integrated morphology for translation

kevin knight

university of southern california

MTMRL / Haifa / January 2011

Automatic Language Translation

Automatic Language Translation

is darn hard

Even for simple sentences…

Input:

贝尔 当场 死亡 , 他 的 两 个 朋友 受伤 .

Correct:

Bell died on the spot and his two friends were injured .

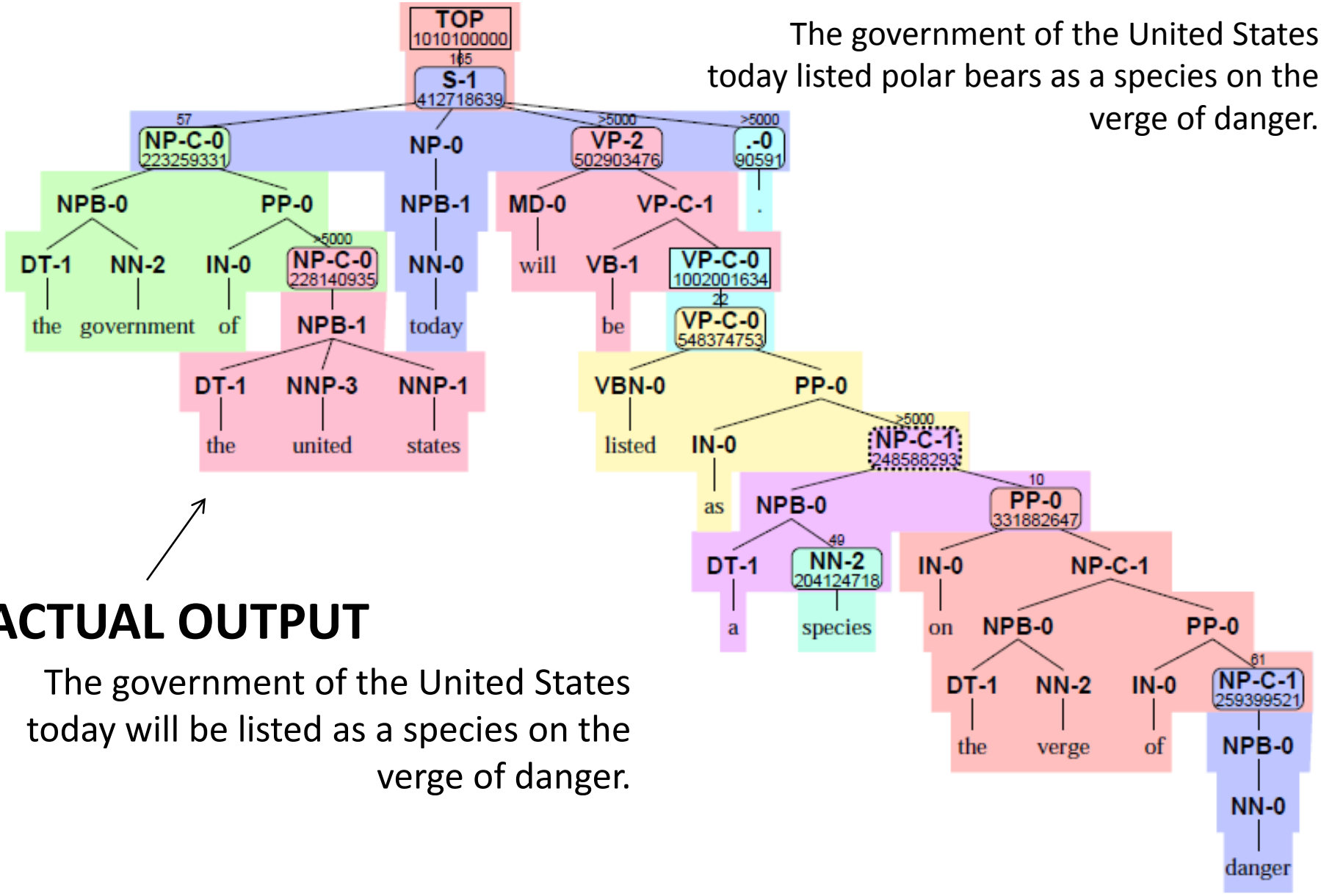
MT output:

Bell died on the spot and injured his two friends.

美国 政府 今天 将 北极熊 列为 濒临 危险 物种 .

DESIRED OUTPUT

The government of the United States today listed polar bears as a species on the verge of danger.



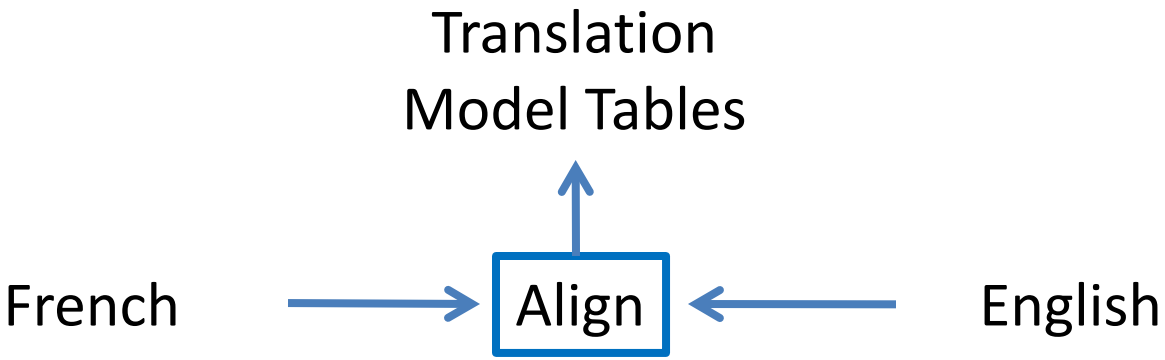
ACTUAL OUTPUT

The government of the United States today will be listed as a species on the verge of danger.

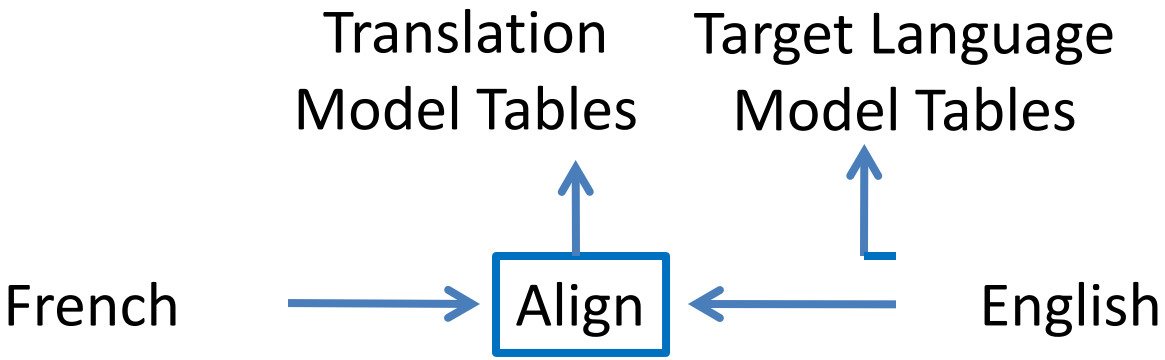
Need Lots of Knowledge

- What do words mean?
- What is the relation between meanings of words and the meaning of a sentence?
- What makes a sentence grammatical?
- **Why does the same word have so many forms?**
- **How do those forms play a role in translation?**

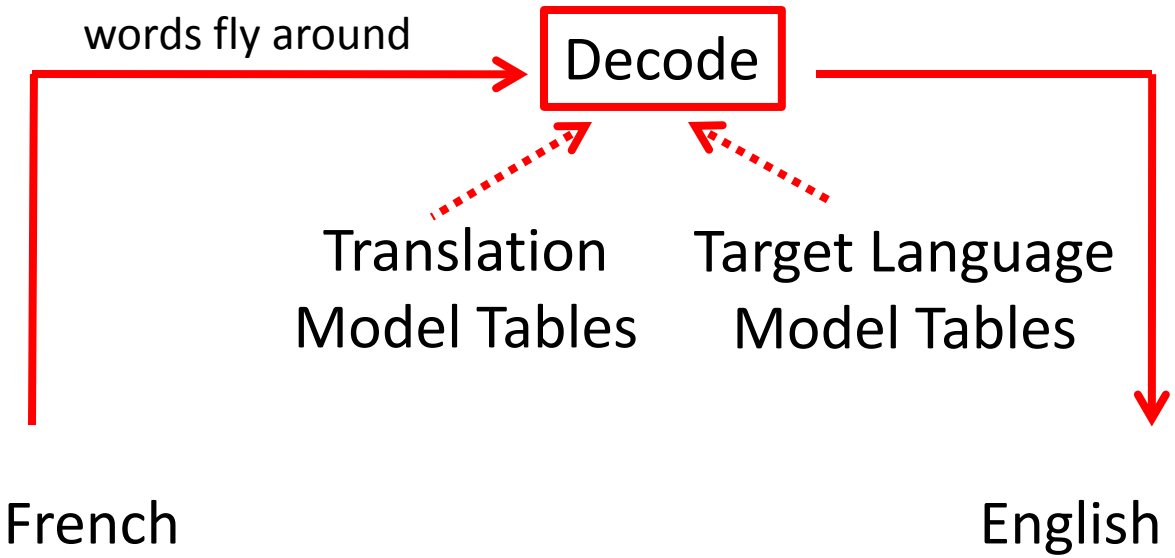
Brown et al, 1988



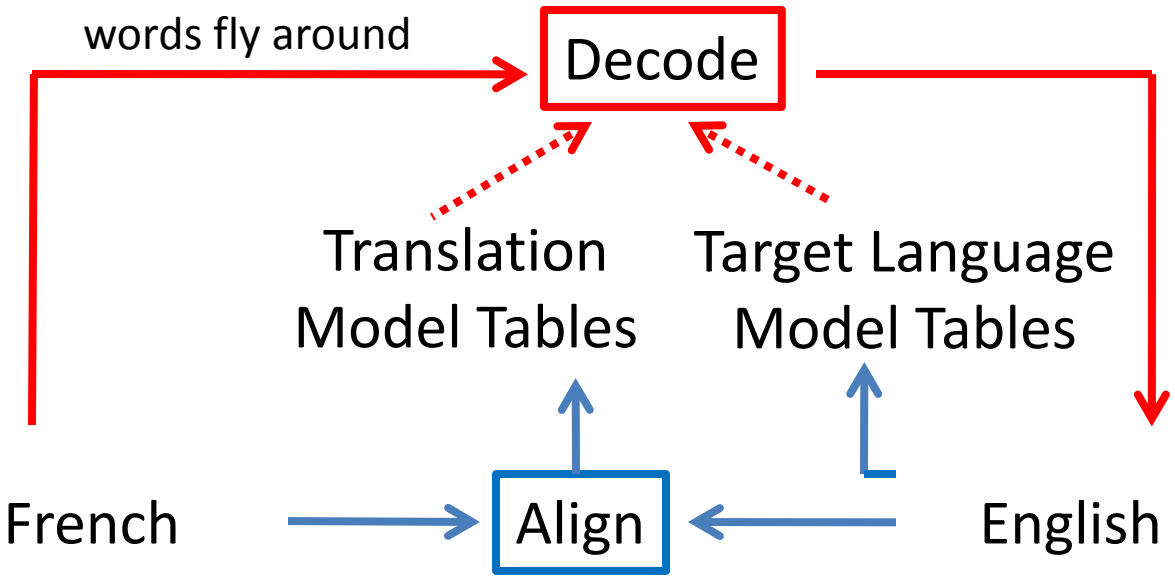
Brown et al, 1988



Brown et al, 1988

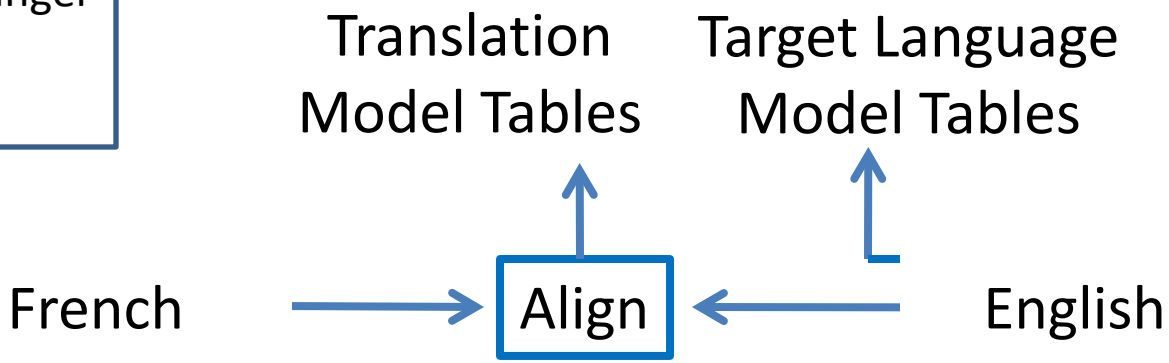


Brown et al, 1988

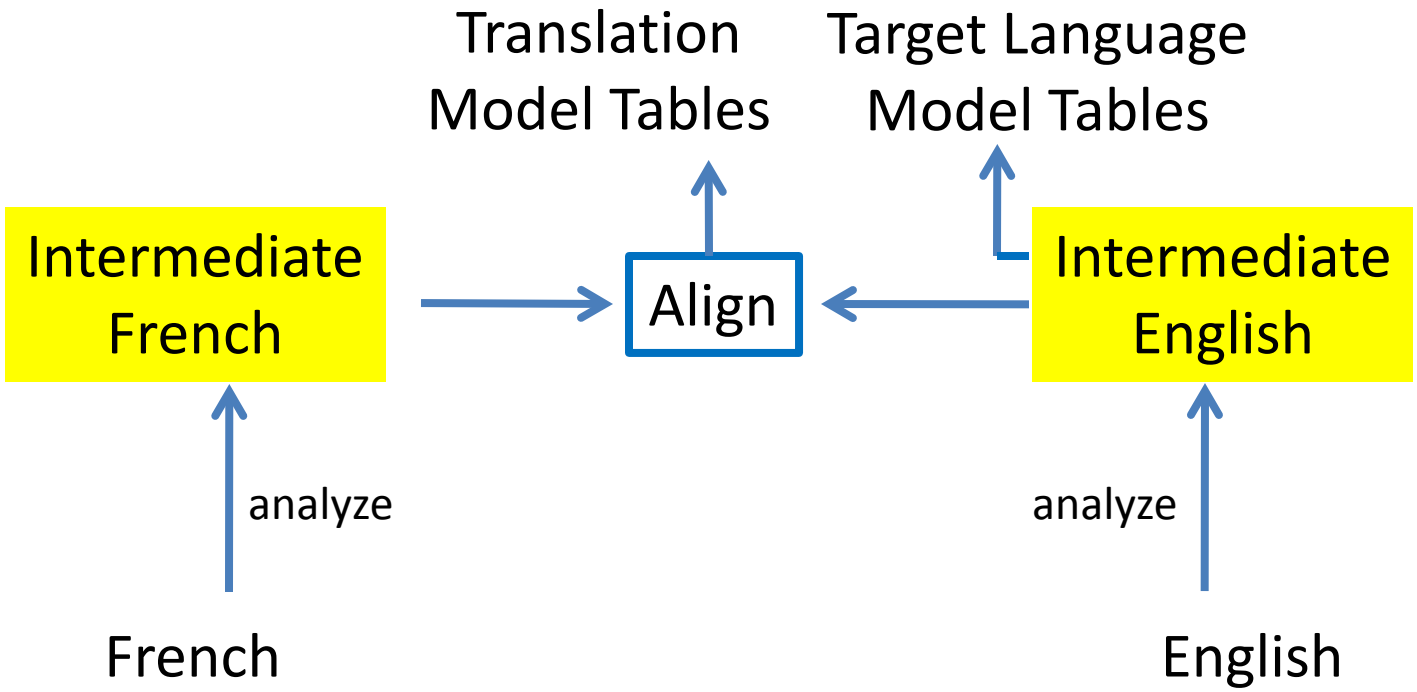


Brown et al, 1988

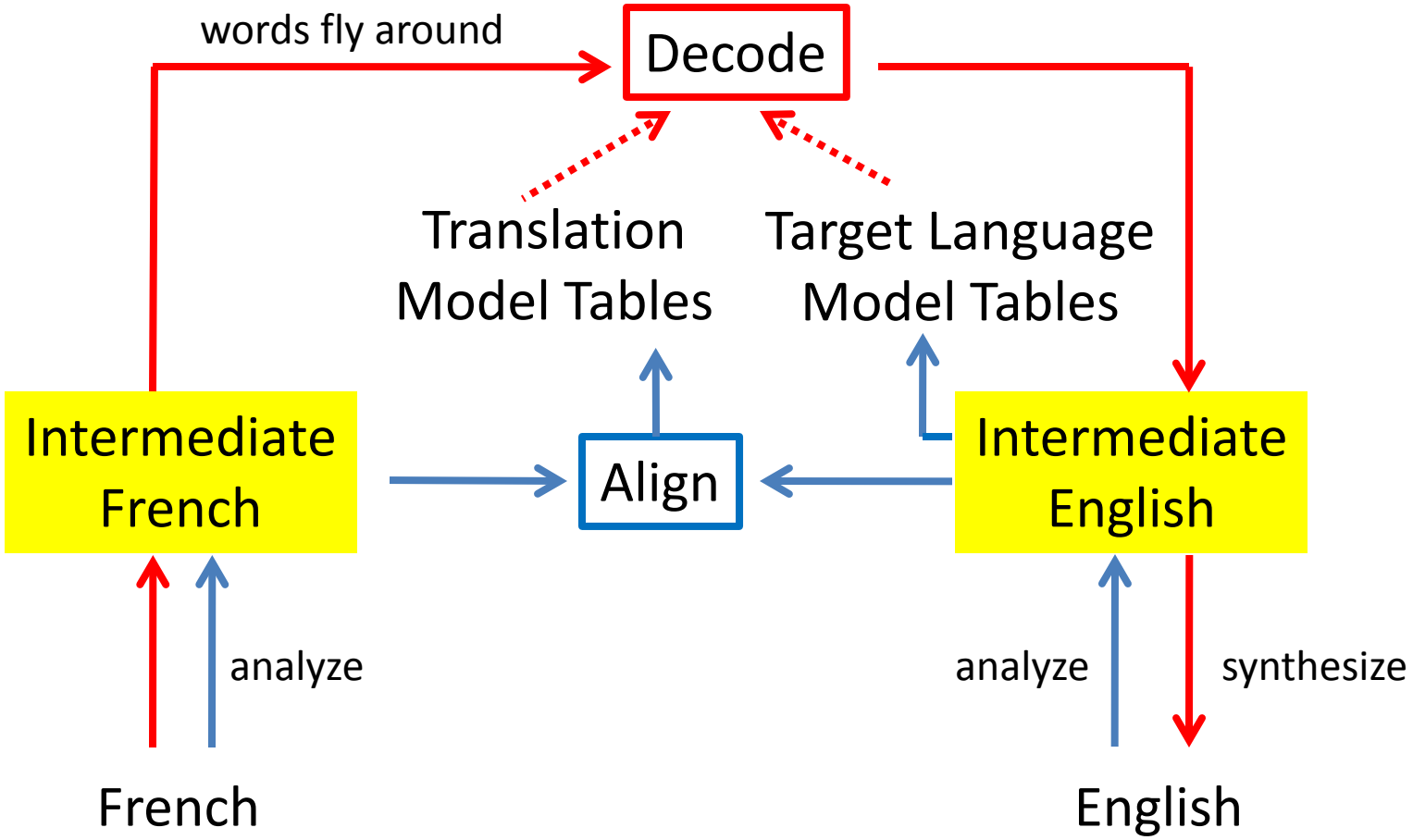
eats	mange
ate	mangé
eating	manger
eat	manger
…	



