

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 .

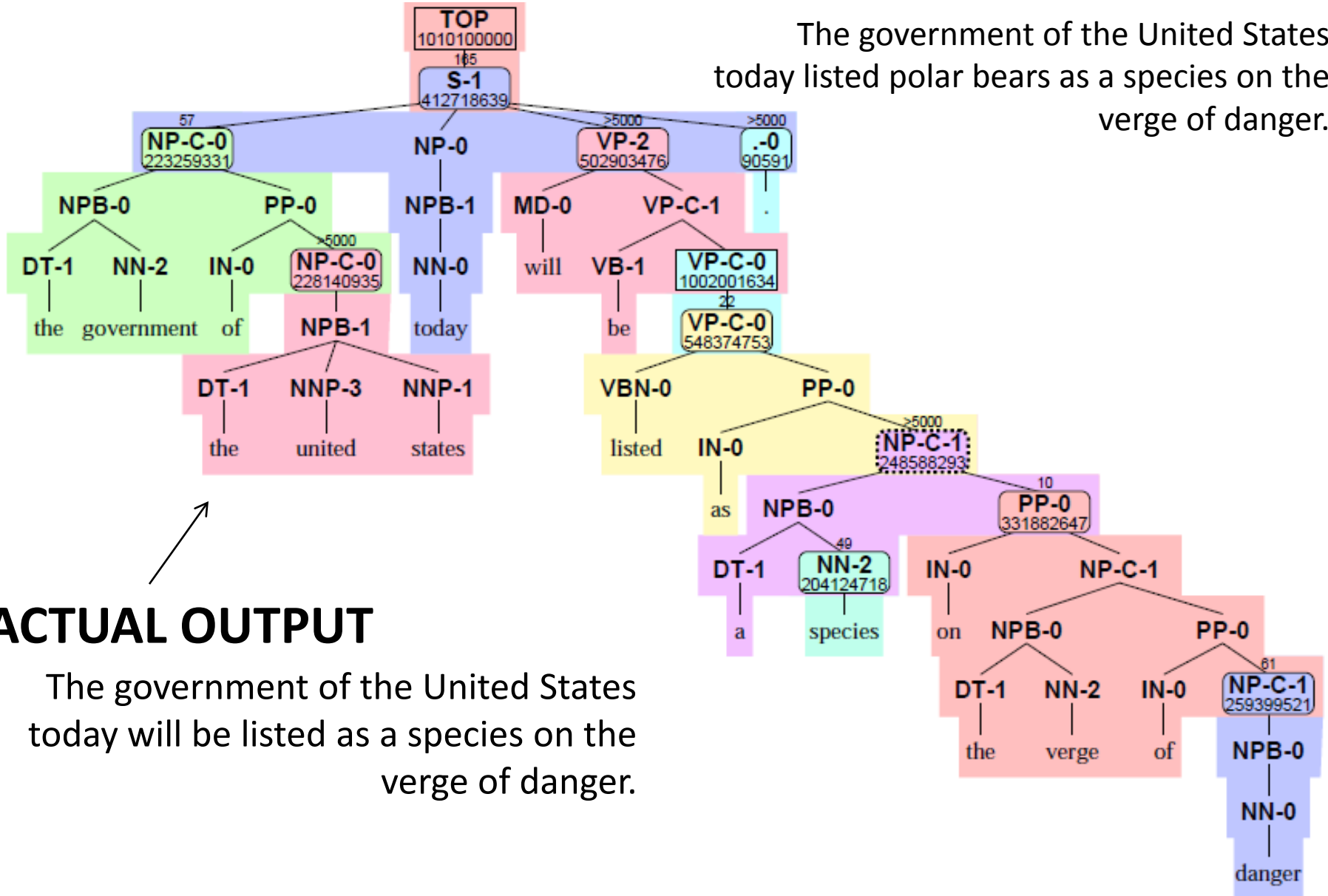
MT output:

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

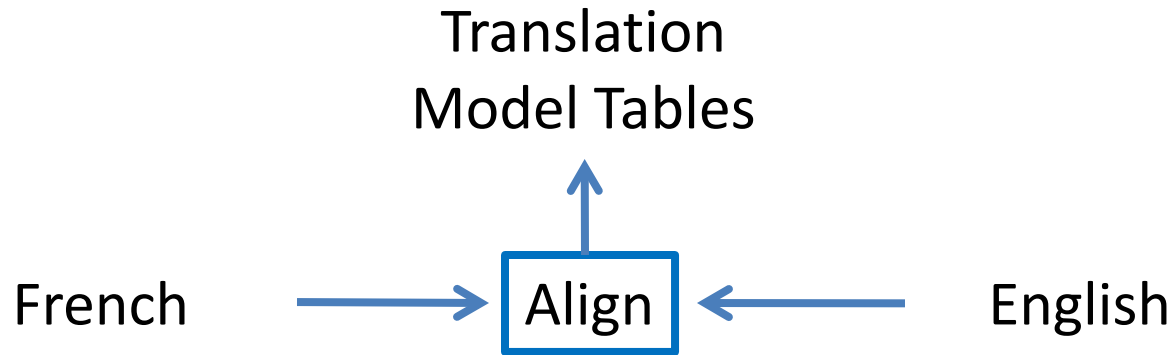
The government of the United States today listed polar bears as a species on the verge of danger.



