems in tree automata and in choosing formalisms for modeling NLP

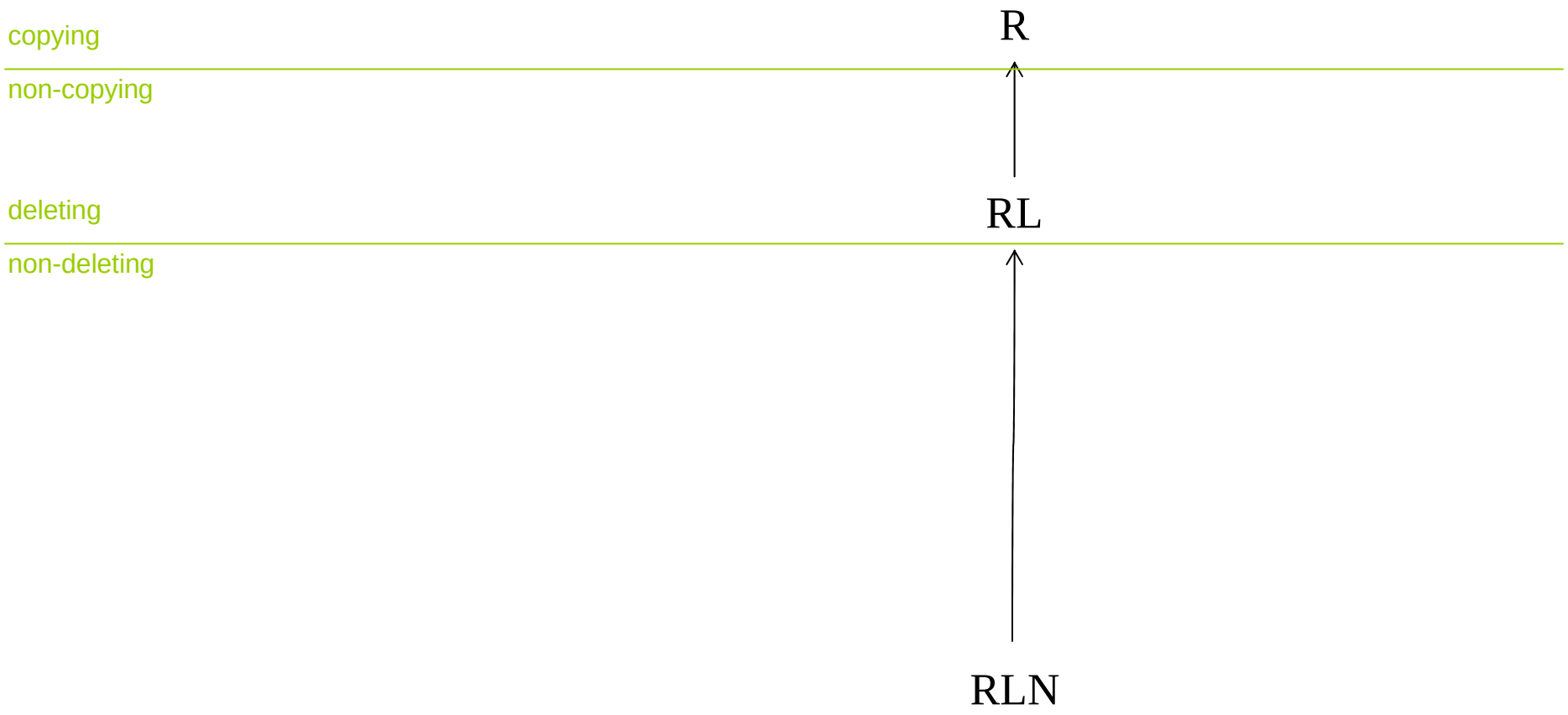
# Tiburon: A Tree Automata Toolkit

| | String World | Tree World |
|---|---|---|
| Weighted Sets | String acceptors (WFSA) | Tree acceptors |
| Weighted Transformations | String transducers (WFST) | Tree transducers |