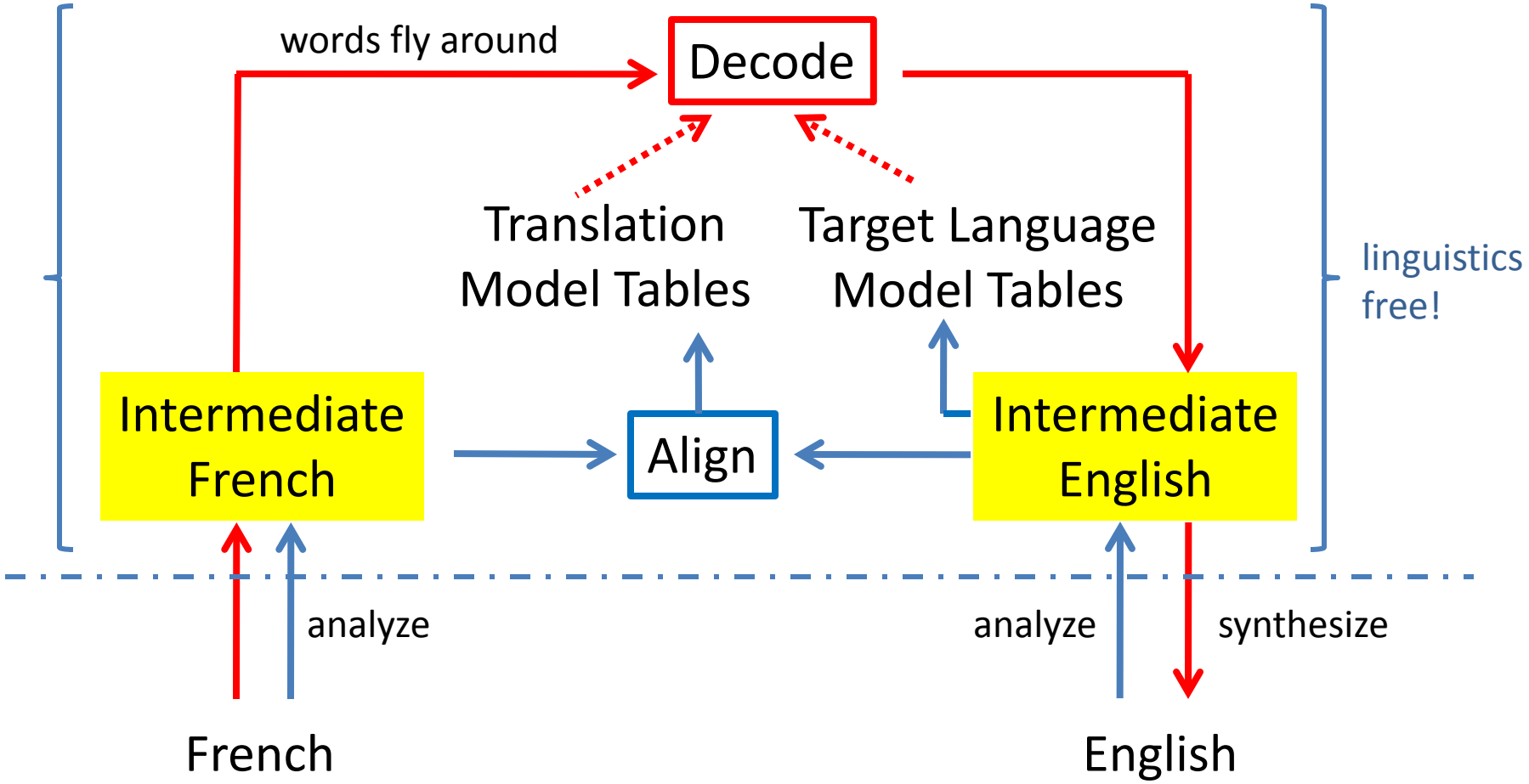
Brown et al, TMI 1992



Brown et al, TMI 1992



Brown et al, TMI 1992



Brown et al, TMI 1992

Peter Brown: "We gotta take the languages and **line 'em up**."

Je ne sais pas.

⇒ *Je sais ne_pas.*

Je vous le donnerai.

⇒ *Je donnerai le_DPRO vous_IPRO.*

Why should farmers be growing less wheat?

⇒ *Why farmers should be growing less wheat QINV*

Because of errors in grammatical tagging, compounded with the primitive nature of the rules that we employ to achieve this goal, we succeed only about 40% of the time.

Mangez-vous des légumes?

⇒ *Vous mangez des légumes QINV1*

Brown et al, TMI 1992

He was eating the peas more quickly than I.

⇒ *He PAST_PROGRESSIVE to_eat the pea N_PLURAL quick er_ADV than I.*

Ils se sont lavés les mains sales.

⇒ *Ils 3RD_PERSON_PLURAL_PAST laver se_RPRO les sale main N_PLURAL.*

Notice in the last example that we retain no indication of the original number on French adjectives. We also discard any distinction in gender. Thus, in the intermediate French, adjectives always appear in their masculine singular form.

Brown et al, TMI 1992

We assign senses to 1000 of the most frequent French words. For example, we map *prendre* to *prendre_1* in the sentence

Je vais prendre ma propre voiture,

but to *prendre_2* in the sentence

Je vais prendre ma propre décision.

Brown et al, TMI 1992

We restrict our attention to vocabularies of 40,809 English words and 57,802 French words. In the enhanced system, morphological analysis reduces these to 33,041 English morphemes and 31,115 French morphemes.

We estimated the parameters of the translation model for each system from a set of 1,778,620 pairs of French and English sentences from the Canadian Hansard data [1, 2]. Each of these sentences is 30 words or less in length. We tested both systems on the same set of 100 randomly selected Hansard sentences each containing at most 10 words. We judged as acceptable 39 of the translations produced by the simpler system as compared with 60 of those produced by the enhanced system.

Brown et al, TMI 1992

In work of this type, it is desirable to be able ascribe certain increments of performance to certain of the steps in the analysis or synthesis component, and thus to assess the value of the various transformations. Making such an assessment would require of us that we construct a series of analysis and synthesis components with different members of the series including different ones of the steps that make up the complete system. Unfortunately, each such construction must have a differently trained statistical transfer component. Because training is a costly undertaking, we have not made any of these collateral investigations and are, therefore, unable to say which of the new analysis and synthesis steps is the most valuable.

Brown et al, TMI 1992

In work of this type, it is desirable to be able ascribe certain increments of performance to certain of the steps in the analysis or synthesis component, and thus to assess the value of the various transformations. Making such an assessment would require of us that we construct a series of analysis and synthesis components with different members of the series including different ones of the steps that make up the complete system. Unfortunately, each such construction must have a differently trained statistica is a costly undertaking, we have not made any of these co unable to say which of the new analysis and synthesis ste

Bob Mercer

Peter Brown

Brown et al, TMI 1992

So long, suckers!

In work of this type, it is desirable to be able … nance to certain of the steps in the analysis or synthesis co… e of the various transformations. Making such an assessment … a series of analysis and synthesis components with different mem… the series including different ones of the steps that make up the complete system. Unfort…ately, each such construction must have a differently trained statistica… is a costly undertaking, we have not made any of these co… unable to say which of the new analysis and synthesis ste…

Bob Mercer

Peter Brown

Berger et al, 1996

to make predictions about the future behavior about the process.

Baseball managers (who rank among the better paid statistical modelers) employ batting averages, compiled from a history of at-bats, to gauge the likelihood that a player will succeed in his next appearance at the plate. Thus informed, they manipulate their lineups accordingly. Wall Street speculators (who rank among the *best* paid statistical modelers) build models based on past stock price movements to predict tomorrow's fluctuations and alter their portfolios to capitalize on the predicted future. At the other end of the pay scale reside natural language researchers, who design language and acoustic models for use in speech recognition systems and related applications.

The pa ant progress toward increasing the

Berger et al, 1996

to make predictions about the future behavior about the process.

Baseball managers (who rank among the be…ploy batting averages, compiled from a history of a…hat a player will succeed in his next appearance at th…nipulate their lineups accordingly. Wall Street specu… paid statistical modelers) build models based on pas…ct tomorrow's fluctuations and alter their portfolios to ca… on the predicted future. At the other end of the pay scale reside natural l…ge researchers, who design language and acoustic models for use in speech r…nition systems and related applications.

The pa…ant progress toward increasing the

Time to switch jobs!

Brown et al, 2010

Wall Street Journal, March 16, 2010

Renaissance Technologies LLC, one of the most successful hedge-fund companies ever… Medallion fund averaged returns of about 45% a year, after fees, since its inception in 1988.

Now [new co-CEOs] **Peter Brown** and **Bob Mercer** … must steer the firm through challenging waters.

Brown et al, 2010

Wall Street Journal, March 16, 2010

Renaissance Technologies LLC, one of the most successful hedge-fund companies ever… Medallion fund averaged returns of about 45% a year, after fees, since its inception in 1988.

Now [new co-CEOs] **Peter Brown** and **Bob Mercer** … must steer the firm through challenging waters.

1988	$ 100
1989	145
1990	210
1991	305
1992	442
1993	641
1994	929
1995	1,348
1996	1,954
1997	2,833
1998	4,108
1999	5,957
2000	8,638
2001	12,525
2002	18,162
2003	26,334
2004	38,185
2005	55,368
2006	80,283
2007	116,410
2008	168,795
2009	244,753

1992-2010: Rising Ambitions

- **Large parallel data!**
  - Eventually observe 99.9% of all word forms that will ever actually occur (& their translations)?
- **Large target language models!**
  - Output “casa roja” or “casa rojo”?
- **Lots of language pairs!**
  - Too much work to “line up” every pair?
- **Avoid medicine with bad side-effects!**
  - What if analyzer/synthesizer makes mistakes, and what if they mess up what was already working fine?

Large Language Models

The weather is cool.

→ El clima es fresco.

(el clima = 6m web hits, la clima = 228k)

That's cool.

→ Eso es genial.

Can you dig it?

→ ¿Puedes creerlo? (i.e., can you believe it?)

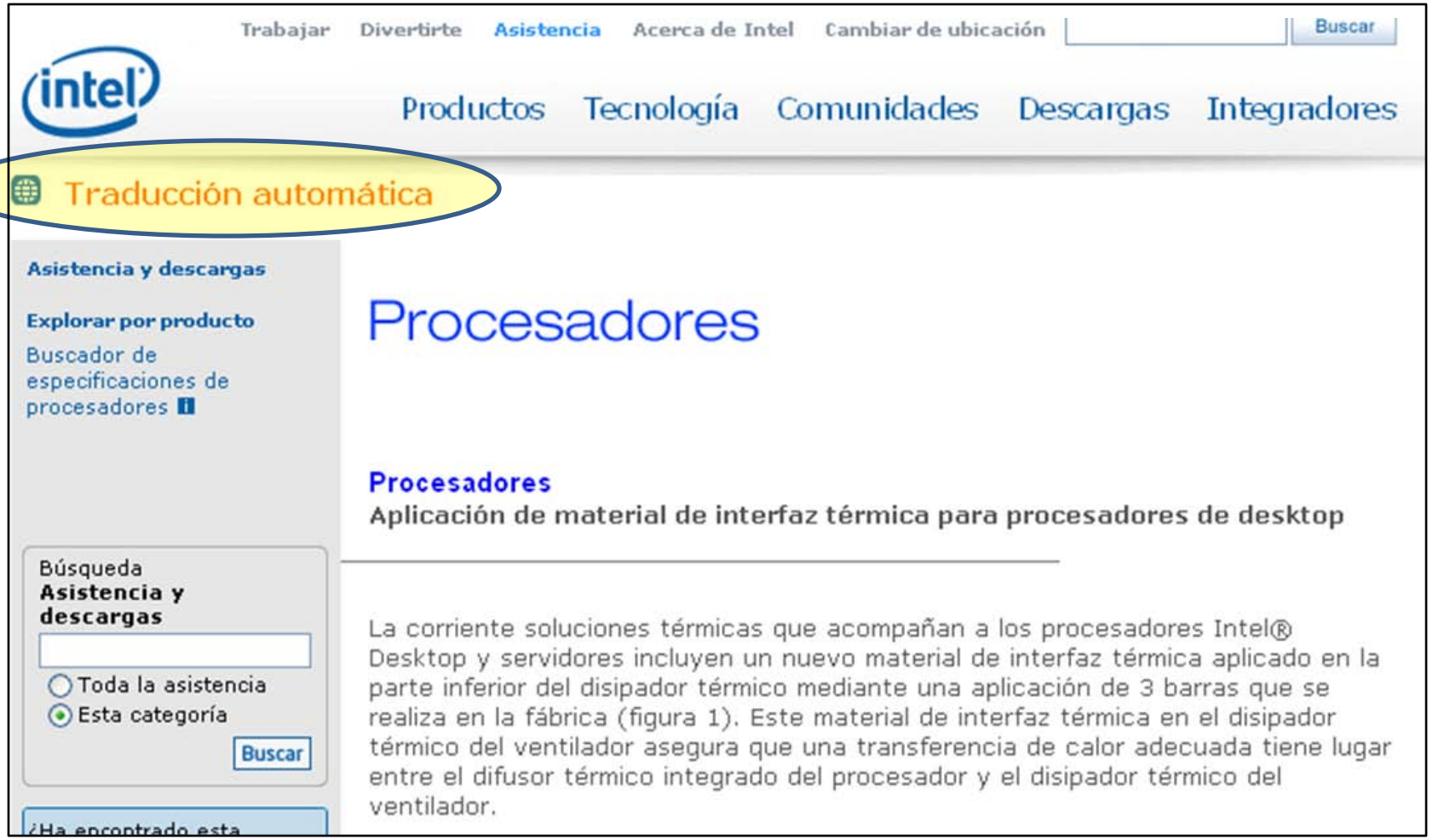
I dig her.

→ Yo la excavación. (i.e., I-her-excavation)

Google Translate, 5/10

Live News Translation

Online Help


Trabajar
Divertirte
Asistencia
Acerca de Intel
Cambiar de ubicación
Buscar
intel
Productos
Tecnología
Comunidades
Descargas
Integradores
Traducción automática
Asistencia y descargas
Explorar por producto
Buscador de especificaciones de procesadores
Procesadores
Procesadores
Aplicación de material de interfaz térmica para procesadores de desktop
Búsqueda
Asistencia y descargas
Toda la asistencia
Esta categoría
Buscar
La corriente soluciones térmicas que acompañan a los procesadores Intel® Desktop y servidores incluyen un nuevo material de interfaz térmica aplicado en la parte inferior del disipador térmico mediante una aplicación de 3 barras que se realiza en la fábrica (figura 1). Este material de interfaz térmica en el disipador térmico del ventilador asegura que una transferencia de calor adecuada tiene lugar entre el difusor térmico integrado del procesador y el disipador térmico del ventilador.

User-Generated Content

TripAdvisor.com: Millions of English user reviews translated into French, Spanish, German, etc.

Deployed Statistical MT Systems

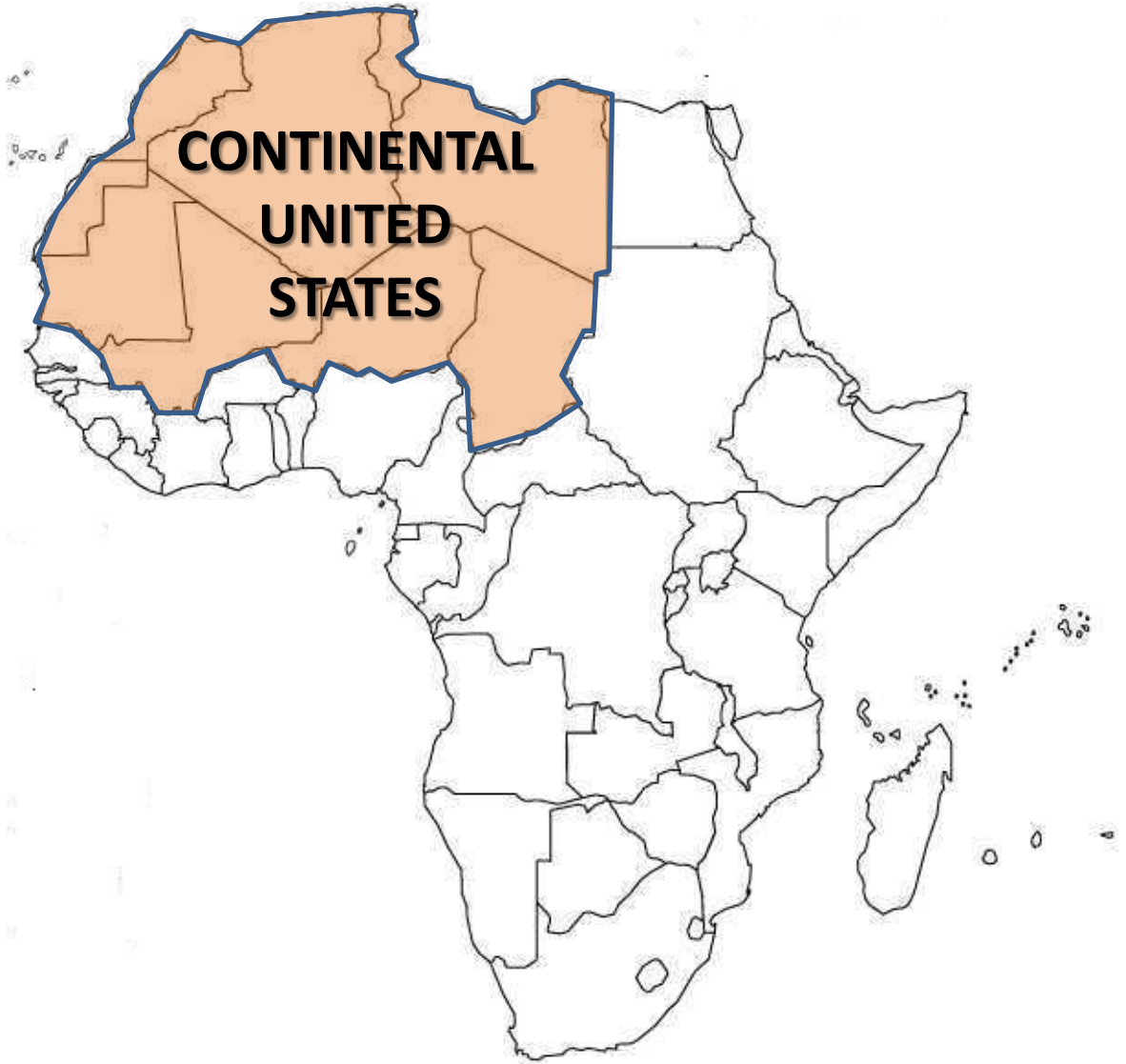
	Company X	Company Y
Western Europe	Danish, Dutch, Finnish, French, German, Greek, Italian, Norwegian, Portuguese, Spanish, Swedish	Catalan, Danish, Dutch, French, Galician, German, Greek, Icelandic, Irish, Italian, Maltese, Norwegian, Polish, Portuguese, Spanish, Swedish, Welsh
Eastern Europe	Bulgarian, Czech, Hungarian, Polish, Romanian, Russian, Serbian, Turkish	Albanian, Belarusian, Bulgarian, Croatian, Czech, Estonian, Finnish, Hungarian, Latvian, Lithuanian, Macedonian, Romanian, Russian, Serbian, Slovak, Slovenian, Turkish, Ukrainian, Yiddish
Middle East & Africa	Arabic, Hausa, Hebrew, Pashto, Persian, Somali, Urdu	Afrikaans, Arabic, Hebrew, Persian, Swahili
Asia	Chinese, Hindi, Indonesian, Japanese, Korean, Thai	Chinese, Filipino, Hindi, Indonesian, Japanese, Korean, Malay, Thai

Not Solved!