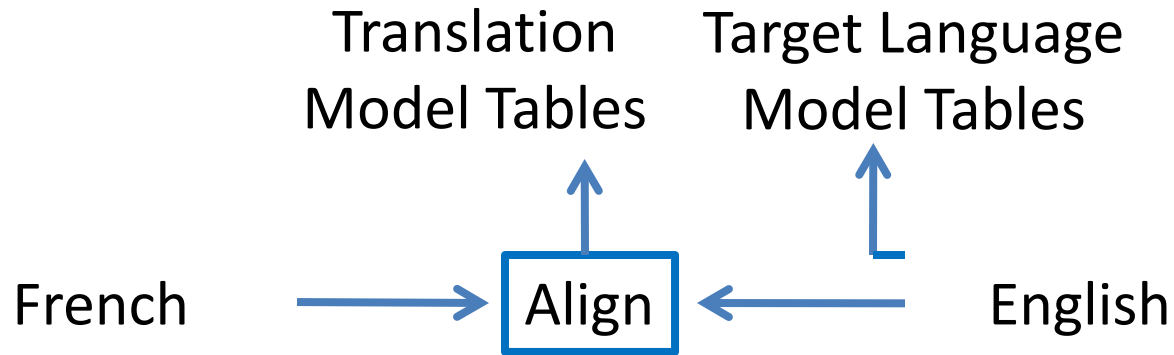
Need Lots of Knowledge

- 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?**

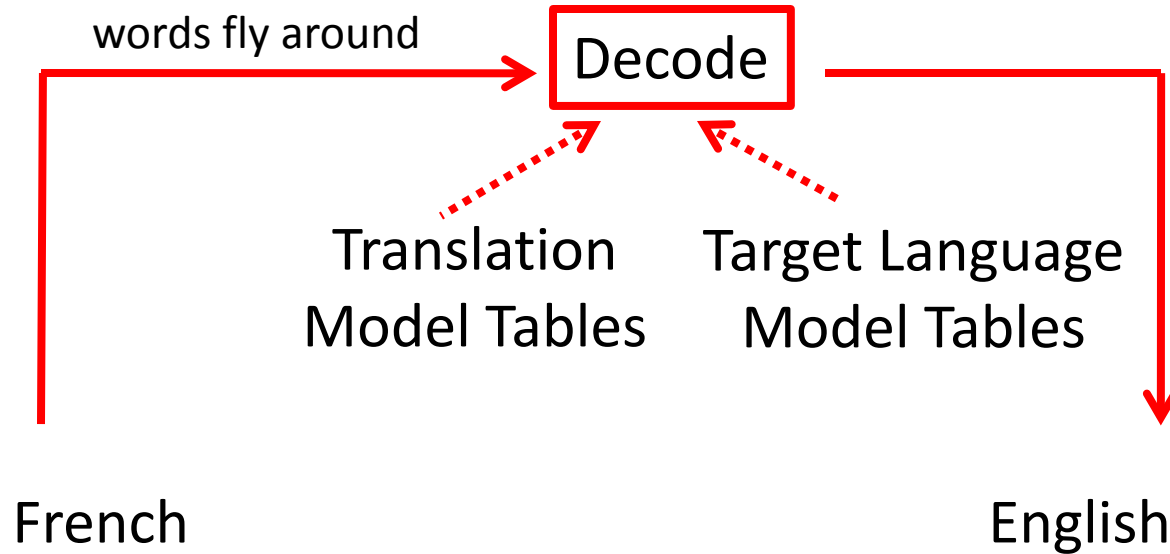
Brown et al, 1988