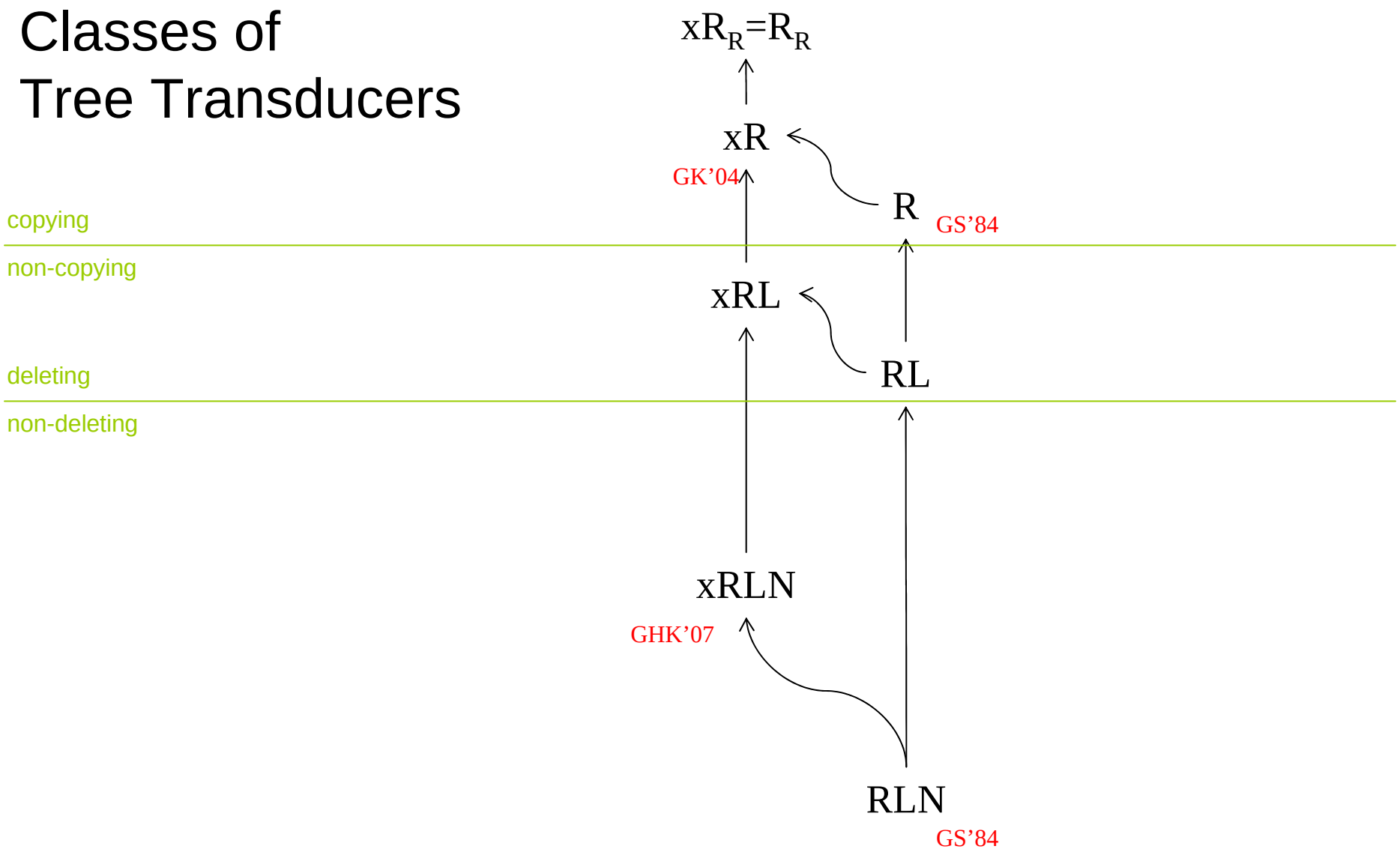
# Tiburon: A Tree Automata Toolkit

| | 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, 1970; Engelfriet, 1975; Graehl, Hopkins, Knight |