English source	Correct Swahili	MT Swahili
I am reading a book.	**Ninasoma kitabu.** I-PRESENT-read book	**I am kusoma kitabu.**
You are reading a book.	**Unasoma kitabu.** you-PRESENT-read book	**Wewe ni kusoma kitabu.**
He is reading a book.	**Anasoma kitabu.** he-PRESENT-read book	**Yeye ni kusoma kitabu.**

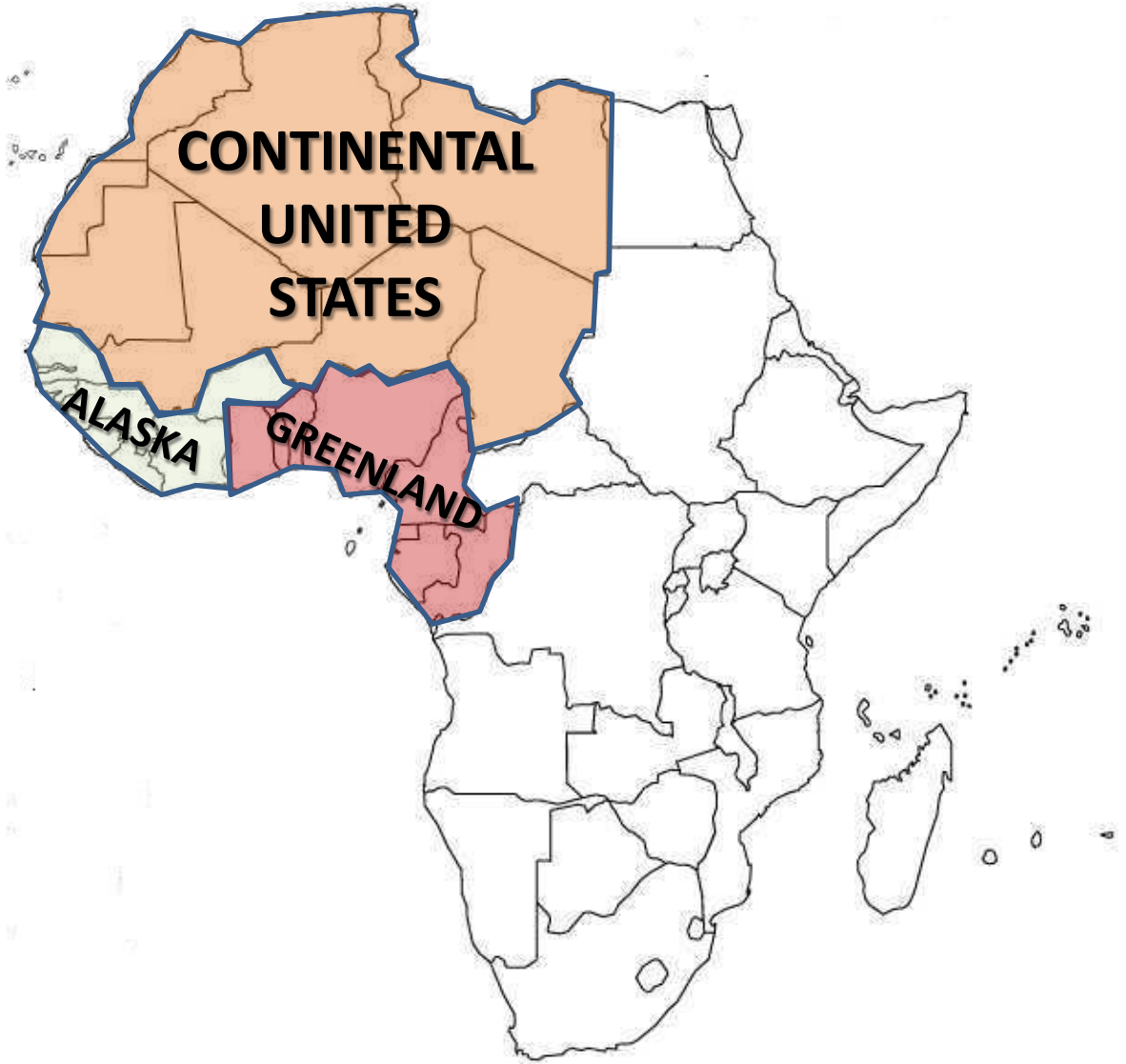
African Languages

African Languages

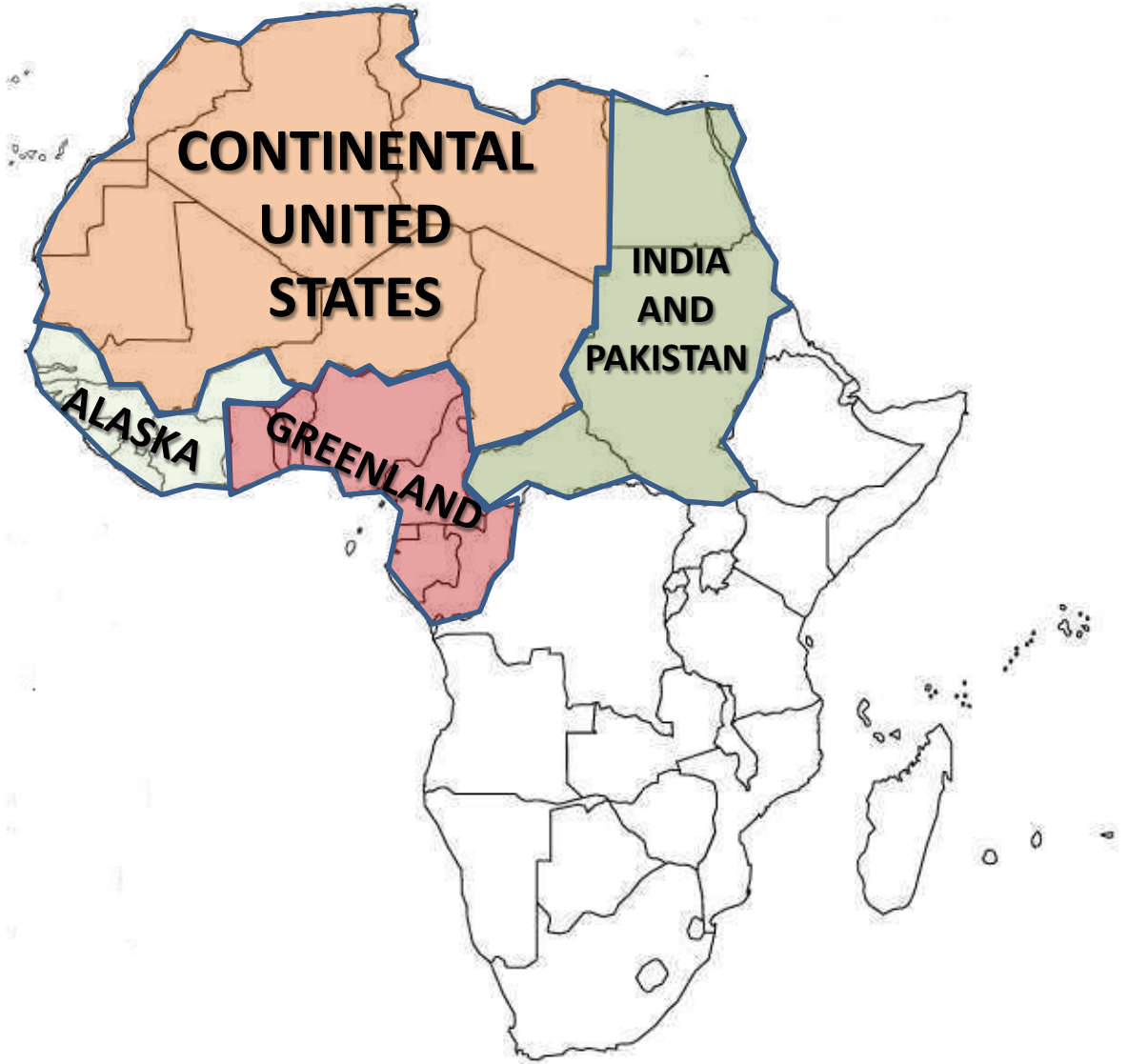


African Languages

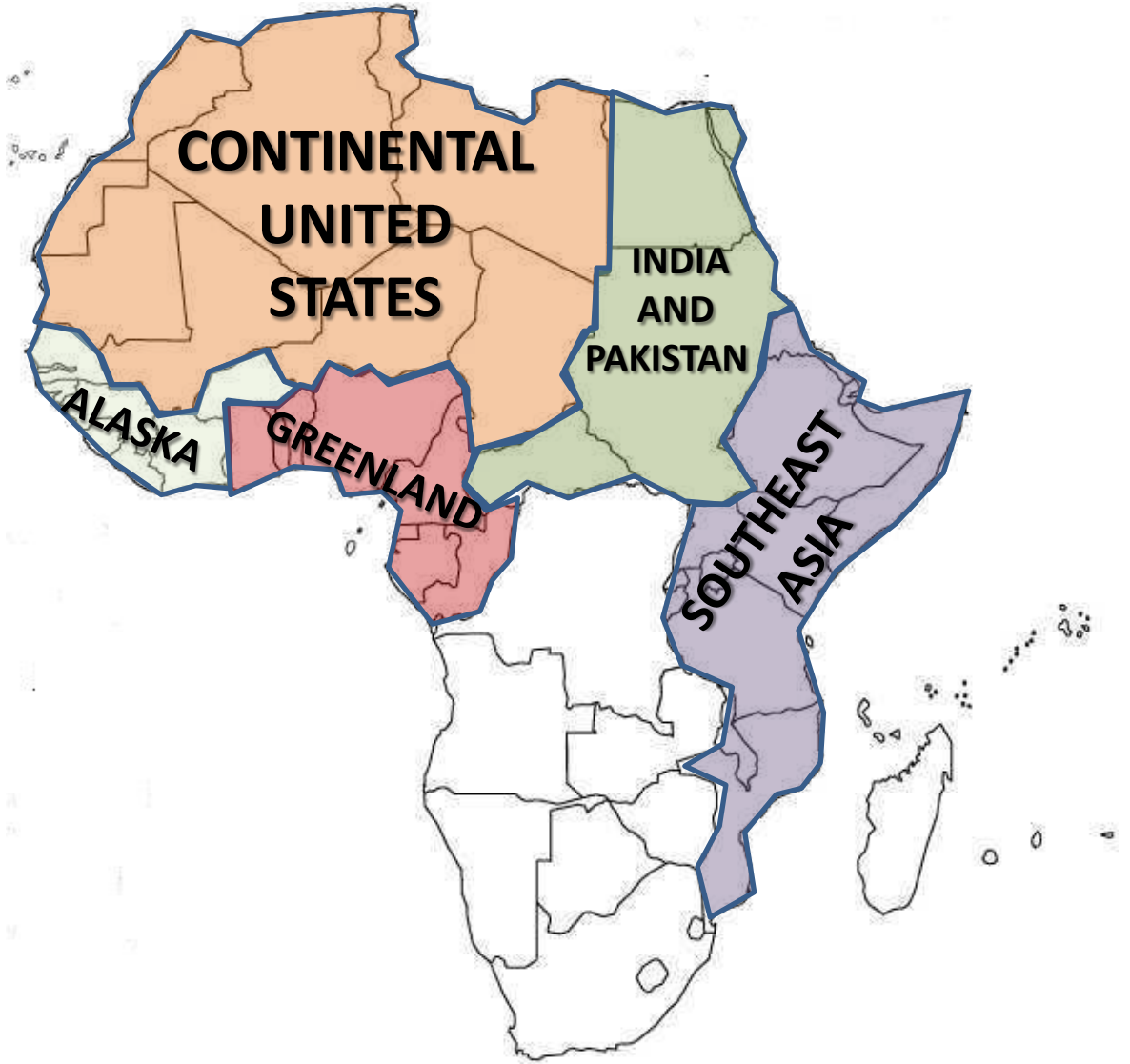
African Languages



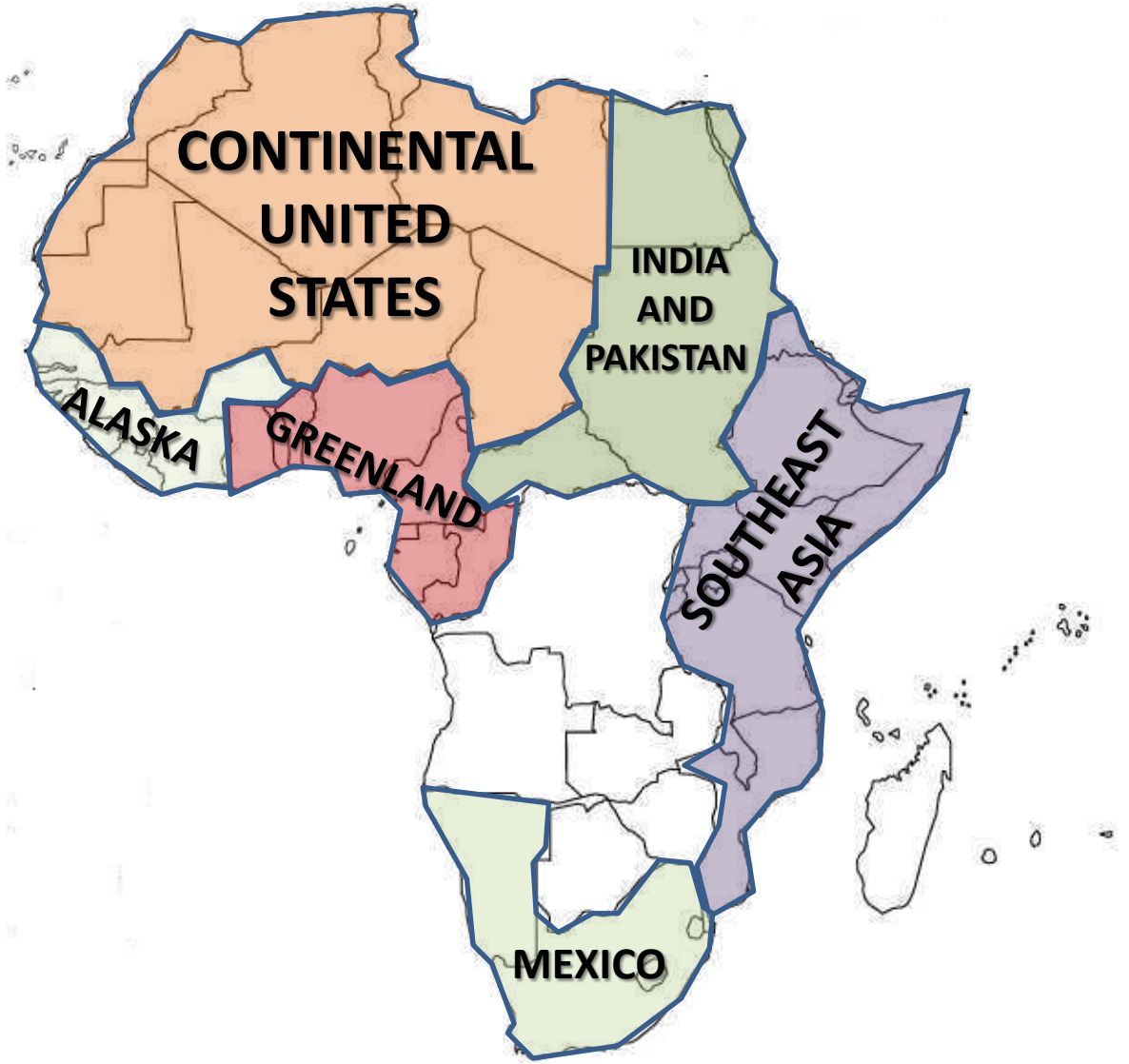
African Languages



African Languages



African Languages



African Languages

African Languages

African Languages

Africa

and
even
more
languages
spoken!

Even if you don’t care about that

- Translation into
  - Czech
  - Russian
  - Japanese
  - German
  - Korean
  - Hebrew
  - Arabic

  is problematic!

Even into Spanish

Gender agreement

English input	Spanish MT
Red house	Casa roja
Yellow house	Casa amarilla
Black house	Casa **negro**
Orange house	**Naranja casa**
An orange house	Una casa de color naranja

Verb inflection

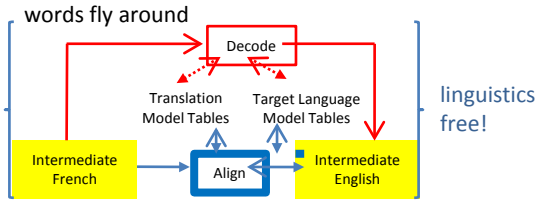
English input	Spanish MT
I saluted the flag.	Saludé a la bandera.
You saluted the flag.	Usted saludó a la bandera.
He saluted the flag.	Saludó a la bandera.
She saluted the flag.	Rindió homenaje a la bandera.
We saluted the flag.	**Saludamos** la bandera.
They saluted the flag.	Se saludaron a la bandera.

Recent Research

- Source → Source-prime
  - Analyzer re-orders source language
  - Makes use of source parsing & morphology
    - Not available in 1992!
  - Can encode uncertainties in **source lattice**
- Target → Target-prime
  - Translate into target lemma sequences
  - Synthesizer guesses inflections
    - Which may have no correlate in English anyway
- Factored Models

Recent Research

- Underlying engine is often still linguistics-free
  - Words fly around at decoding time
  - "Distortion" cost
  - Target language model supposed to sort it out



- In other words:
  - We already have a linguistics-free, distortion-based MT system
  - Let's bolt on some syntax/morph without disturbing the engine

Still, Lots of Details …

北极熊 → polar bear
北极熊 → polar bears

Chinese-to-English, 200m words training	Test Bleu
Baseline state-of-the-art MT	32.0
1) Separate English affixes in training data 2) Train, decode **3) Re-join affixes** in decoder output	30.6
1) Remove English affixes in training data 2) Train, decode **3) Guess inflections** for decoder output	31.6
1) Decode **2) Remove and re-generate** affixes in decoder output	31.4

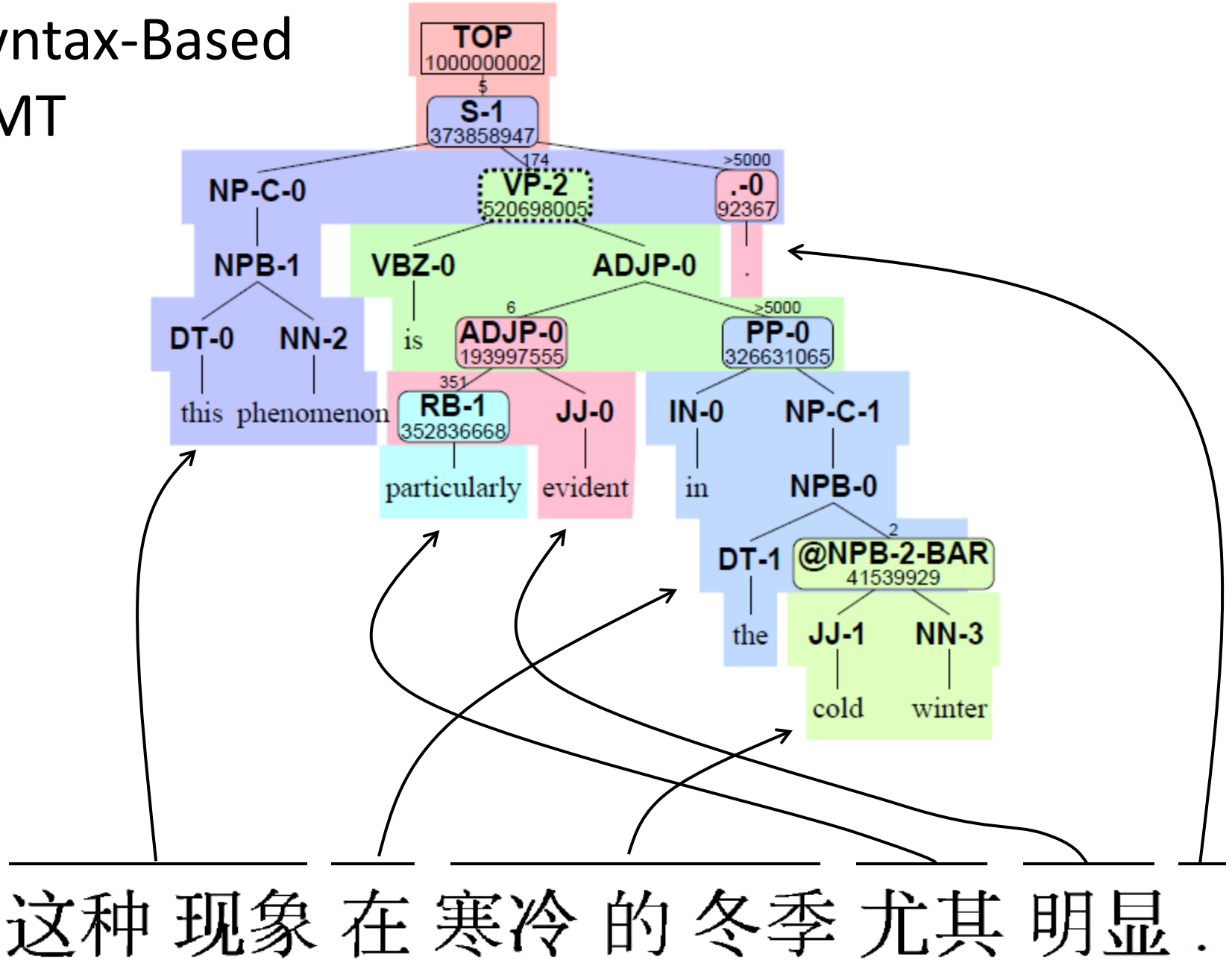
Not Super-Satisfying?