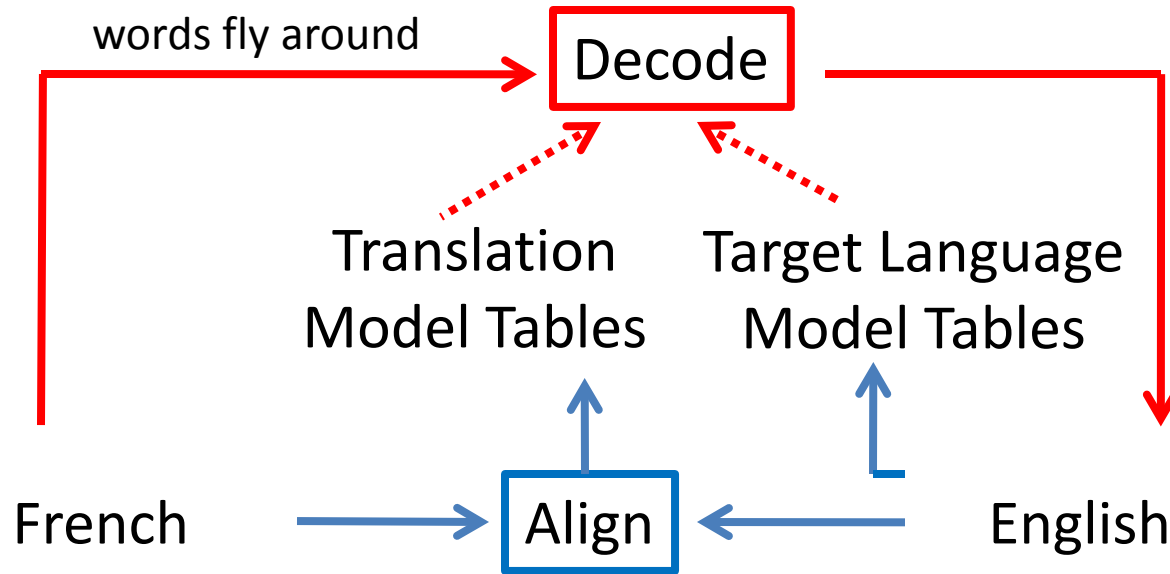
Brown et al, 1988



Brown et al, 1988

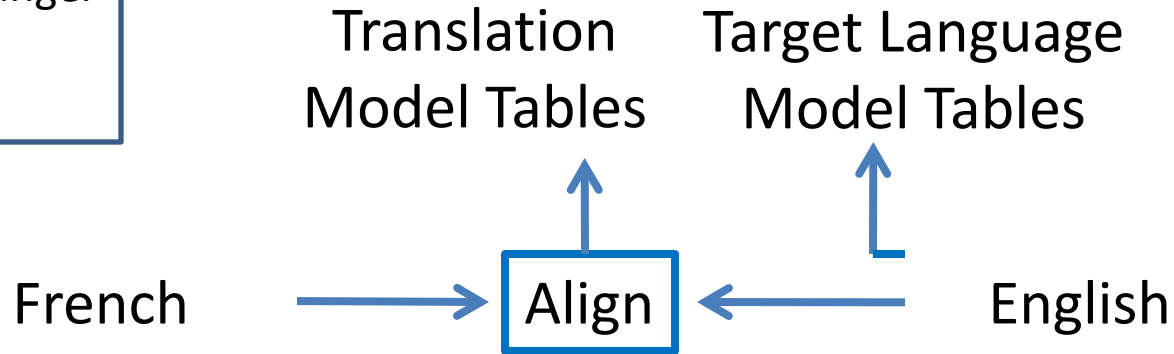


Brown et al, 1988

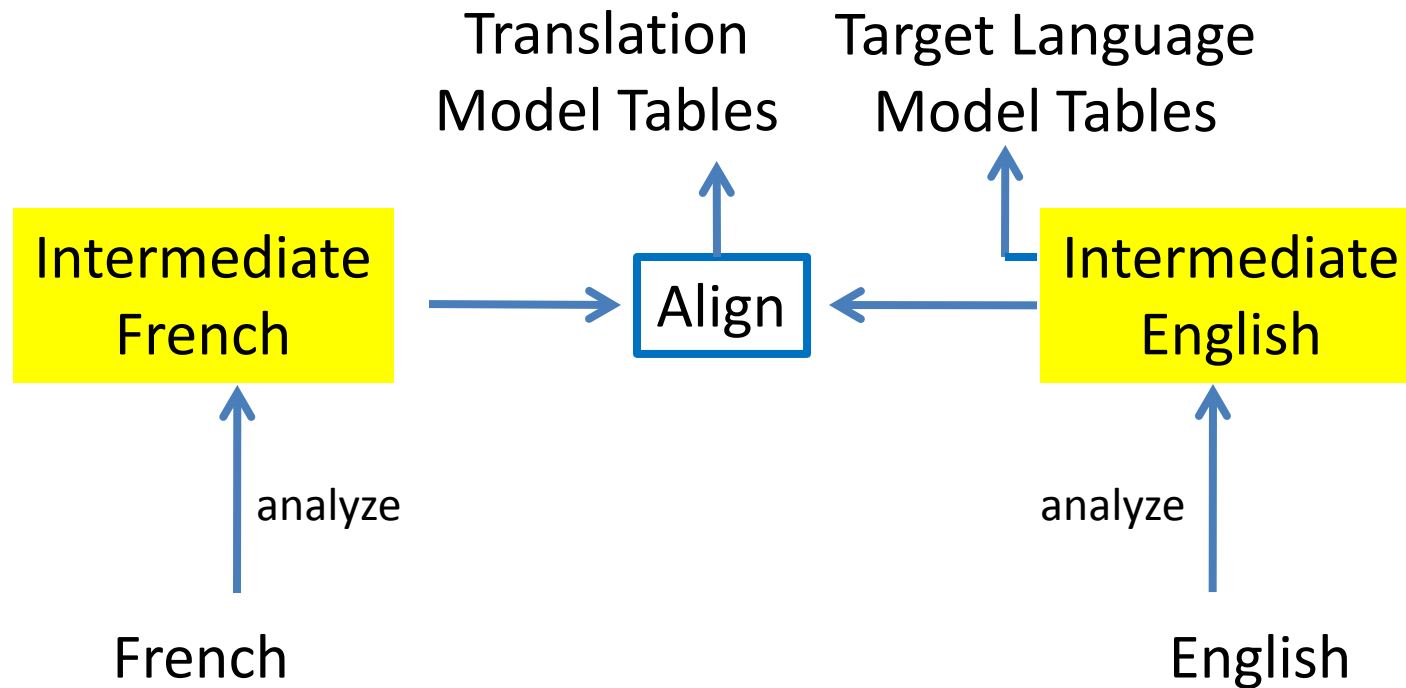


Brown et al, 1988