# Classes of Tree Transducers

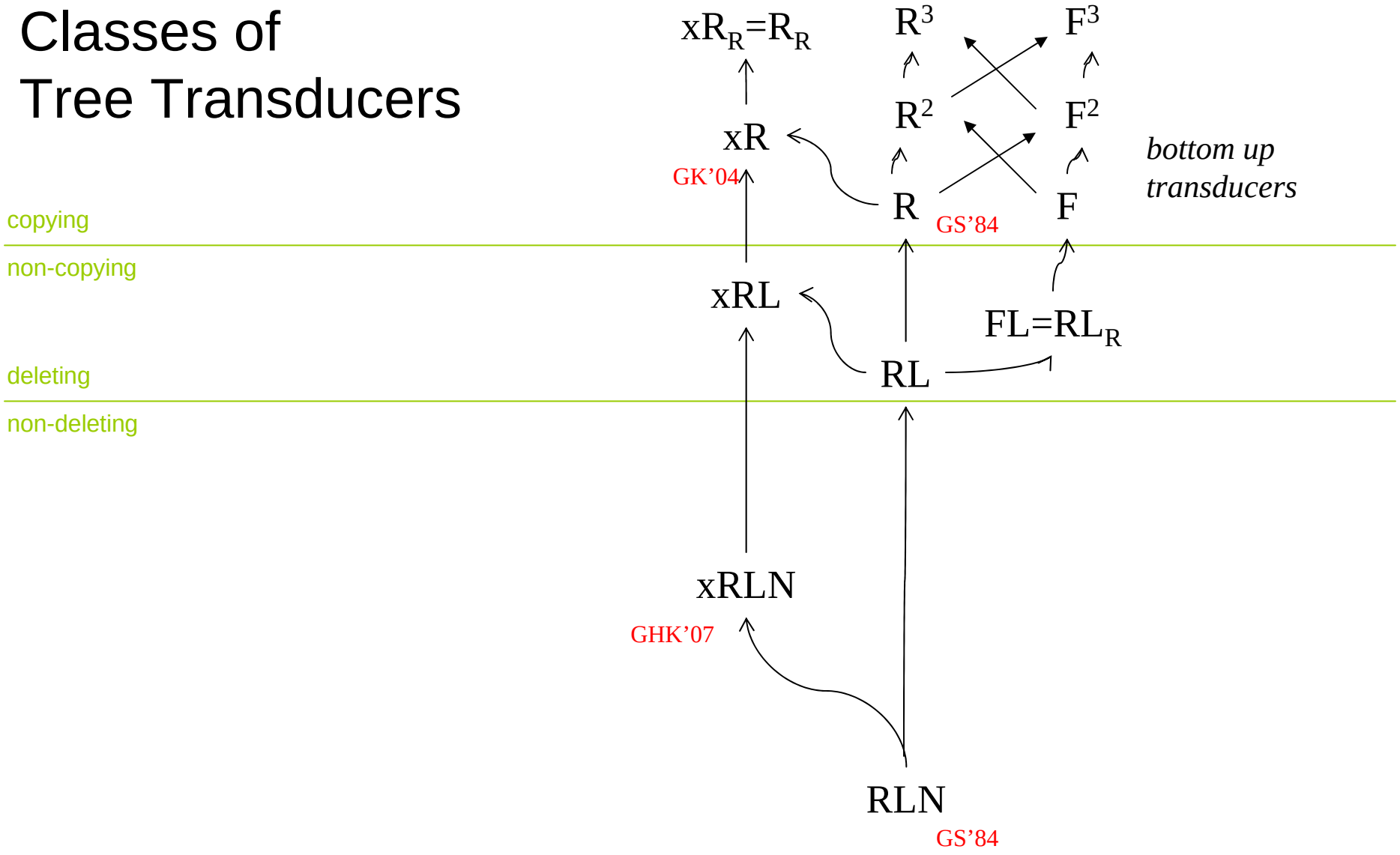


# Classes of Tree Transducers

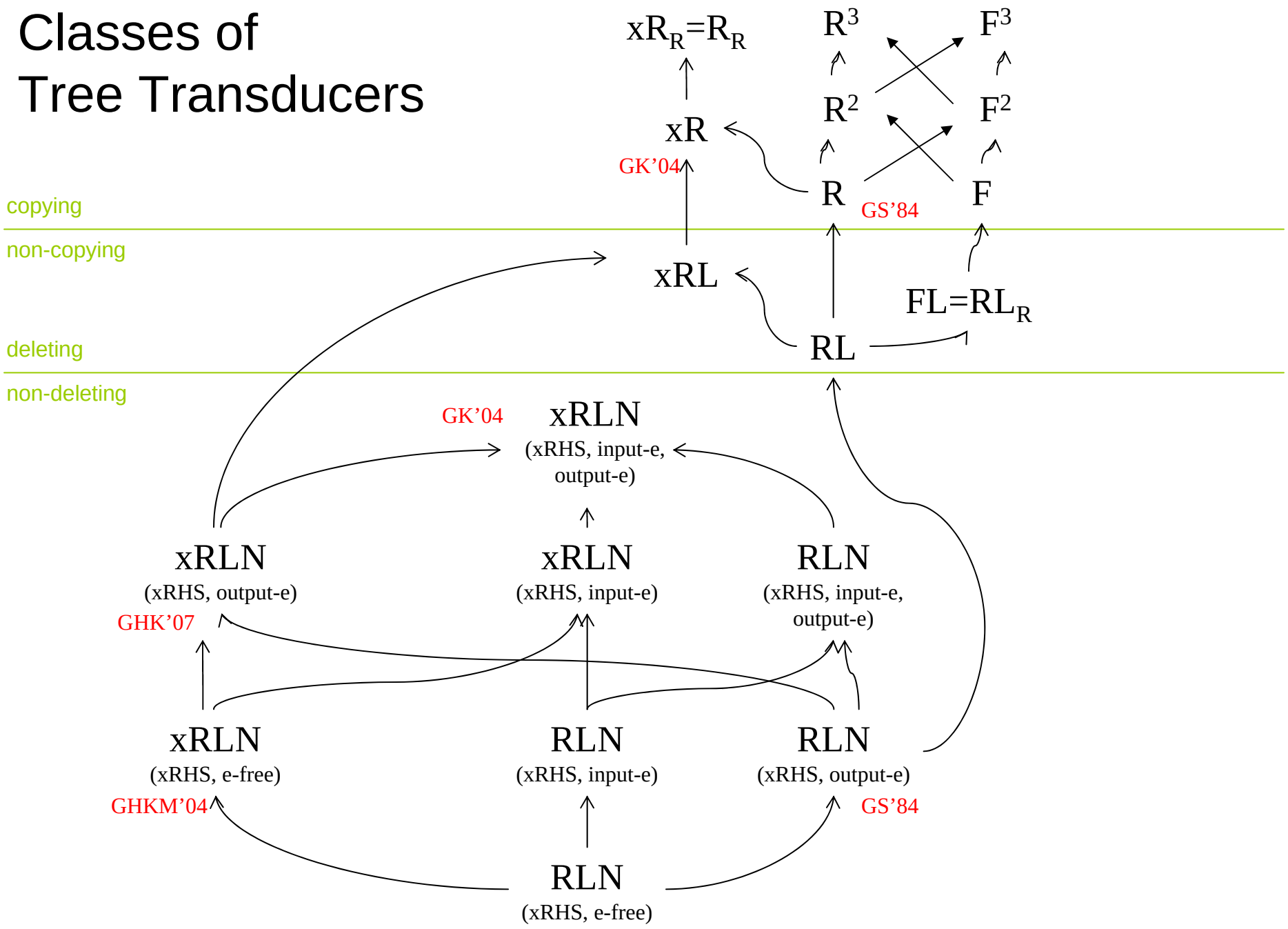



Expressive power theorems in Graehl, Hopkins, Knight (submitted)

# Classes of Tree Transducers



# Classes of Tree Transducers



# Conclusion

- Making progress on machine translation
- Opening up field of tree automata to NLP
- Interdisciplinary Research
  - Machine Learning
  - Engineering
  - Linguistics
  - Efficient search algorithms
  - Automata theory
  - Grid computing

the ~~end~~

beginning!