- Want rules like “Translate direct object by moving it before the verb and marking it with an accusative affix”
- We don’t have such morphological processes integrated in
- Syntactic movement has been getting attention, though
  - Next: description of syntax-based MT
  - Then: possible directions for morphology

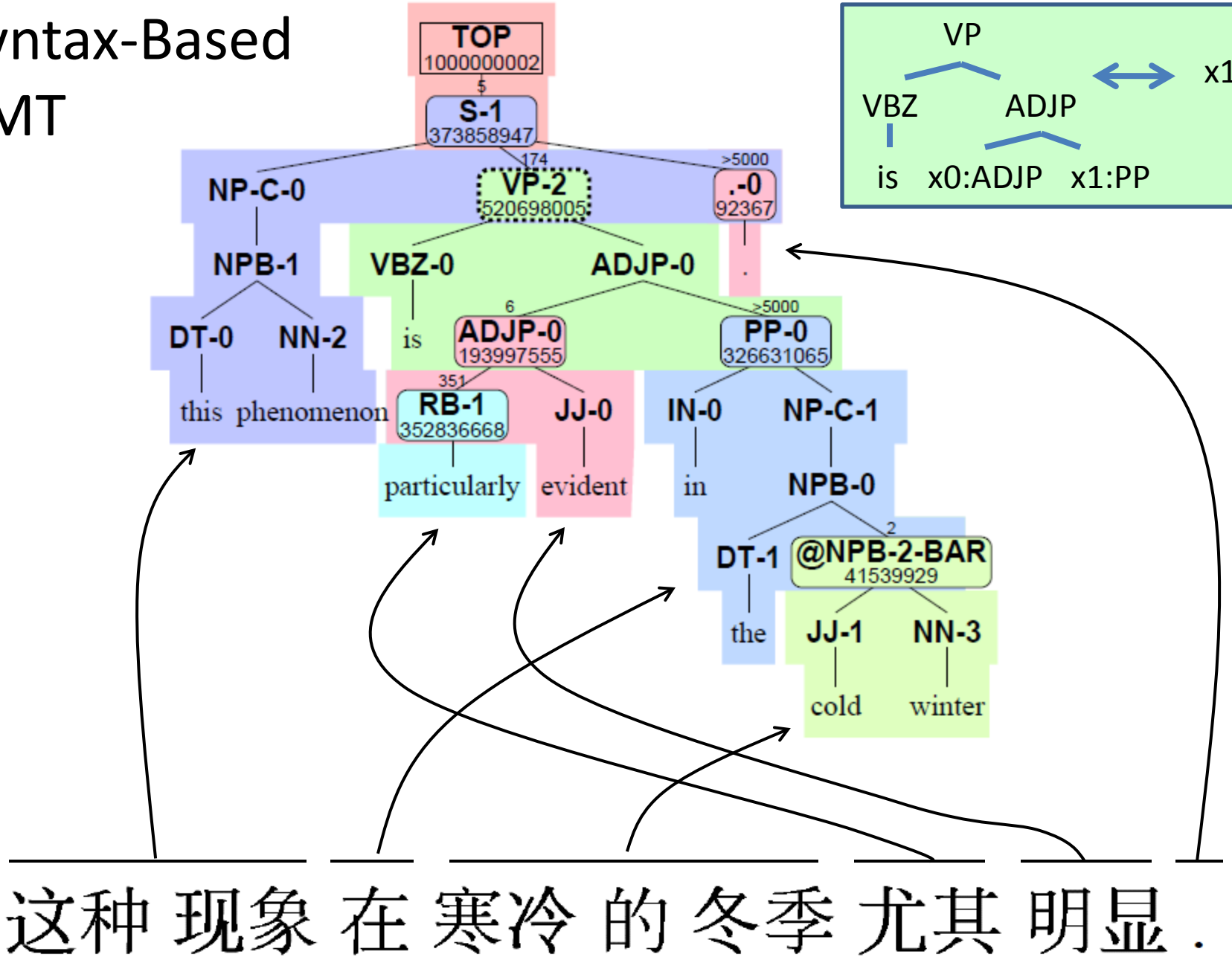
Syntax-Based
SMT

这种 现象 在 寒冷 的 冬季 尤其 明显 .

Syntax-Based SMT



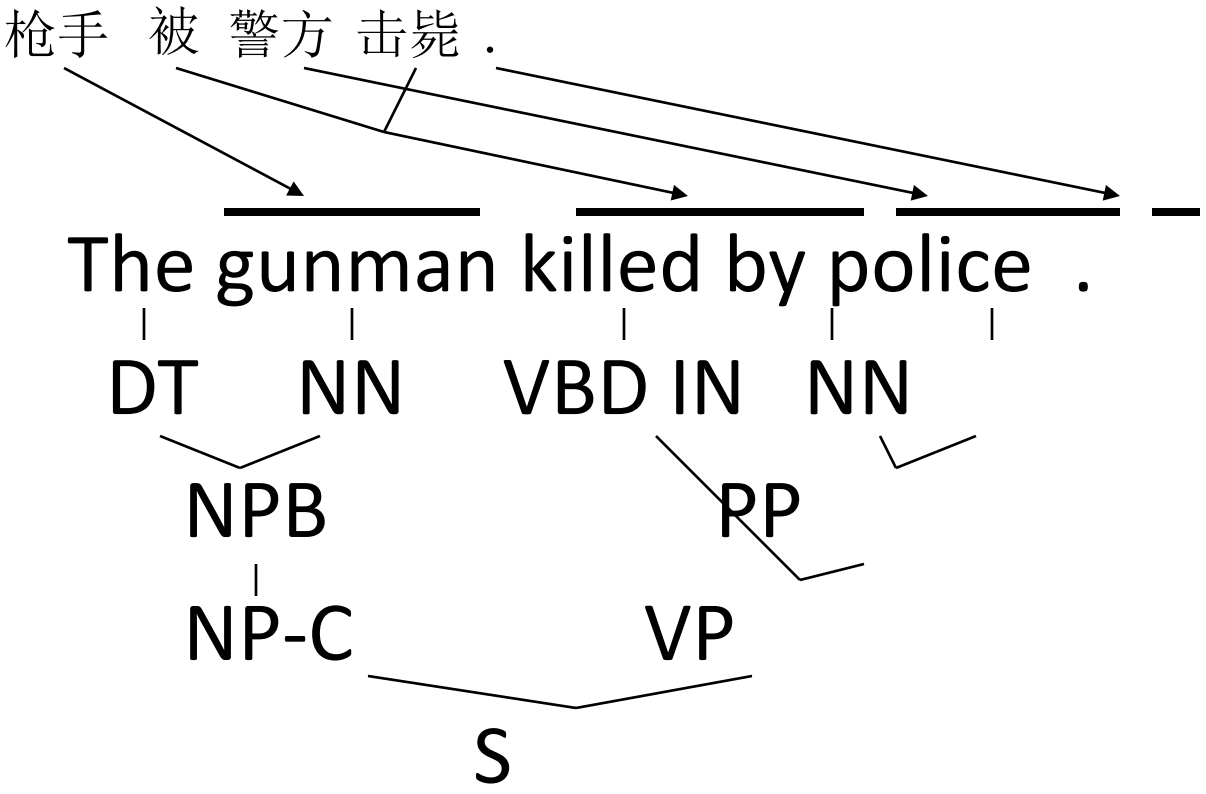
Syntax-Based SMT



Tree-Based Output

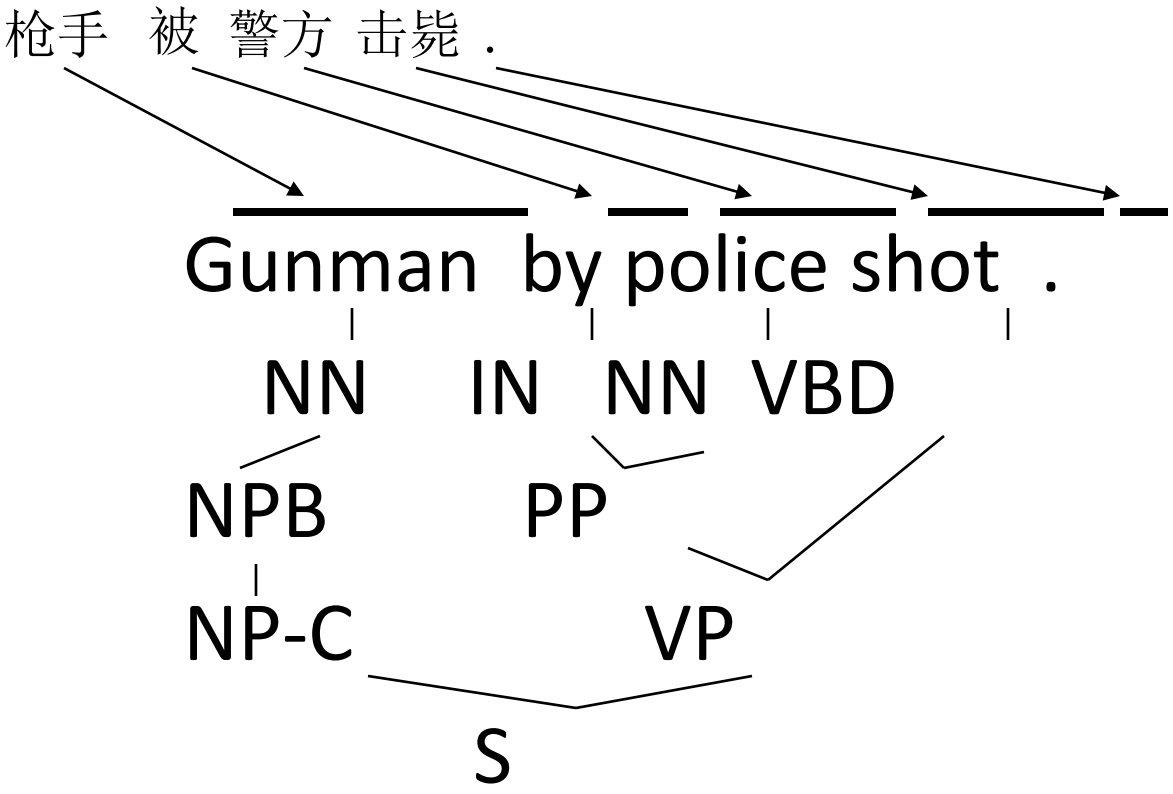
枪手 被 警方 击毙 .

Tree-Based Output



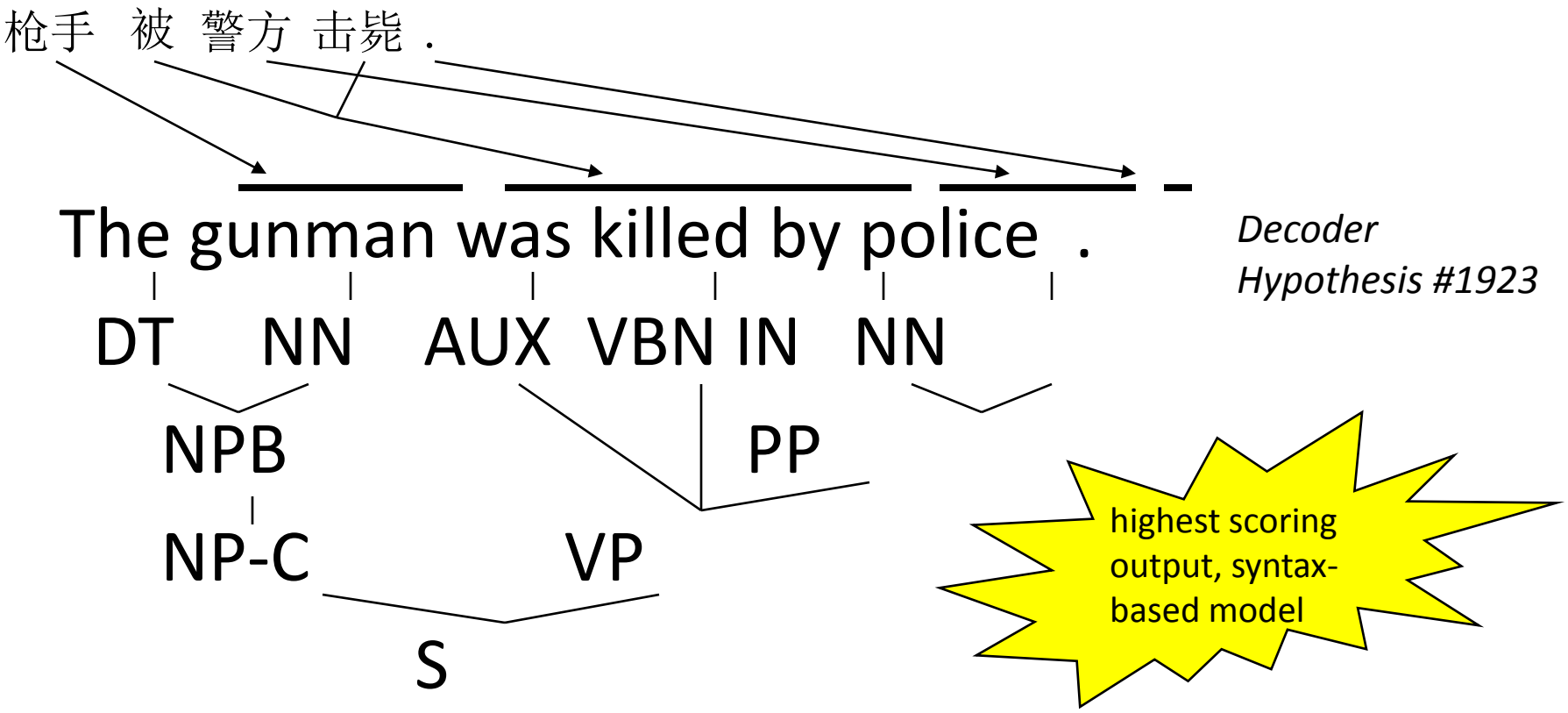
Decoder Hypothesis #1

Tree-Based Output



Decoder Hypothesis #16

Tree-Based Output

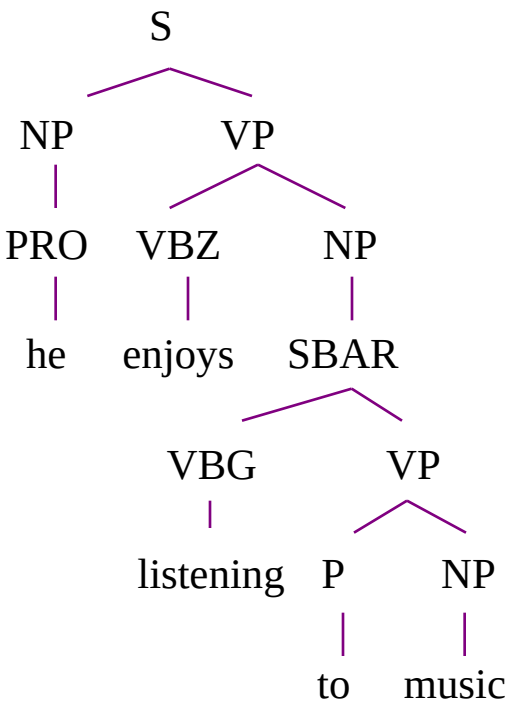


How does a Chinese string become an English tree, or vice-versa?

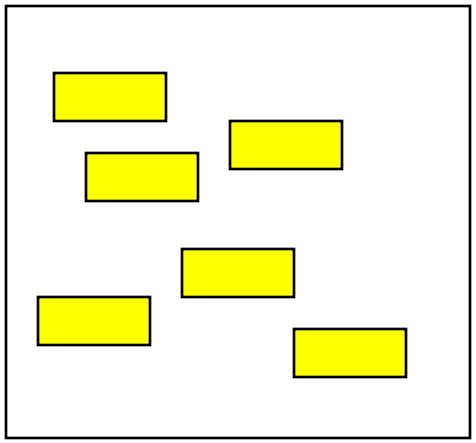
Top-Down Tree Transducer

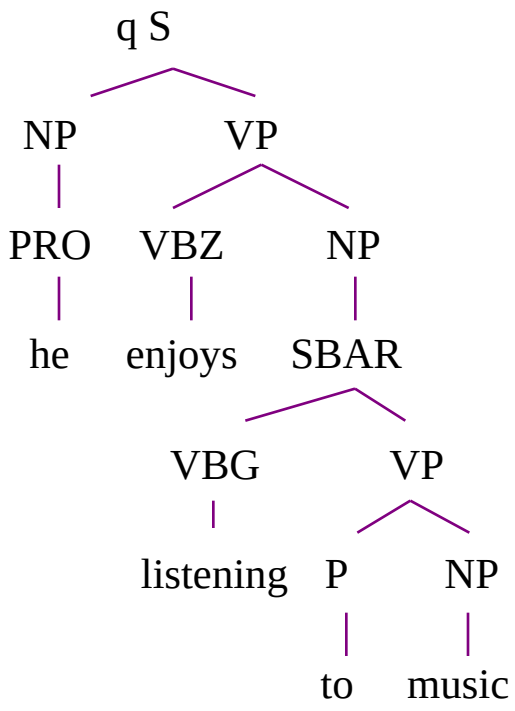
(W. Rounds 1970; J. Thatcher 1970)

Original input:



Transformation:

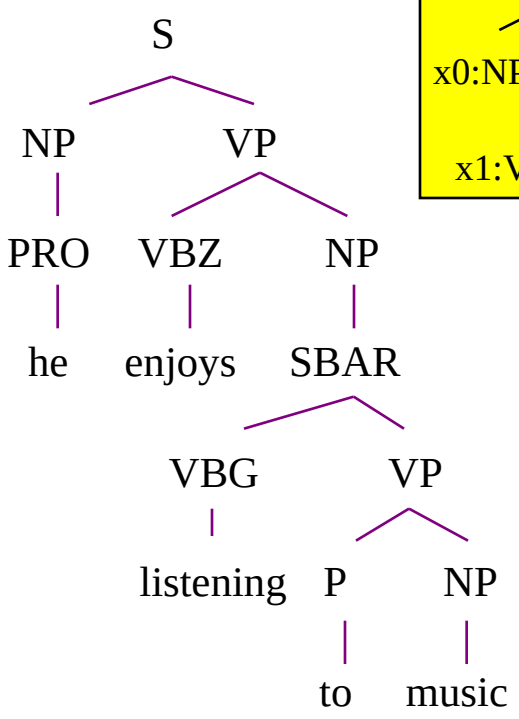


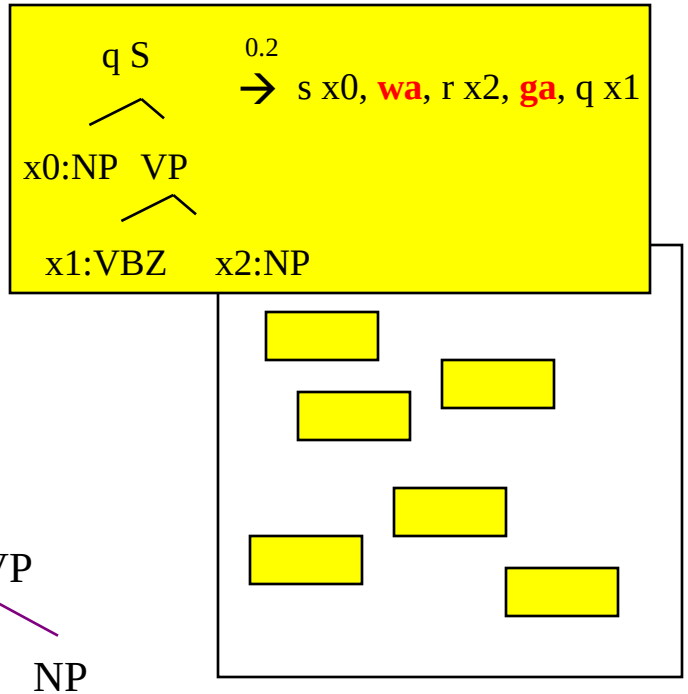


Top-Down Tree Transducer

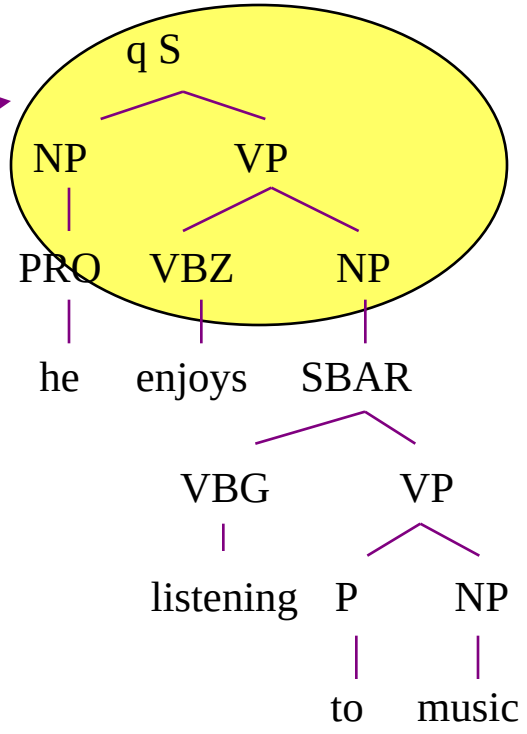
(W. Rounds 1970; J. Thatcher 1970)