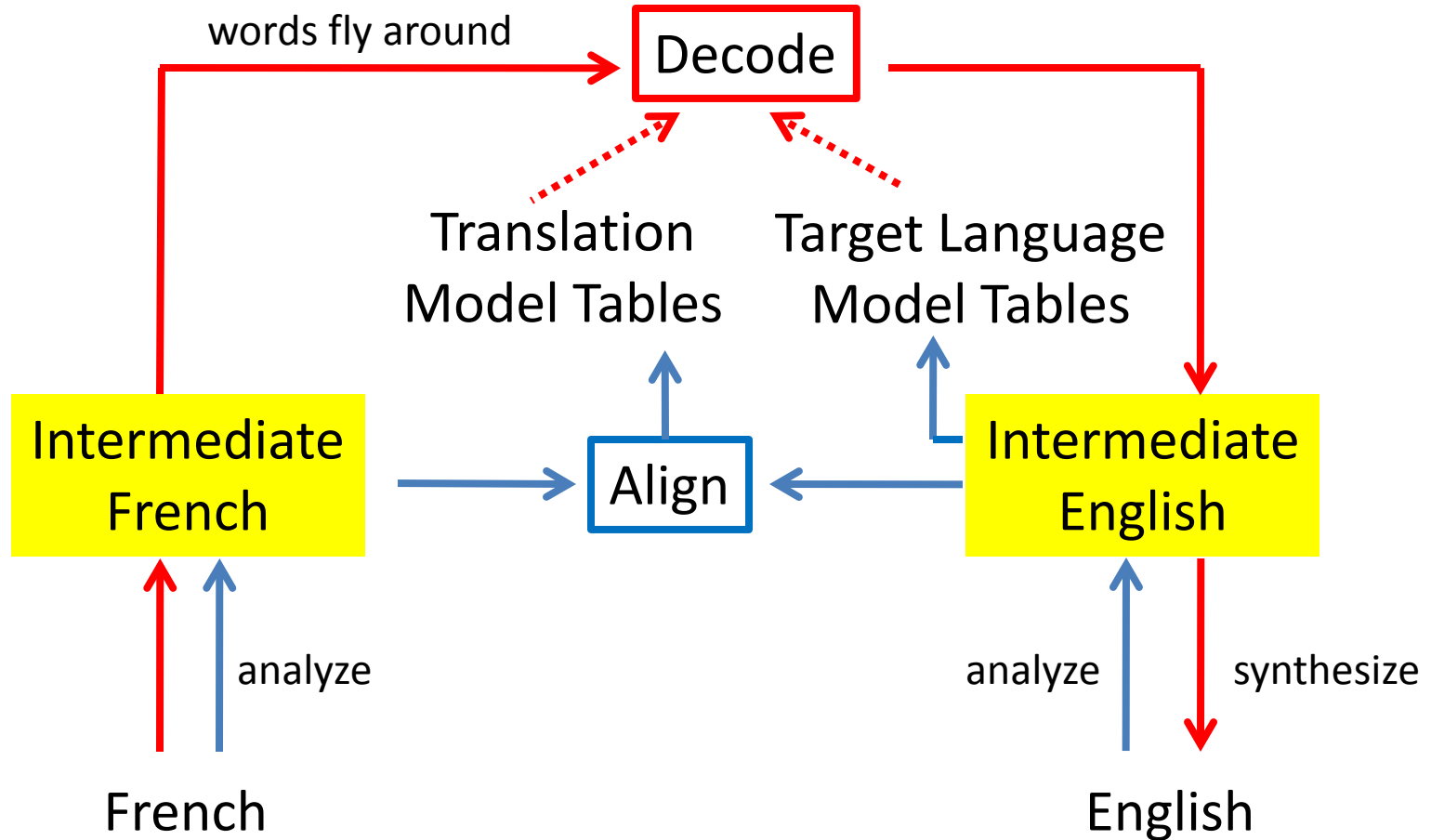
eats	mange
ate	mangé
eating	manger
eat	manger
...	



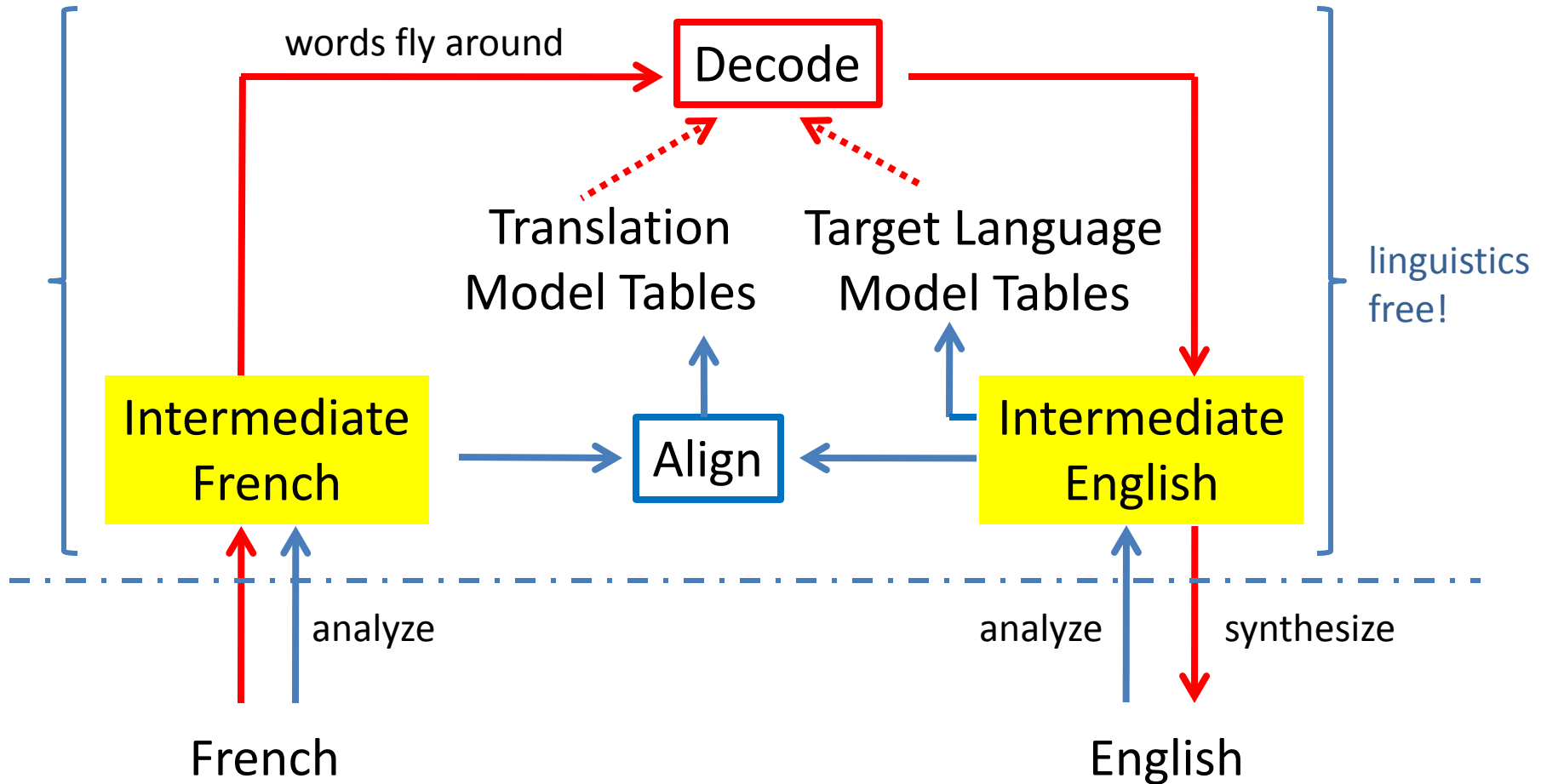
Brown et al, TMI 1992



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.