Original input:





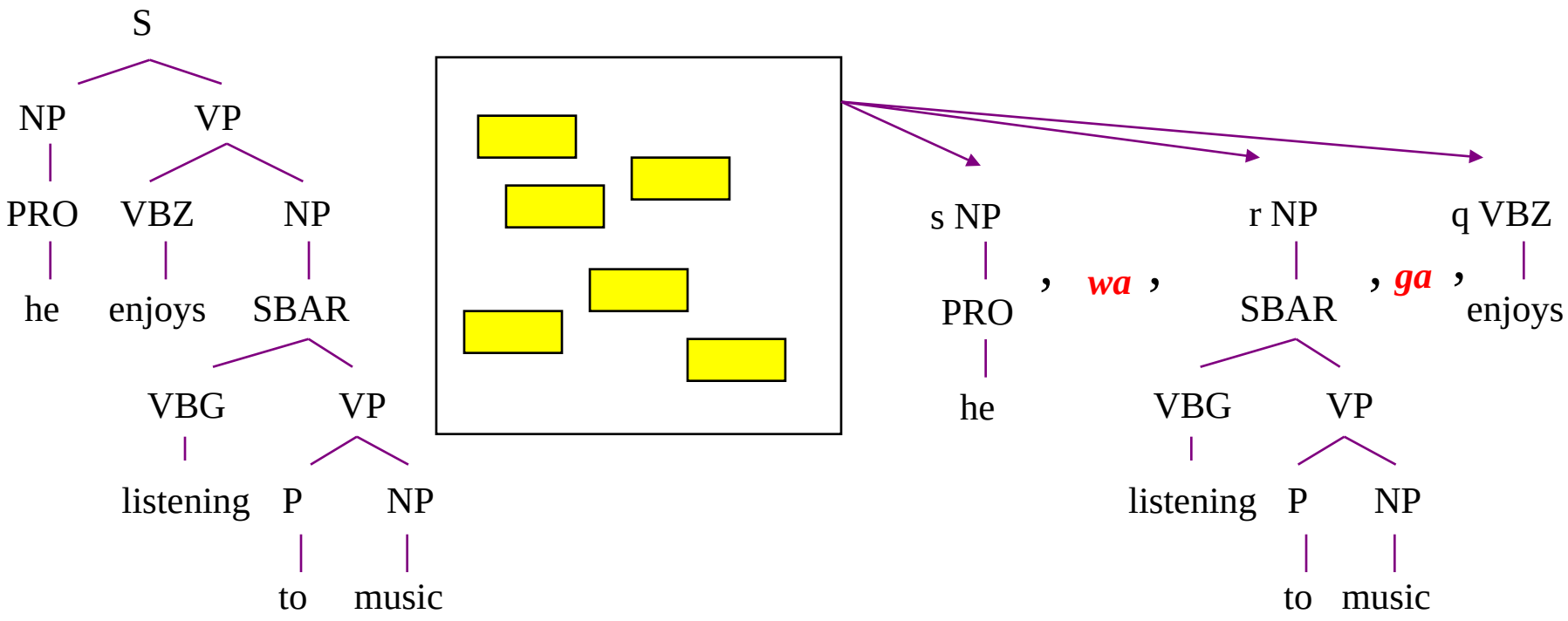
Transformation:



Top-Down Tree Transducer

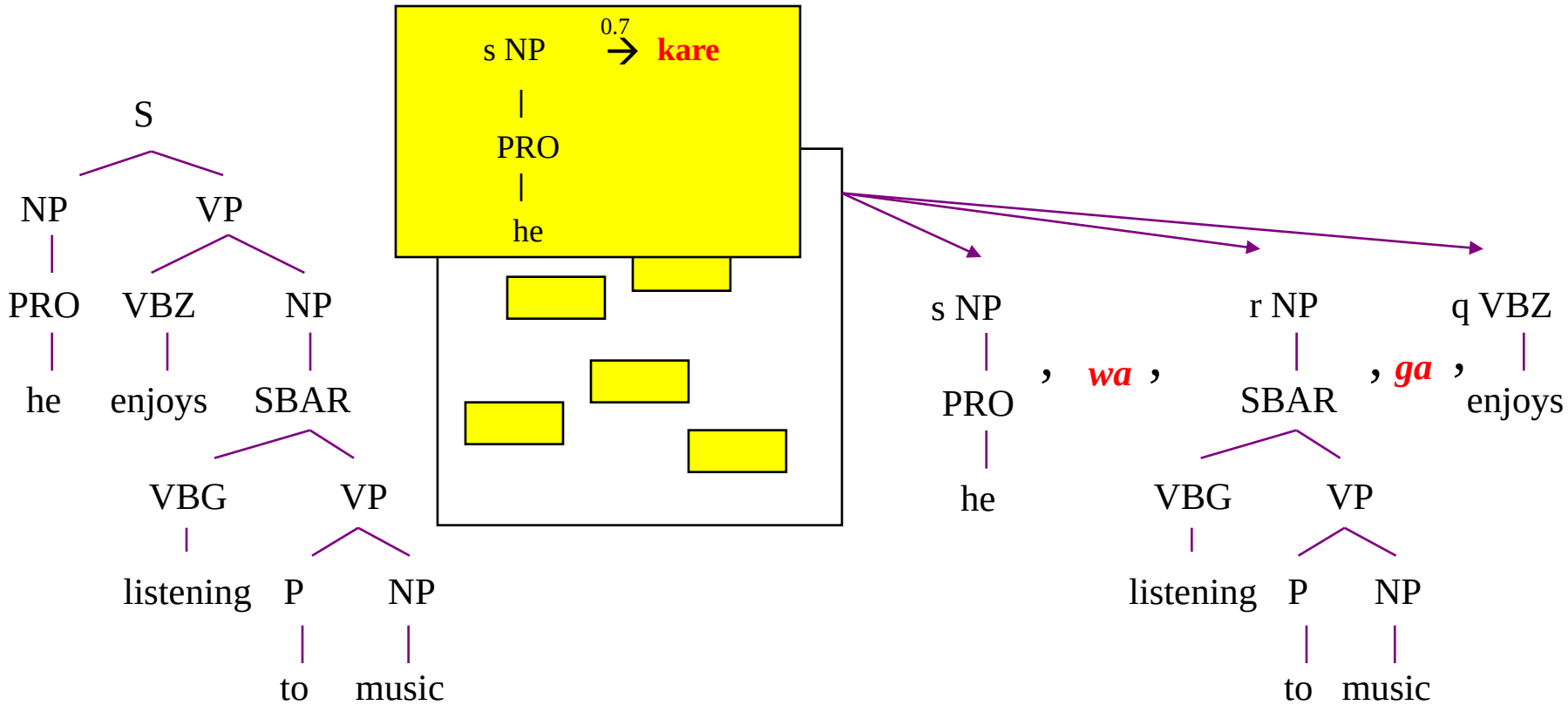
(W. Rounds 1970; J. Thatcher 1970)

Original input: Transformation:



Top-Down Tree Transducer

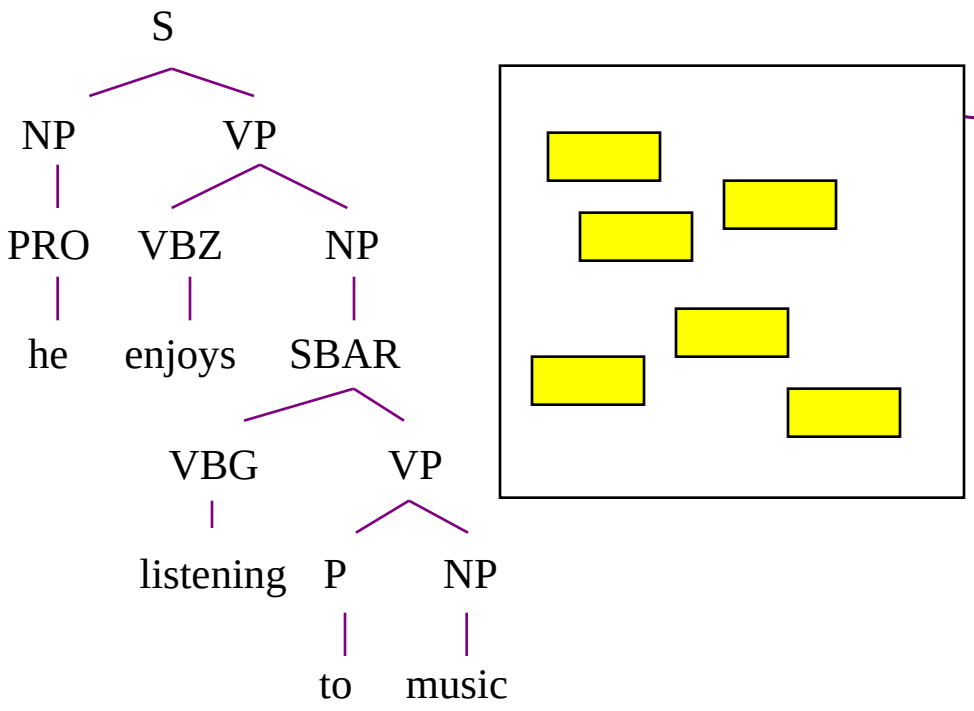
(W. Rounds 1970; J. Thatcher 1970)



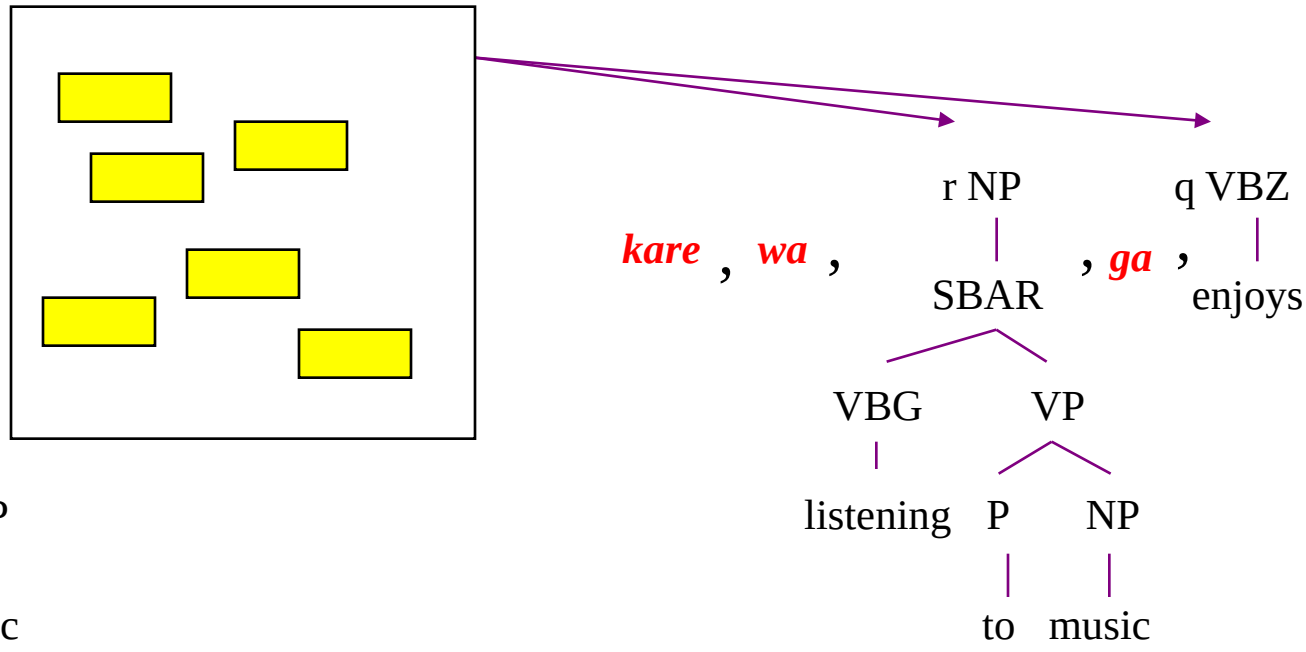
Top-Down Tree Transducer

(W. Rounds 1970; J. Thatcher 1970)

Original input:



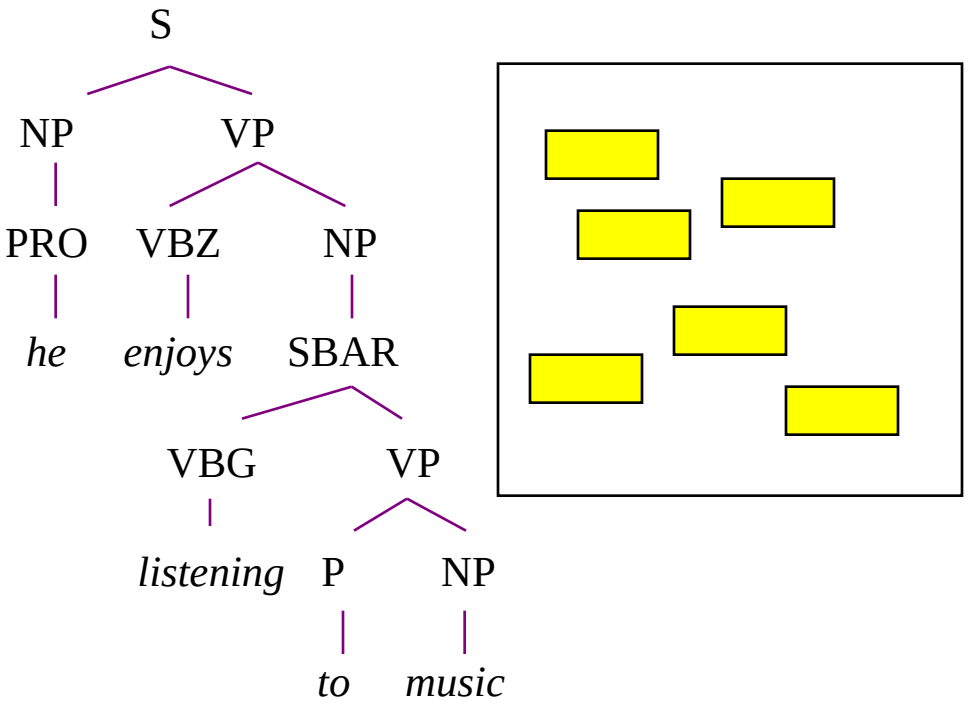
Transformation:



Top-Down Tree Transducer

(W. Rounds 1970; J. Thatcher 1970)

Original input:



Final output:

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

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

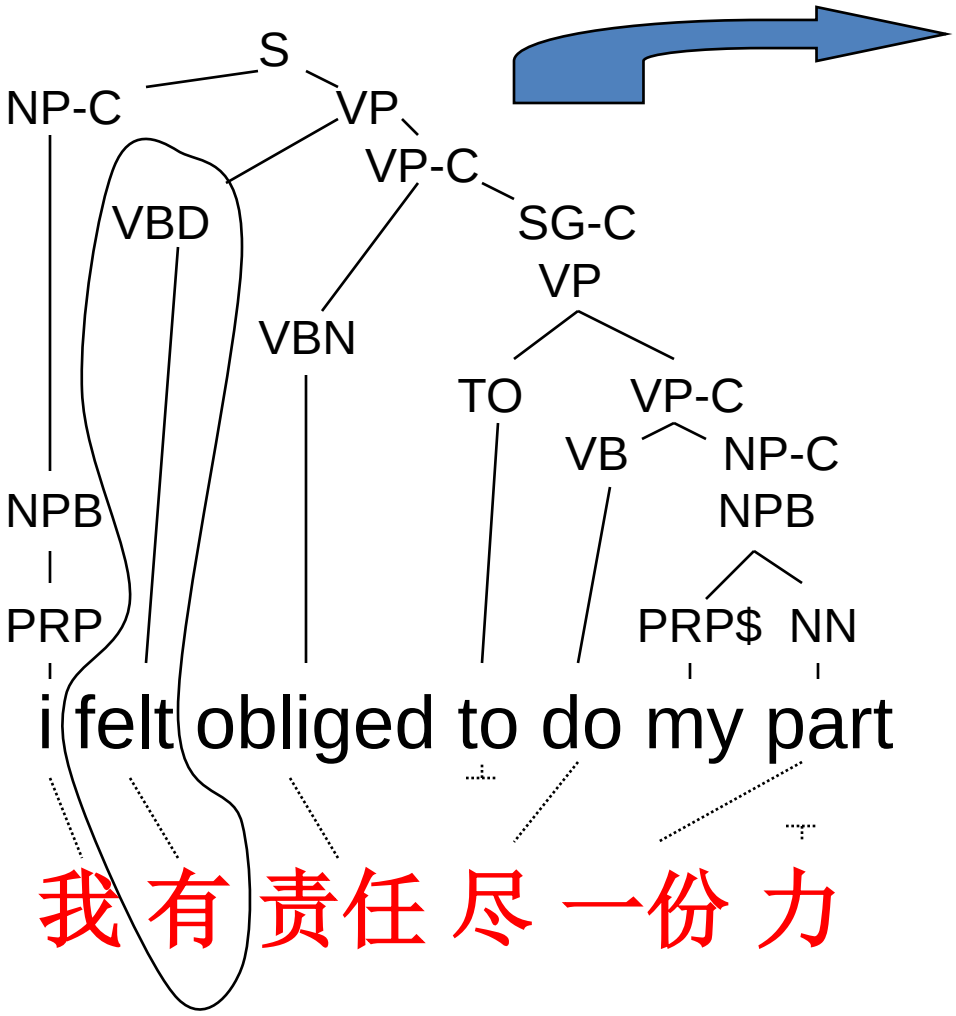
Top-Down Tree Transducer

- Introduced by Rounds (1970) & Thatcher (1970)

  "… parts of **mathematical linguistics can be formalized easily** in a tree-automaton setting …" (Rounds 1970, "Mappings on Grammars and Trees", *Math. Systems Theory* 4(3))

- Large theory literature
  - e.g., Gécseg & Steinby (1984), Comon et al (1997)
- Once again re-connecting with NLP practice
  - e.g., Knight & Graehl (2005), Knight (2007), May & Knight (2006), Maletti (2010), ATANLP workshop at ACL 2010, etc.

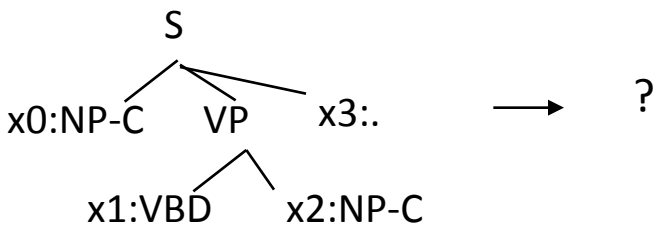
Tree Transducers Can be Extracted from Bilingual Data
(Galley, Hopkins, Knight, Marcu, 2004)



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

Sample Subject-Verb-Object 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

Decoding

- argmax P(etree | cstring)
  etree

- Difficult search problem
  - 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

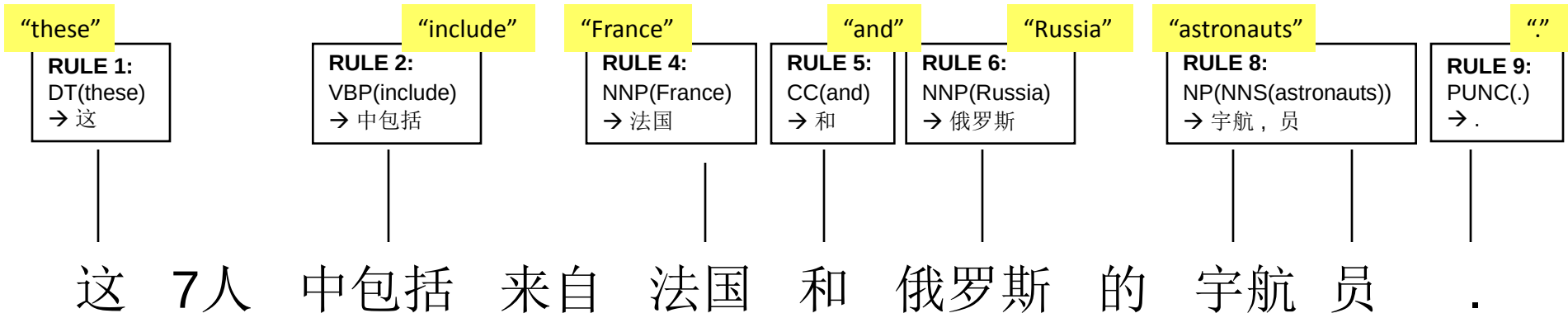
Syntax-Based Decoding

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

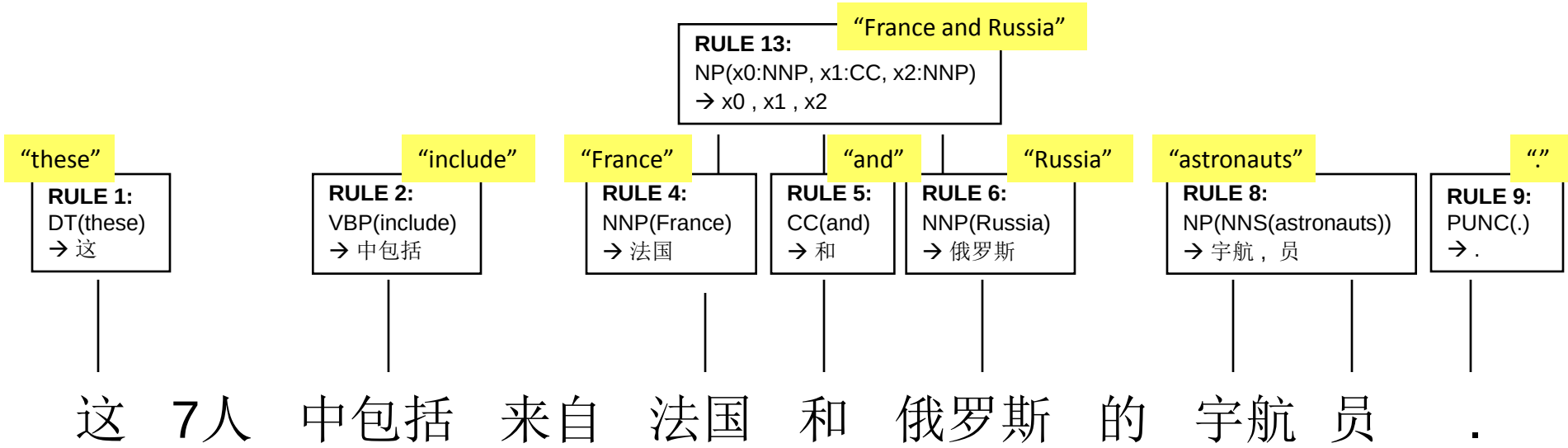
这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

Syntax-Based Decoding

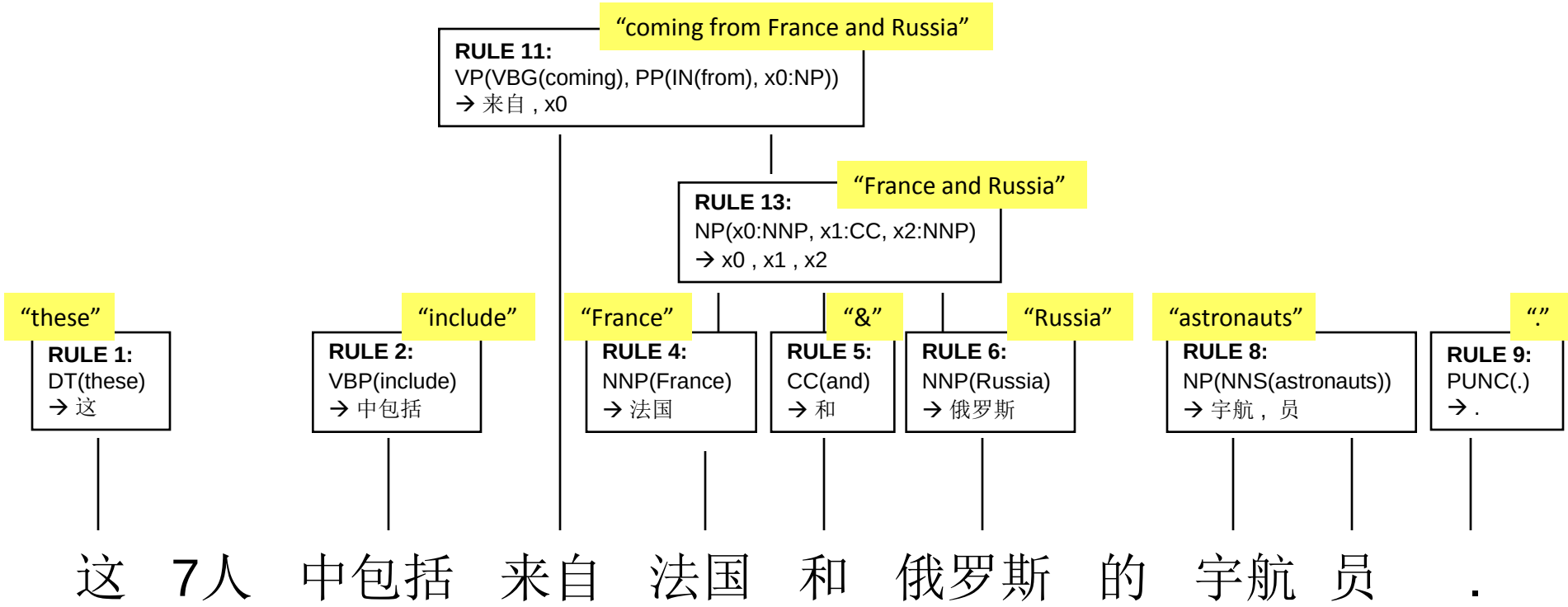
Rules apply when their right-hand sides (RHS) match some portion of the input.