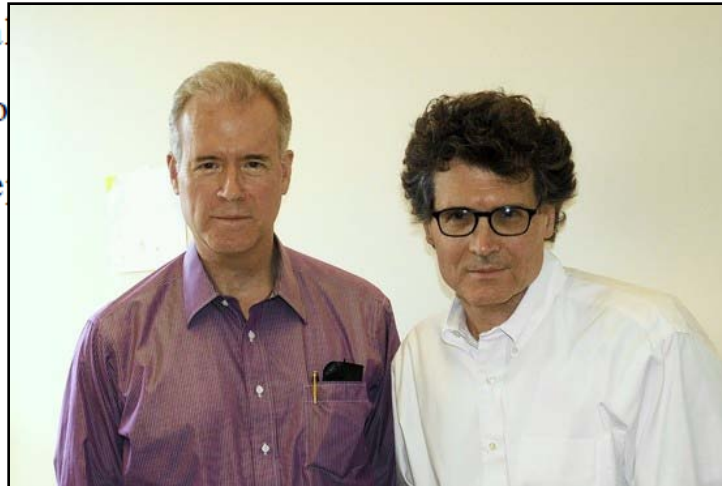
We estimated the parameters of the translation model for each system from a set of 1,778,620 pairs of French and English sentences from the Canadian Hansard data [1, 2]. Each of these sentences is 30 words or less in length. We tested both systems on the same set of 100 randomly selected Hansard sentences 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 statistical model. This is a costly undertaking, we have not made any of these constructions, and are unable to say which of the new analysis and synthesis steps



**Bob
Mercer**

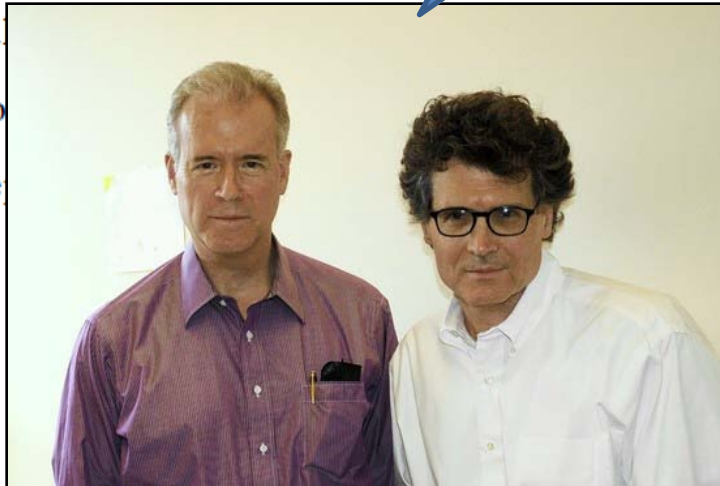
**Peter
Brown**

Brown et al, TMI 1992

In work of this type, it is desirable to be able to assess the performance to certain of the steps in the analysis or synthesis components of the various transformations. Making such an assessment requires 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 model. This is a costly undertaking, we have not made any of these constructions, and are unable to say which of the new analysis and synthesis steps

So long, suckers!

**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 na



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 best-paid employees in the world) employ batting averages, compiled from a history of a player's performance, to predict that a player will succeed in his next appearance at the plate. They use these statistics to manipulate their lineups accordingly. Wall Street speculators (who rank among the best-paid financial statisticians) build models based on past market data to predict tomorrow's fluctuations and alter their portfolios to cash in 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.

Time to switch jobs!

The na

ant progress toward increasing the



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)

Live News Translation

The screenshot displays the VIRAGE interface for live news translation. It features a central video player showing a news anchor, with a search bar on the left and a list of news bulletins. The right side shows a translation window with English text and its Arabic equivalent. The interface includes navigation controls like 'Previous Clip', 'Storyboard', and 'Next Clip'.

news broadcast

foreign language speech recognition

English translation

searchable archive

Administration

Search: bin laden

Showing: 1-5 of 9

NWI : International Newsfir... 48.49%

Thu Feb 26 13:14:20 EST 2004
- Michael Eisner
- Rio De Janeiro, Mexico City, Madrid
- United Technologies, Walt Disney, Gene Motors

Aljazeera : News Bulletin 96.67%

Tue Feb 24 19:00:20 EST 2004
حسني مبارك أمامه بن لادن عبد الله بن عبد العزيز -
المرحوم بخلاف القاضية -
الجزيرة تنظيم القاعدة

Aljazeera : News Bulletin 56.62%

Tue Feb 24 19:02:20 EST 2004

Aljazeera : News Bulletin 96.67%

Wed Feb 25 12:14:20 EST 2004
أسامة بن لادن يحيى باللقطة شوية سلطان -
الولايات المتحدة والجزائر القاعدة -
تنظيم القاعدة وزارة الدفاع الإسرائيلية -
الجزيرة الأمم المتحدة