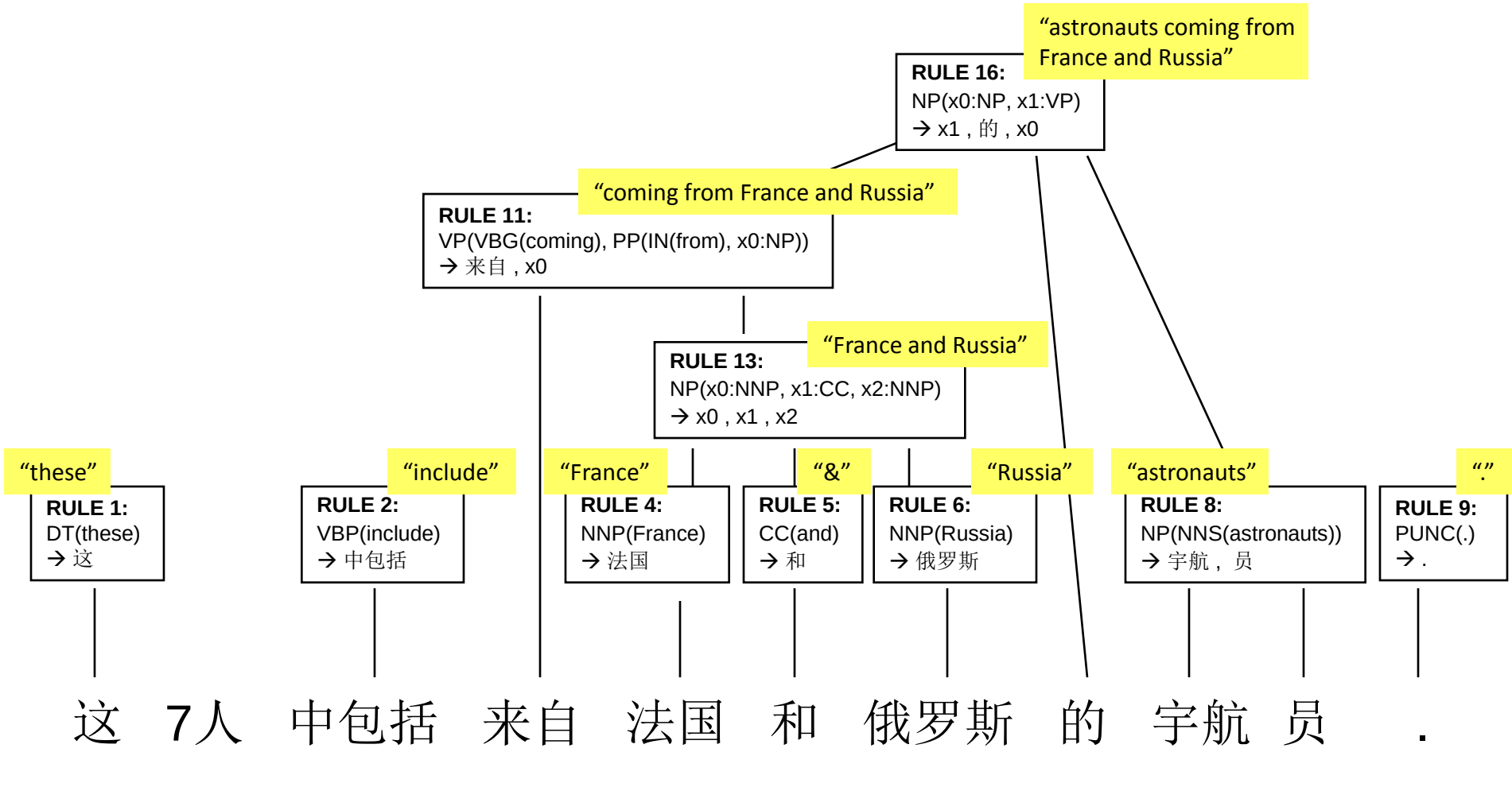
Syntax-Based Decoding

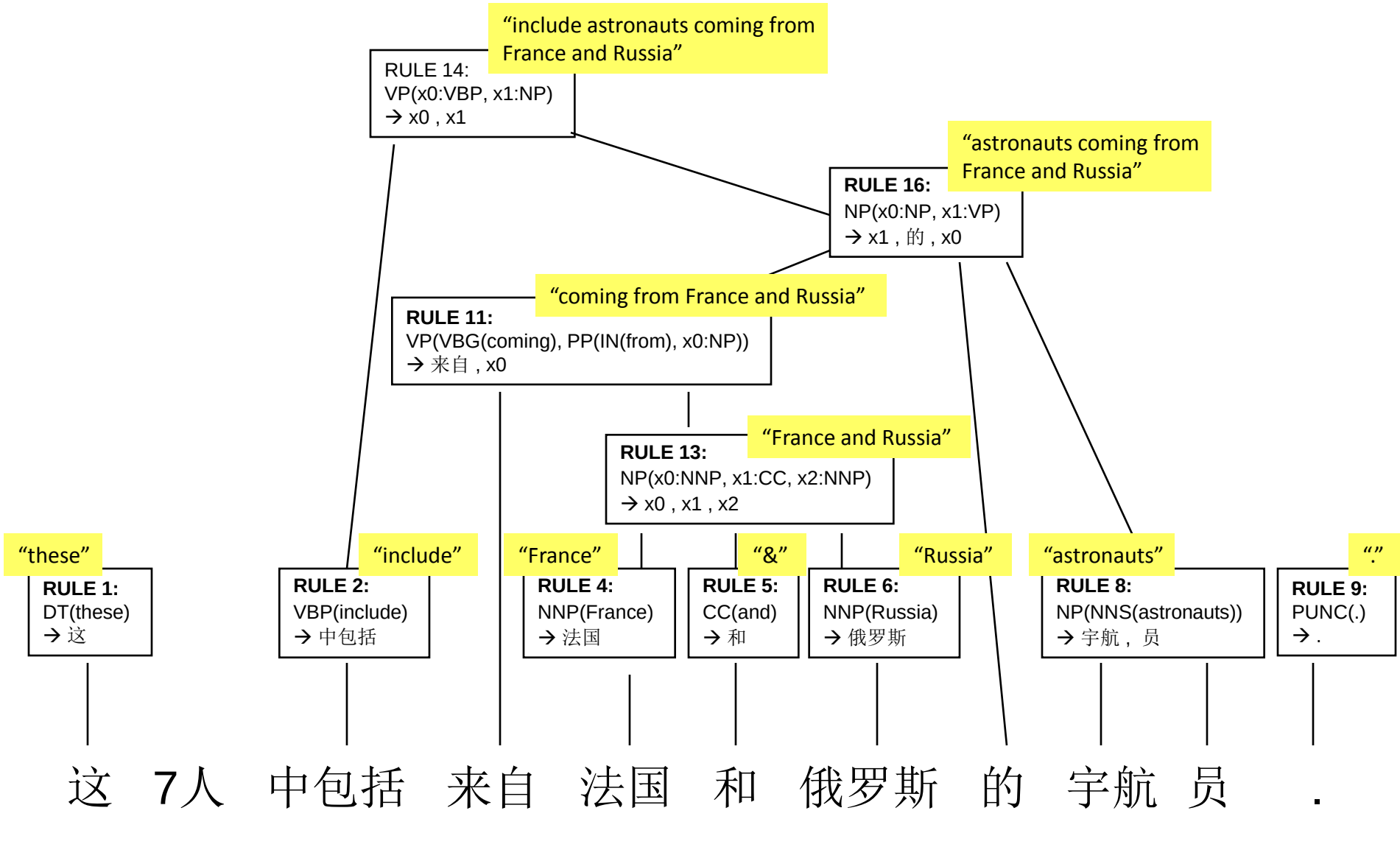


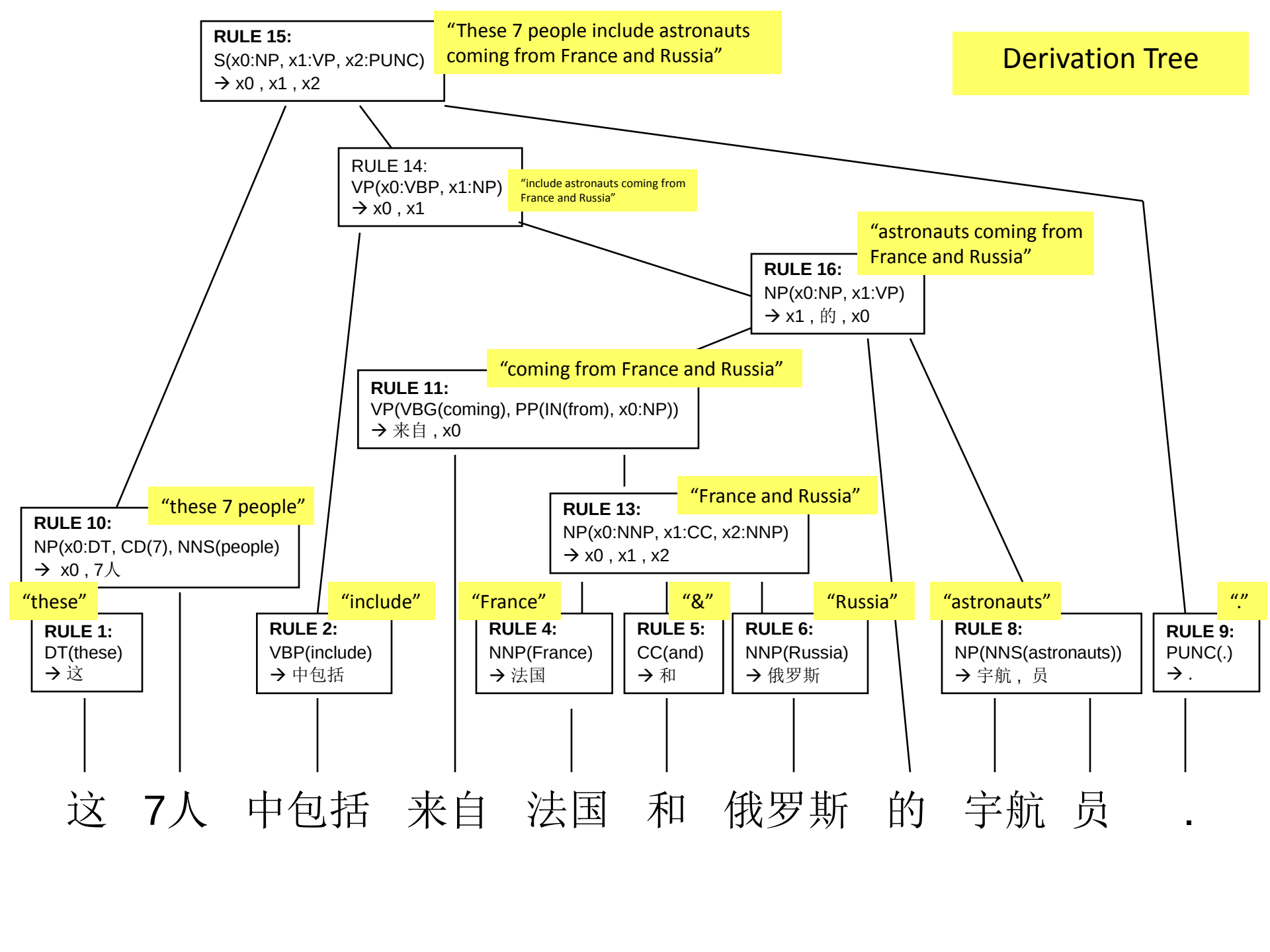
Syntax-Based Decoding



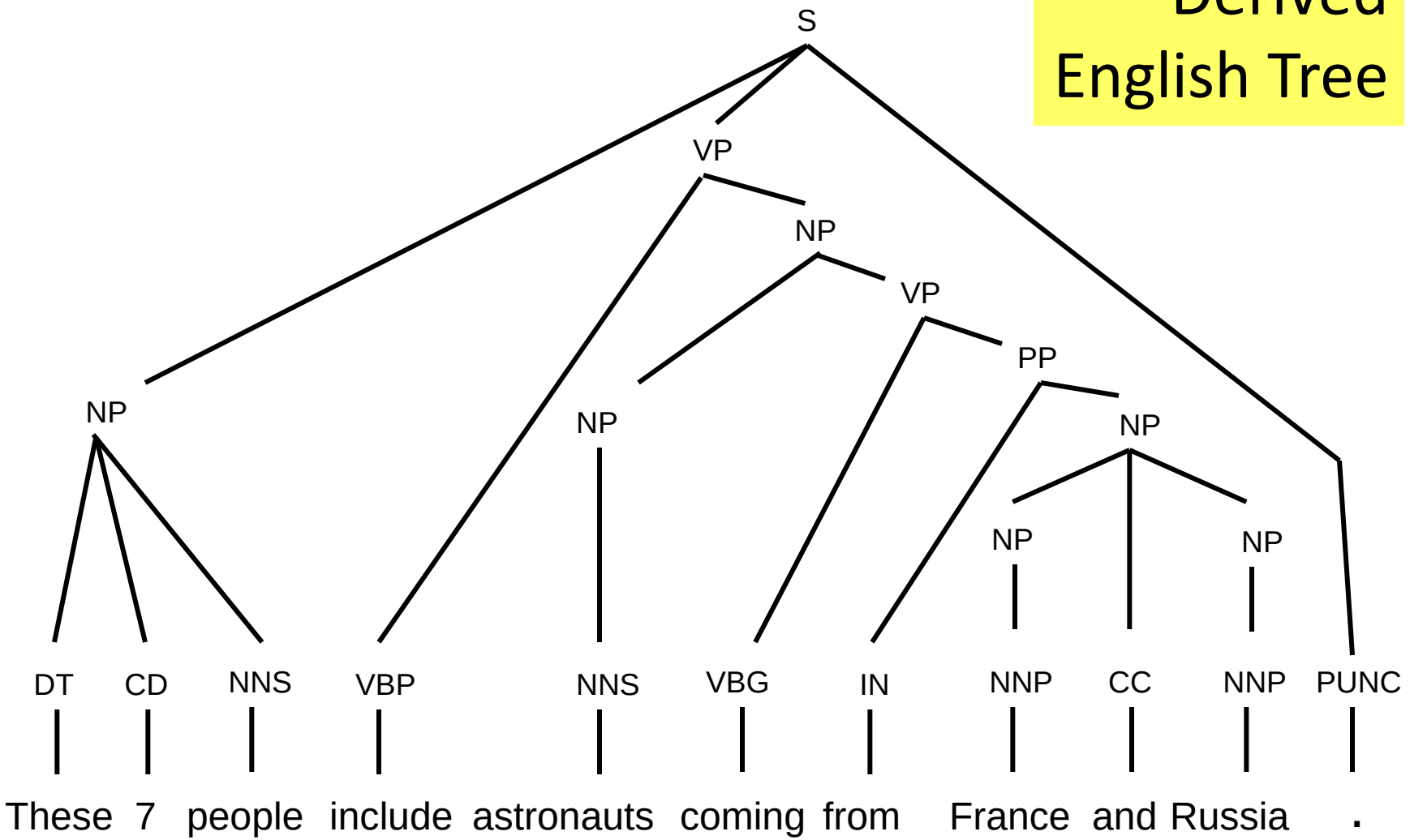
Syntax-Based Decoding



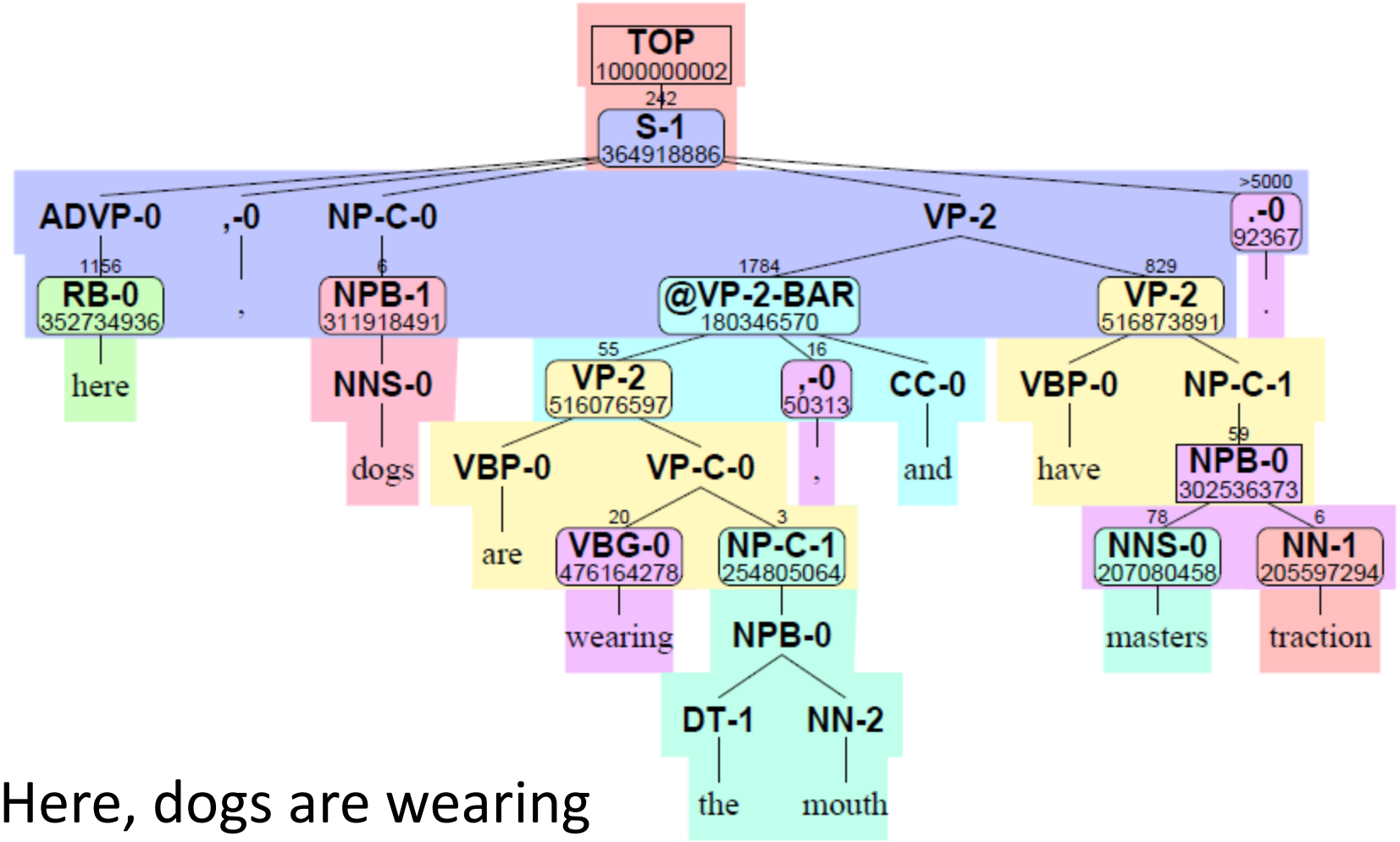

"include astronauts coming from France and Russia"
RULE 14:
VP(x0:VBP, x1:NP)
→ x0 , x1
"astronauts coming from France and Russia"
RULE 16:
NP(x0:NP, x1:VP)
→ x1 , 的 , x0
"coming from France and Russia"
RULE 11:
VP(VBG(coming), PP(IN(from), x0:NP))
→ 来自 , x0
"France and Russia"
RULE 13:
NP(x0:NNP, x1:CC, 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
RULE 15:
S(x0:NP, x1:VP, x2:PUNC)
→ x0 , x1 , x2
“These 7 people include astronauts coming from France and Russia”
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”
RULE 10:
NP(x0:DT, CD(7), NNS(people)
→ x0 , 7人
“these 7 people”
“these”
“include”
“France”
“&”
“Russia”
“astronauts”
“.”
RULE 1:
DT(these)
→ 这
RULE 2:
VBP(include)
→ 中包括
RULE 4:
NNP(France)
→ 法国
RULE 5:
CC(and)
→ 和
RULE 6:
NNP(Russia)
→ 俄罗斯
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 English Tree



在这里，狗 都 配 戴 嘴 套，并 有 主人 牵引 .



“Here, dogs are wearing the mouth, and have masters traction.”

Syntax-Based Summary

- Lots of technology…
- Critical part: underlying "generative story"

How does a tree in language X become a string in language Y?

"… through a sequence of steps of some tree automaton operating on Penn Treebank-style trees…"

Morphology?

- Syntax-based MT systems currently work at the word level
- Possible direction:
  - Morpho-syntactic translation models

How does a **morpho-syntactic tree** in language X become a **string of characters** in language Y?

Current Syntax-Based SMT

RULE BASE

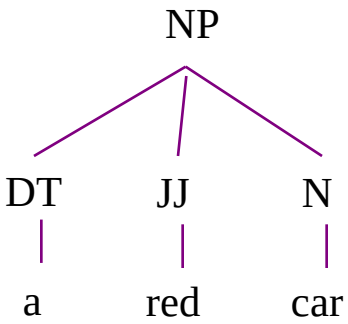
```
q.JJ(red) <-> rojo      q.JJ(green) <-> verde
q.JJ(red) <-> roja      q.JJ(green) <-> verdes
q.JJ(red) <-> rojos     q.N(cat) <-> gato
q.JJ(red) <-> rojas     q.N(cats) <-> gatos
q.N(car) <-> coche      q.N(moon) <-> luna
q.N(cars) <-> coches    q.N(moons) <-> lunas
q.DT(a) <-> un          q.N(light) <-> luz
q.DT(a) <-> una         q.N(lights) <-> luzes

q.NP(x0:DT x1:JJ x2:N) <-> q.x0 q.x2 q.x1
```

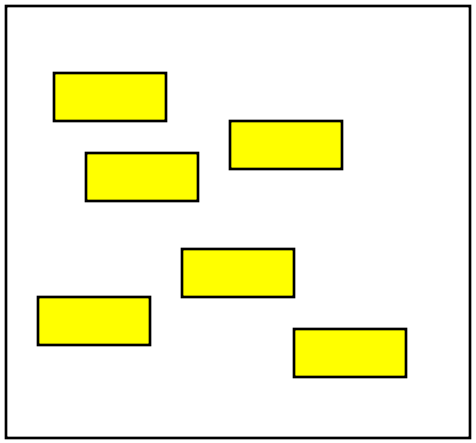
Very large wordform-to-wordform dictionary

Simple syntactic combination

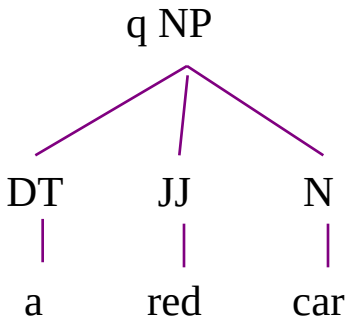
Current Syntax-Based SMT

Original input:



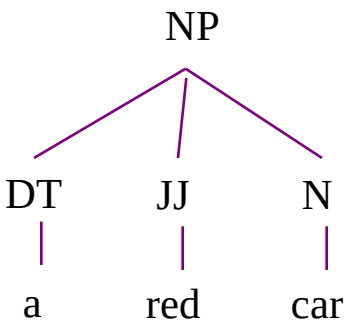


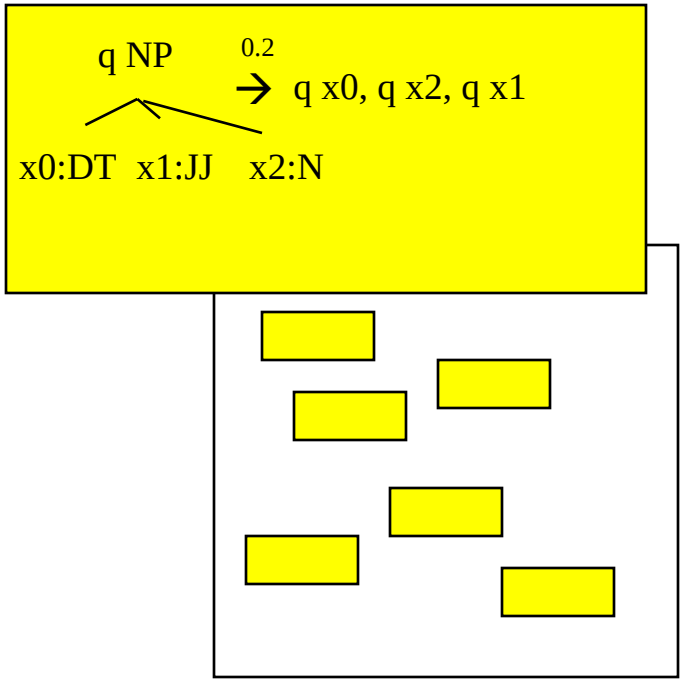
Transformation:



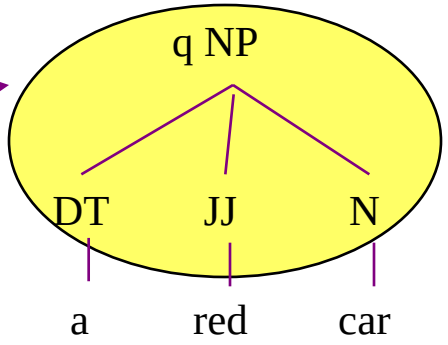
Current Syntax-Based SMT

Original input:



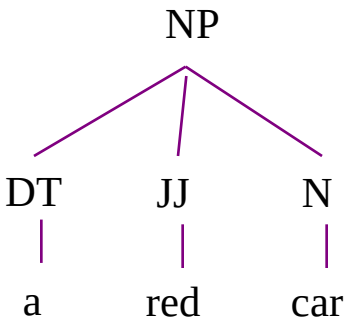


Transformation:

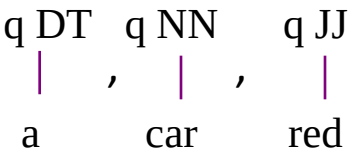


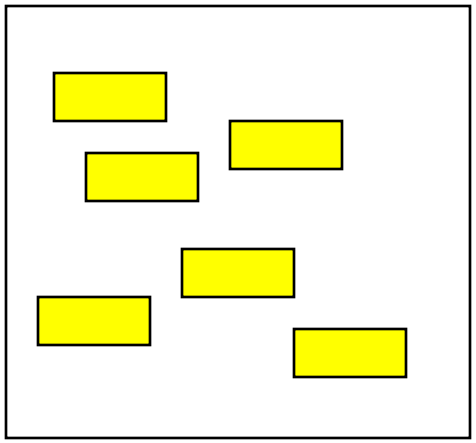
Current Syntax-Based SMT

Original input:



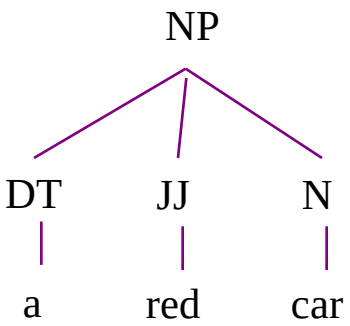
Transformation:

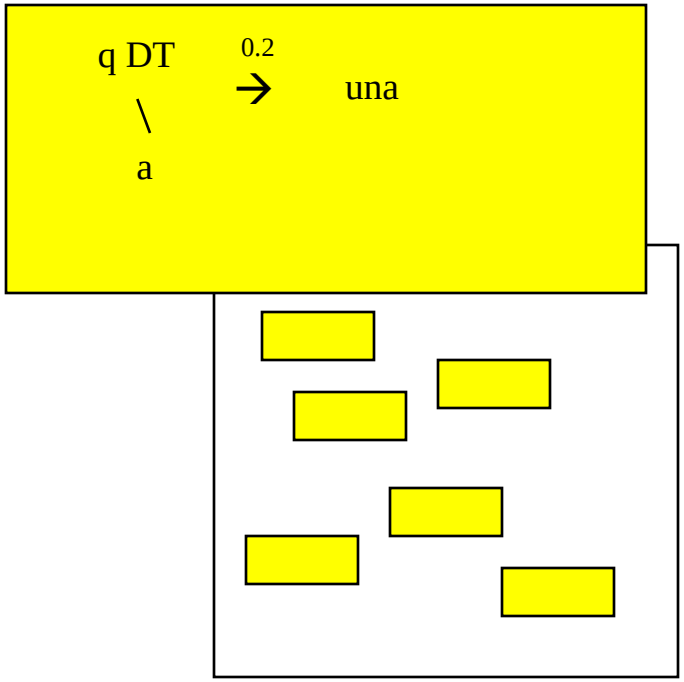




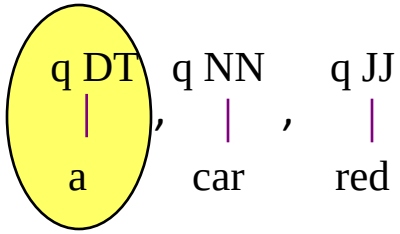
Current Syntax-Based SMT

Original input:



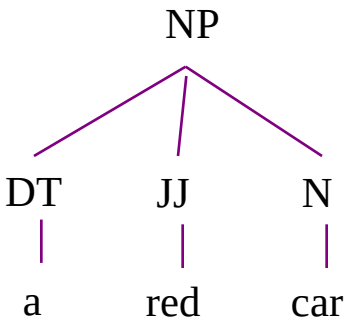


Transformation:

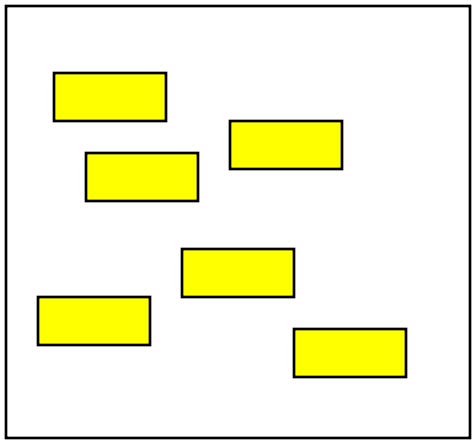


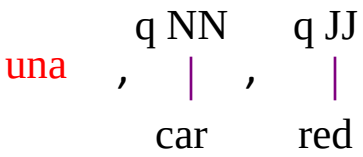
Current Syntax-Based SMT

Original input:



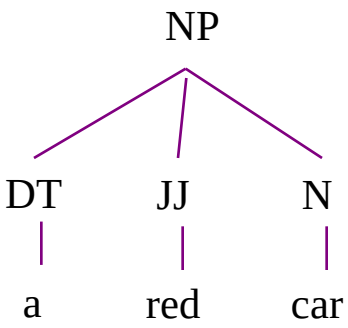
Transformation:



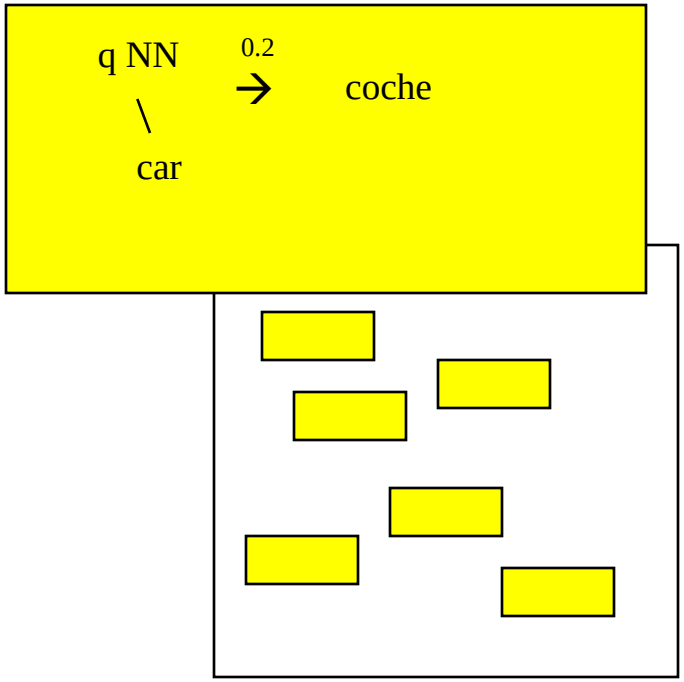


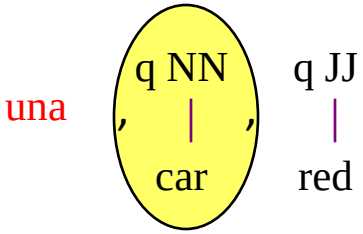
Current Syntax-Based SMT

Original input:



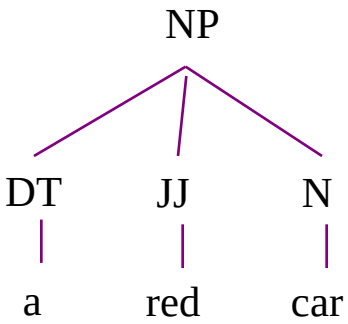
Transformation:

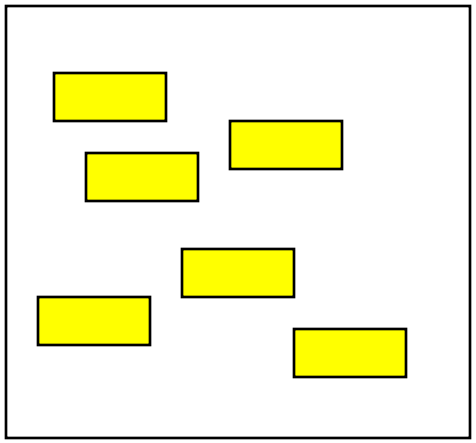




Current Syntax-Based SMT

Original input:





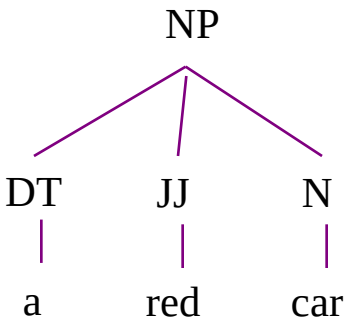
Transformation:

una , coche, q JJ

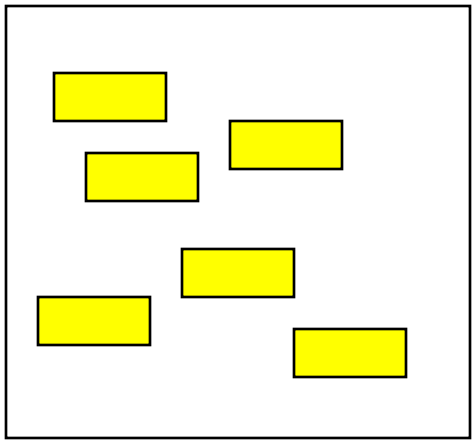
red

Current Syntax-Based SMT

Original input:



Transformation:



una , coche, rojas

Current Syntax-Based SMT

```
% echo 'NP(DT(a) JJ(red) N(car))' |
  tiburon -l -k 8 - sbmt.xlnts

OUTPUTS:

una coche rojo # 1.0      un coche rojas # 1.0
una coche rojas # 1.0     una coche roja # 1.0
un coche rojos # 1.0      un coche rojo # 1.0
una coche rojos # 1.0     un coche roja # 1.0
```

Overgeneration
(rely on language model
to catch problems)

Possible Morpho-Syntactic SMT?

```
RULE BASE

qjo.red <-> r o j              qje.green <-> v e r d e
qnsmasc.car <-> c o c h e      qnsmasc.cat <-> g a t o
qnsfem.moon <-> l u n a        qnesfem.light <-> l u z
qdmasc.a <-> u n               qdfem.a <-> u n a

qmasc.JJ(x0:) <-> qjo.x0 o     qmasc.N(x0:) <-> qnsmasc.x0
qmasc.JJ(x0:) <-> qje.x0       qmasc.N(x0:) <-> qnesmasc.x0
qplmasc.x0:JJ <-> qmasc.x0 s   qplmasc.N(x0: x1:pl) <-> qnsmasc.x0 s
                               qplmasc.N(x0: x1:pl) <-> qnesmasc.x0 e s
...
q.NP(x0:DT x1:JJ x2:N) <-> qdmasc.x0 _ qmasc.x2 _ qmasc.x1


% echo 'NP(DT(a) JJ(red) N(car))' | tiburon -l -k 1 - msmt.xlnts

OUTPUTS:

u n _ c o c h e _ r o j o  # 1.0 (no other outputs)
```

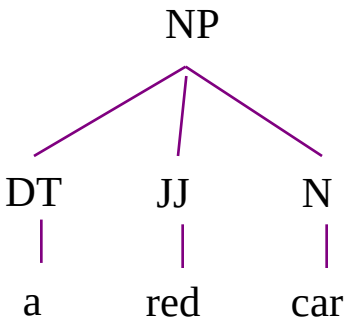
Compact root-to-root dictionary

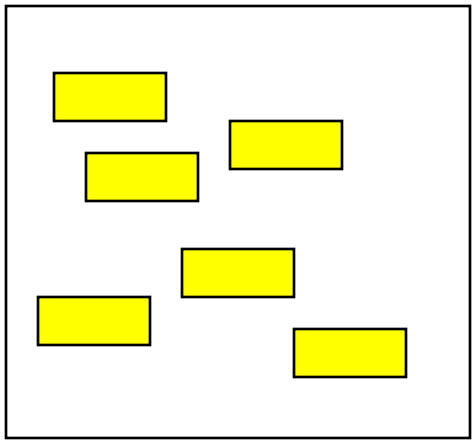
Morpho-syntax

Translation

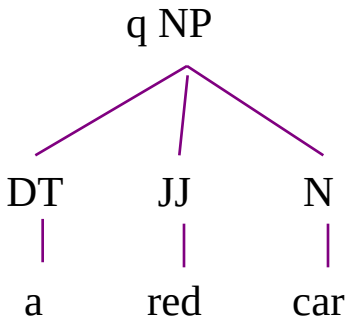
Possible Morpho-Syntactic SMT?

Original input:



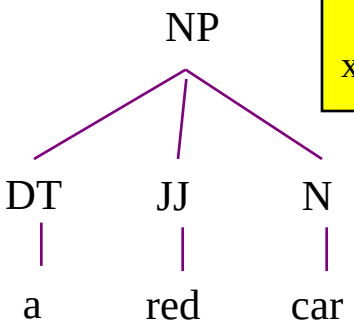


Transformation:

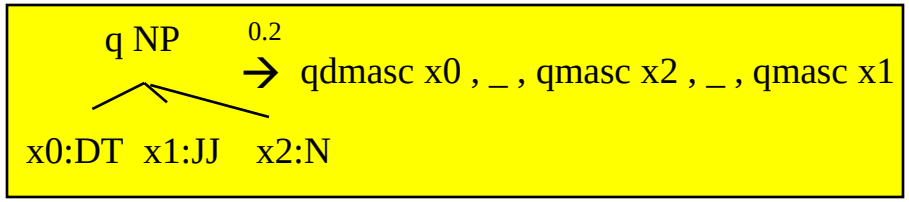


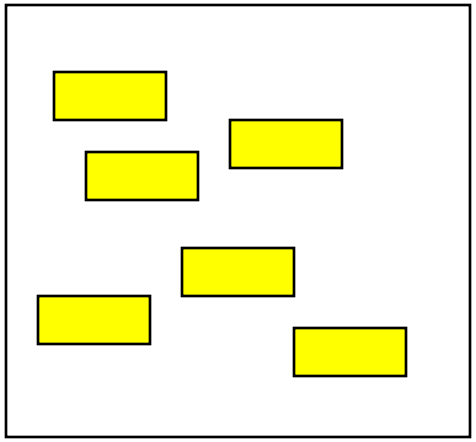
Possible Morpho-Syntactic SMT?

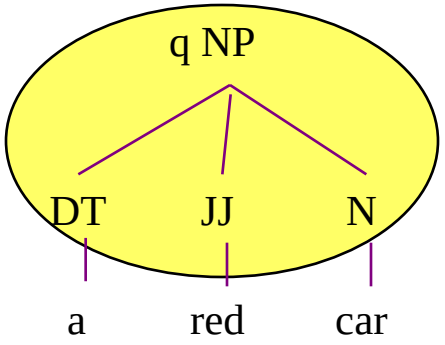
Original input:



Transformation:

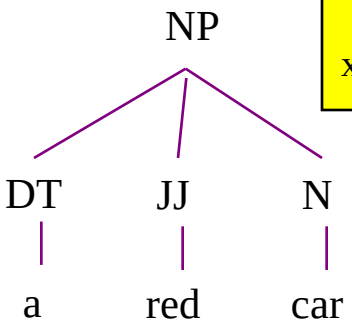




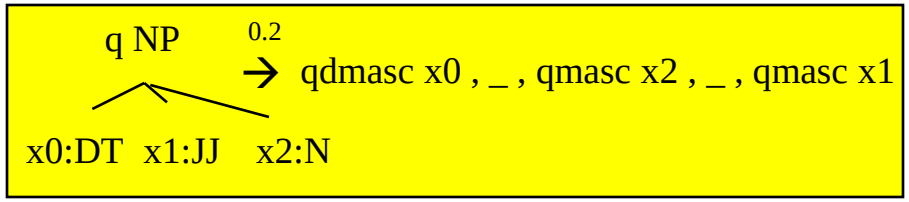


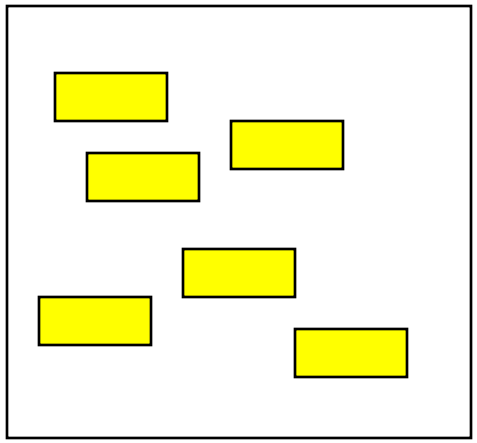
Possible Morpho-Syntactic SMT?

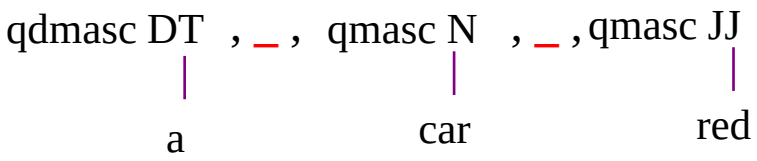
Original input:



Transformation:

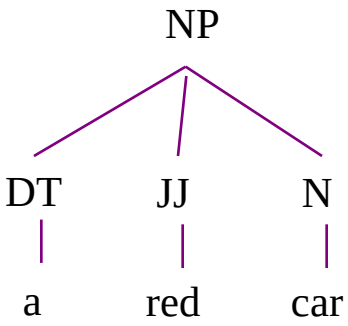




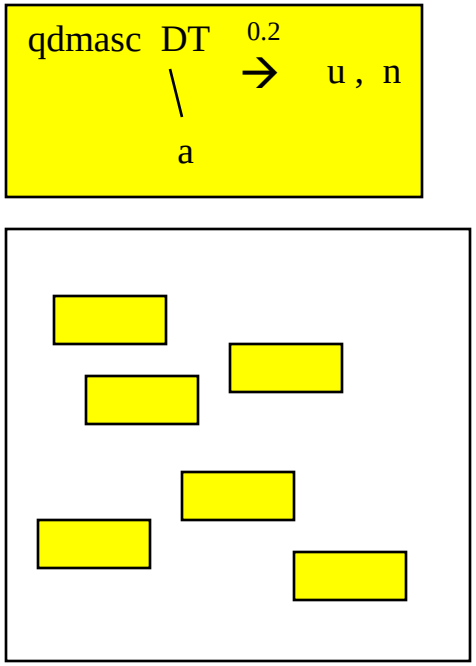


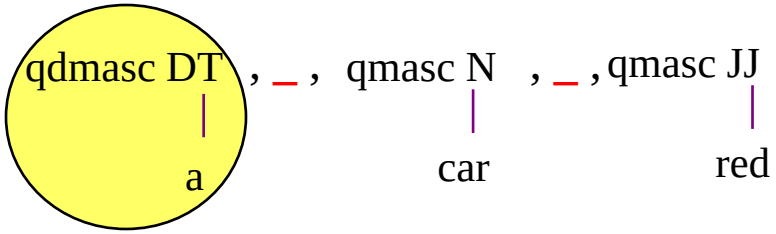
Possible Morpho-Syntactic SMT?

Original input:



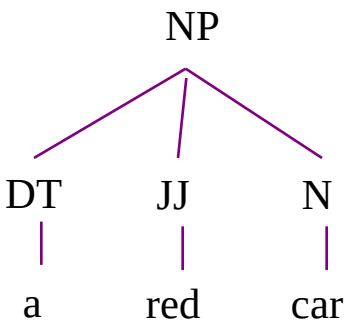
Transformation:

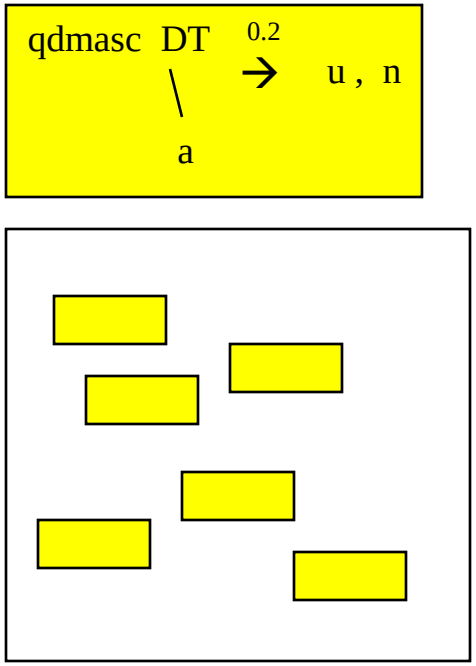




Possible Morpho-Syntactic SMT?

Original input:





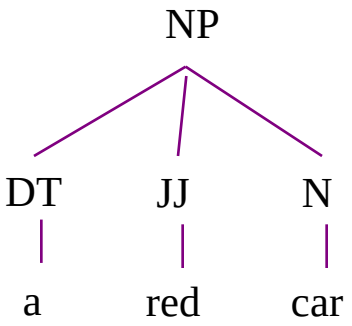
Transformation:

u , n , _ , qmasc N , _ , qmasc JJ

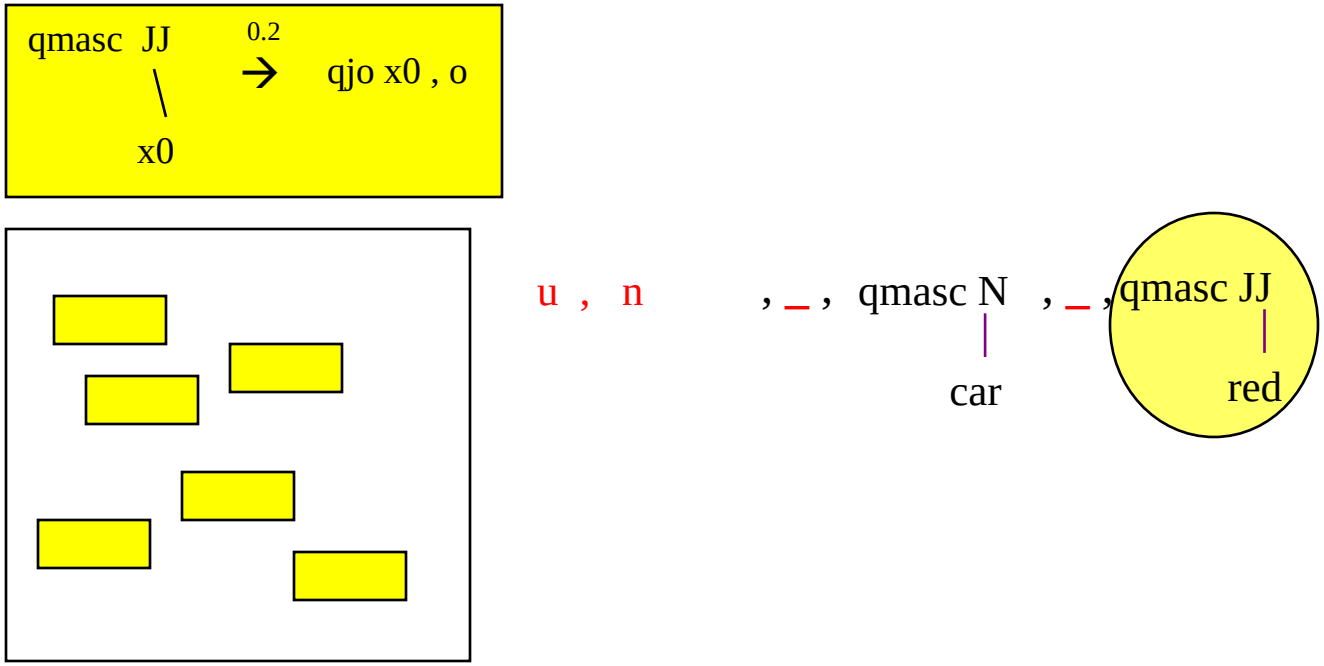
car red

Possible Morpho-Syntactic SMT?

Original input:

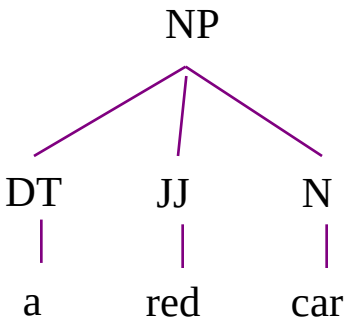


Transformation:

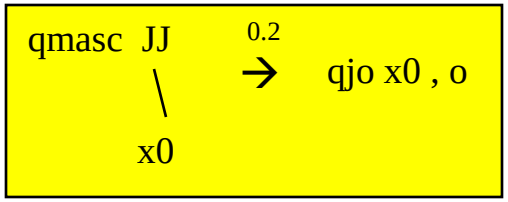


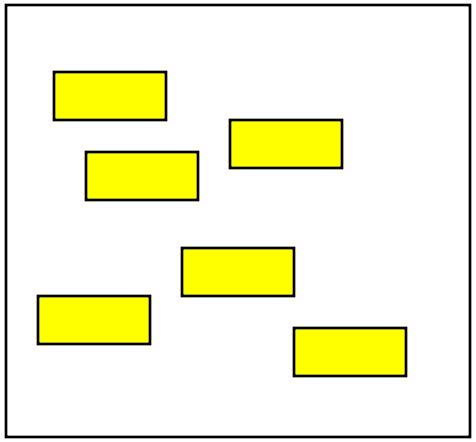
Possible Morpho-Syntactic SMT?

Original input:



Transformation:



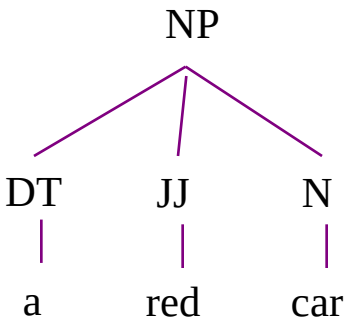


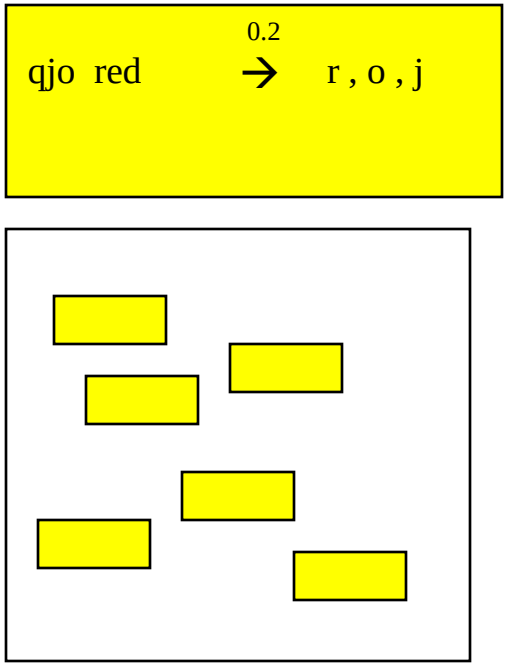
u , n , _ , qmasc N , _ , qjo red , o

car

Possible Morpho-Syntactic SMT?

Original input:





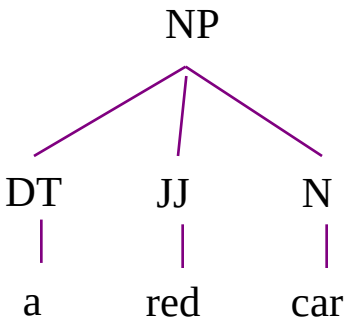
Transformation:

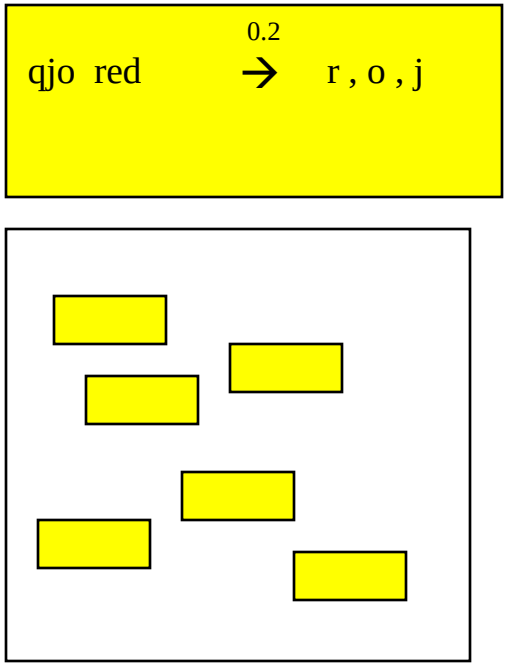
u , n , _ , qmasc N , _ , qjo red , o

car

Possible Morpho-Syntactic SMT?

Original input:





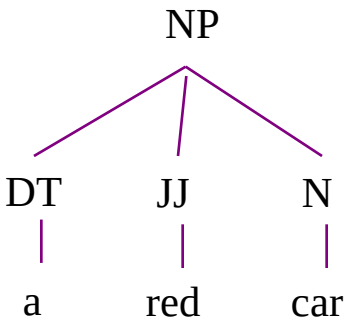
Transformation:

u , n , _ , qmasc N , _ , r , o, j , o

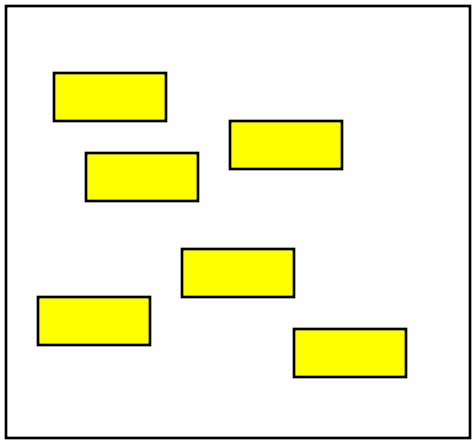
car

Possible Morpho-Syntactic SMT?

Original input:



Transformation:



u , n , _ , c , o , c , h e , _ , r , o, j , o

Possible Morpho-Syntactic SMT?

```
% echo 'NP(DT(a) JJ(red) N(car))' |
  tiburon -l -k 1 - msmt.xlnts
```

Translation

OUTPUTS:

```
u n _ c o c h e _ r o j o  # 1.0 (no other outputs)
```

Another Open Problem

- Syntax-based MT systems acquire rules from aligned, parsed parallel text
- But alignments are still done without syntax, via GIZA++ and are poor
  - Brown et al 1990 technology!

Another Open Problem

- Syntax-based MT systems acquire rules from aligned, parsed parallel text
- But alignments are still done without syntax, via GIZA++ and are poor
  - Brown et al 1990 technology!

Another Open Problem

- Syntax-based MT systems acquire rules from aligned, parsed parallel text
- But alignments are still done without syntax, via GIZA++ and are poor
  - Brown et al 1990 technology!

Use generalized morpho-syntactic framework to explain (align) parallel data at character level.

Like GIZA++, do it unsupervised.

Summary

- Some ancient history
- Some of what's happening in morphology and MT
- Some possible directions

thanks