5 12:00:20 EST 2004
أسامة بن لادن يحيى من
الولايات المتحدة
الجزيرة الأمم المتحدة

Alerts Projects

Welcome, Sys

Show: Words

Stop Highlighting Translate

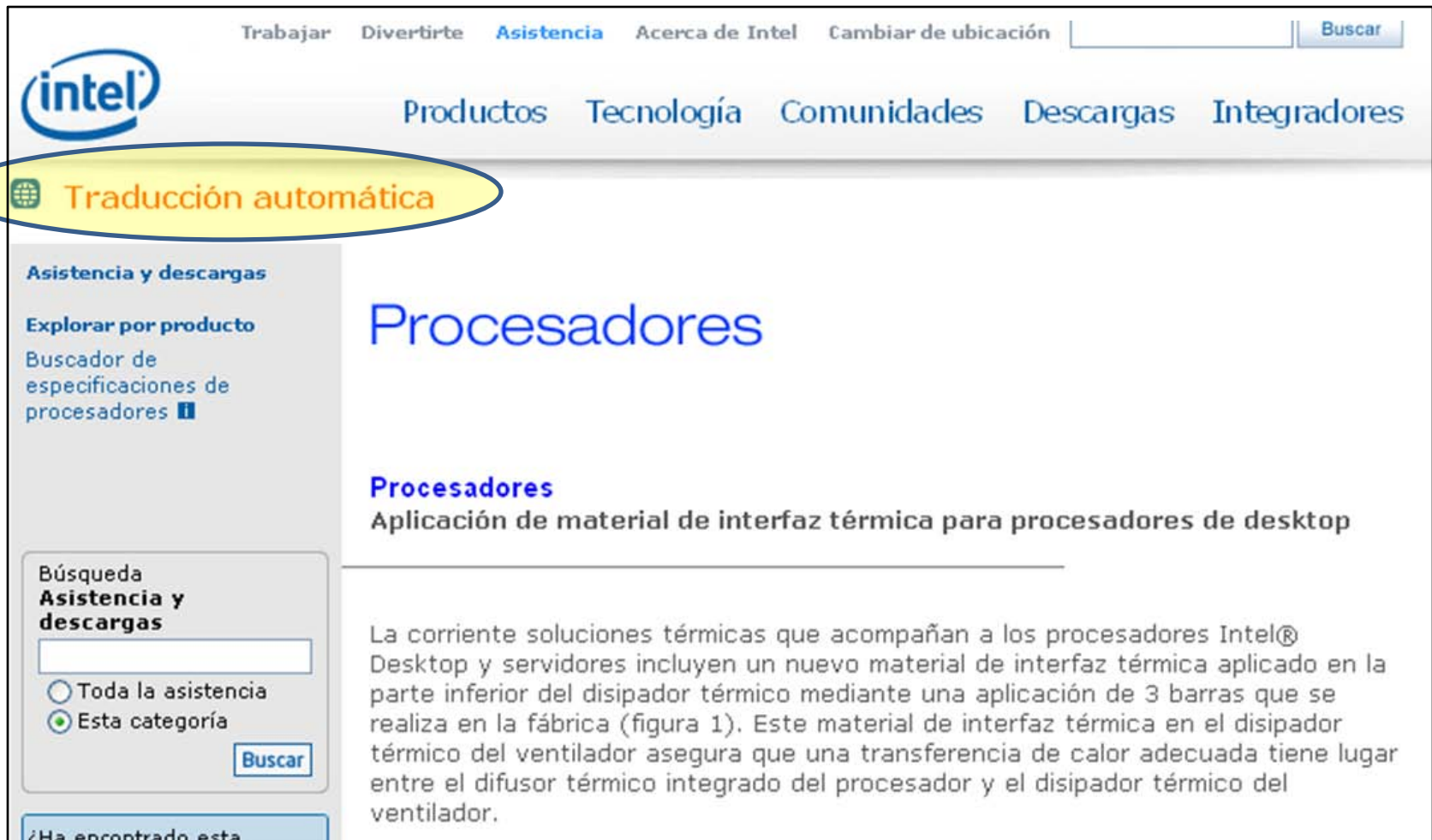
which touched several schools and villages, a number of victims of the first keep mwqtan [موقتان] to While emerged with proceeds from the final, had been sent from within the island, one of the centres to shelter Al-Hussein City Pakistani military spokesman confirmed the two, General Sultan, said the Ouane northwest Pakistan carried out by the army against the Taliban and al-Qaeda had expired today and resulted in the arrest of a number of foreigners without specify the nationality of those on both, but he denied the presence of senior leaders of al-Qaeda in the area of aggression which returned to the front events of a new force has undergone intensive Pakistani army operations in search of al-Qaeda elements

التي نهائية أرسل
أه منبهة الحسين
شوية سلطان
في نفذها الجيش
وم وأسفرت عن
جنسية المقوض
عليهما بحثه نعي وجود رعاء حرب من تنظيم القاعدة في
المنطقة بمنطقة عوان التي عدلت إلى واجهة الأحداث من
جديد شهدت عمليات مكثفة لقوات الجيش الباكستاني بحثا
عن عناصر تنظيم القاعدة والتي نشنته السلطات
الباكستانية بوجودهم في هذه المنطقة المجاورة للحدود مع
أفغانستان حيث تنكرت الهجمات على القوات الأميركية
الموجودة هناك شهيد عوان أكورا للجزيرة هم القوات
الباكستانية عددا من نوبت السكان المحليين بتهمة إيواء
ومساعدة عناصر القاعدة بينما استهدفت المصارف
الرسمية الباكستانية وجود كبار زعماء القاعدة في هذه
المنطقة كما شهدت المصارف على مشاركة المروحيات
المسكوبة والأسلحة الثقيلة خلال العملية التي تم خلالها
القبض على بعض الأجانب دون تحديد جنسياتهم الرقيق
بلورثها للتعرف ليبيا اعظت خلال العملية بعض
الأشخاص من بينهم اجانب وثانكت من وجود بعض
الاجانب في المنطقة من خلال الوثائق والجوازات التي تم
التحور عليها خلال العملية ويأتي اطلاق هذه العملية بعد
عمليتين طرقت والسيدان التي اعظت عنها القوات
الأمريكية داخل أفغانستان وهو ما دفع بعض المحللين إلى
الاعتقاد بأن هذه العمليات قد تسفر عن القبض على أسماك
كبيرة بداية كيشن وقت منح عمليات القبض على أسامة
بن لادن والملا محمد عمر وهناك فرص كبيرة لذلك
الذي أهداف من يقومون بهذه العملية اسلام إباد المنقلب
نوب انها وجزارات الداخلية والولاية تحيض أياها صحبة
في ظل الملفات الشائكة التي تنتظرها ما بدا من أزمة تسريب
تفريعات نووية إلى الخارج وانتهاءا بتفاديها في
الحرب على ما يوصف بالإرهاب يكاد يفتق معظم
المرافقين السياسيين على أن العملية العسكرية الباكستانية
إنما جاءت المستوطنات الأميركية خصوصا في ظل
تسريبات أميركية عن عدم جدية باكستان بالتعاون للحرب
على ما يوصف بالإرهاب أحمد زيدان الجزيرة اسلام
أباد وقد أكد وزير الخارجية الباكستاني خورشيد محمود
تصوري أن العملية الأخيرة التي قام بها الجيش الباكستاني

Previous Clip Storyboard Next Clip

0:18:32:04 00:18:32:28 00:18:40:10 00:18:45:25 00:18:59:27 00:19:22:17 00:19:24:11 00:19:26:16 00:19:26:23 00:19:29:13 00:19:34:22 00:19:59:09 00:2


Online Help



Trabajar Divertirte **Asistencia** Acerca de Intel Cambiar de ubicación

intel

Productos Tecnología Comunidades Descargas Integradores

 Traducción automática

Asistencia y descargas

Explorar por producto

Buscador de especificaciones de procesadores

Búsqueda

Asistencia y descargas

Toda la asistencia

Esta categoría

¿Ha encontrado esta

Procesadores

Procesadores

Aplicación de material de interfaz térmica para procesadores de desktop

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.

Plaza

is

009

n West

os

ity -

009

en Los

009

is en Los

009

is en Los

ngeles

is

009

n West

Traducción automática ?

Original en inglés

Traducido por: 
Language Weaver

"Uno de los mejores sitios de sushi en el área de los Ángeles"

Sushi Katsu-ya



Sworn2Fun

Los Angeles

3 ene 2007

1/1 considera esta crítica muy útil

Éste es un lugar favorito de almuerzo de la mía. La ubicación es estupenda Sherman Oaks también, pero yo preferiría un poco este, quizás porque es tu clásico en el un-strip-sushi-gourmet de centro comercial. Como tal, presume la decoración mismo no de la decoración, pero ¿qué importa? Lujoso lugares de sushi son normalmente no muy bueno (algunas excepciones, pero en la verdad demasiado a menudo). No te pierdas los bollos de cangrejo o el arroz crujiente, pero es todo bien. El servicio es amable, pero está siempre lleno, reservarlas, reservarlas, reservarlas!

Esta crítica es la opinión subjetiva de un miembro de TripAdvisor,

Por l

Hotel

HOTEL

El Pati

Beverl

Garlan

Holida

Sports

Lodge

Colony

Hollyw

Best W

Mikad

TODOS

*Preci

Resta
puntu

RESTAL

Amir's

Firefly

Sushi |

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

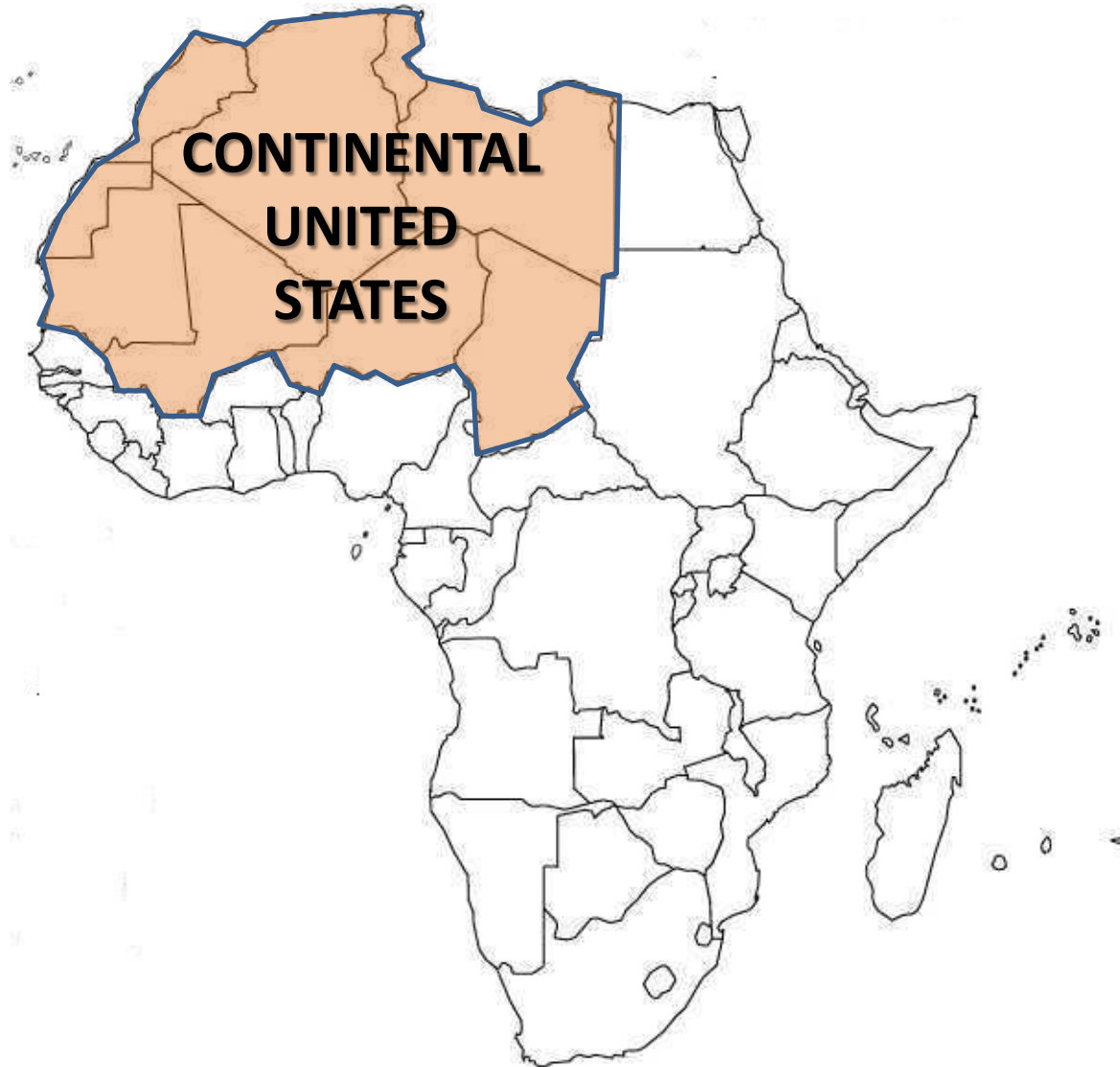
Not Solved!

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.

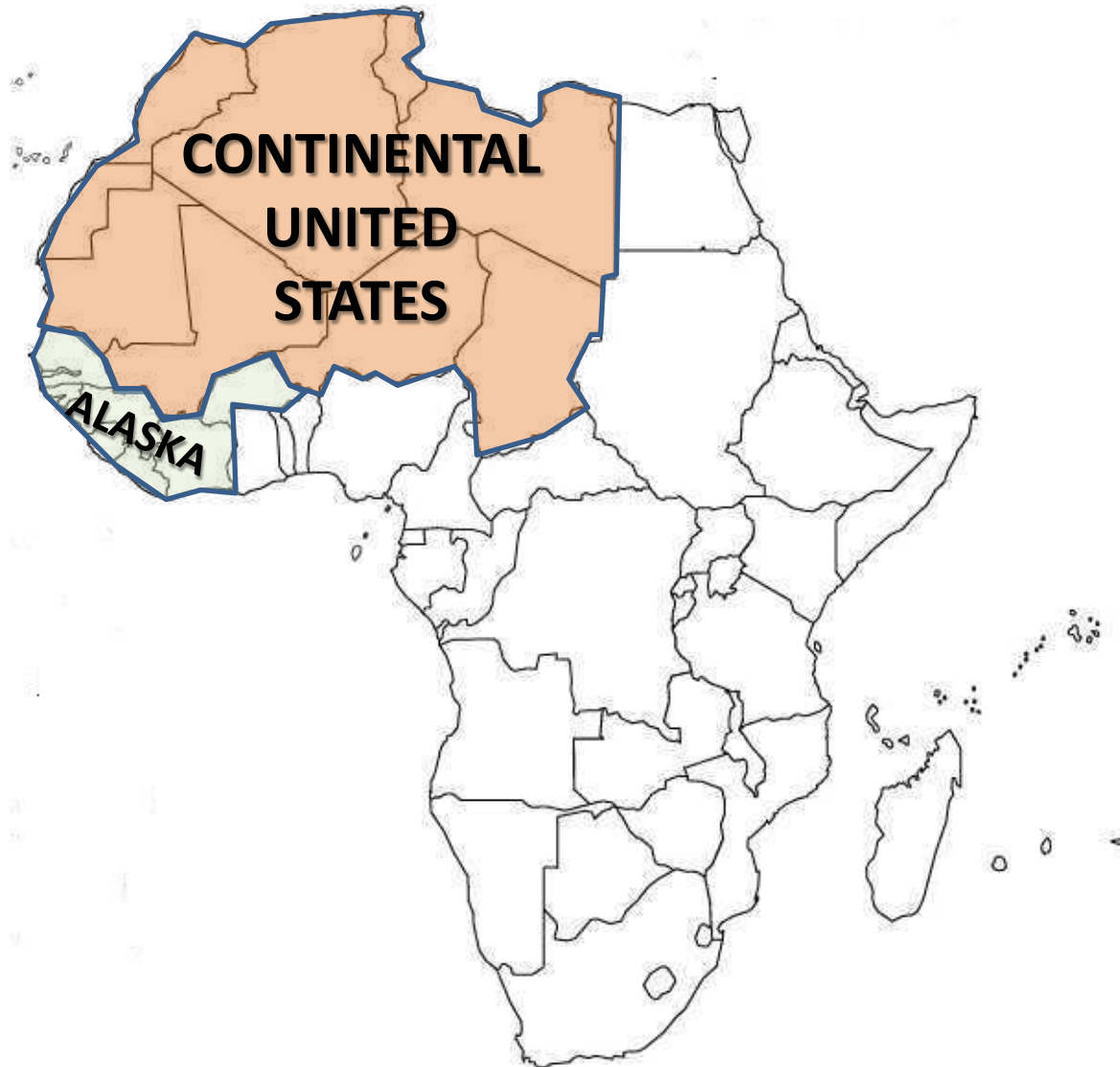
African Languages



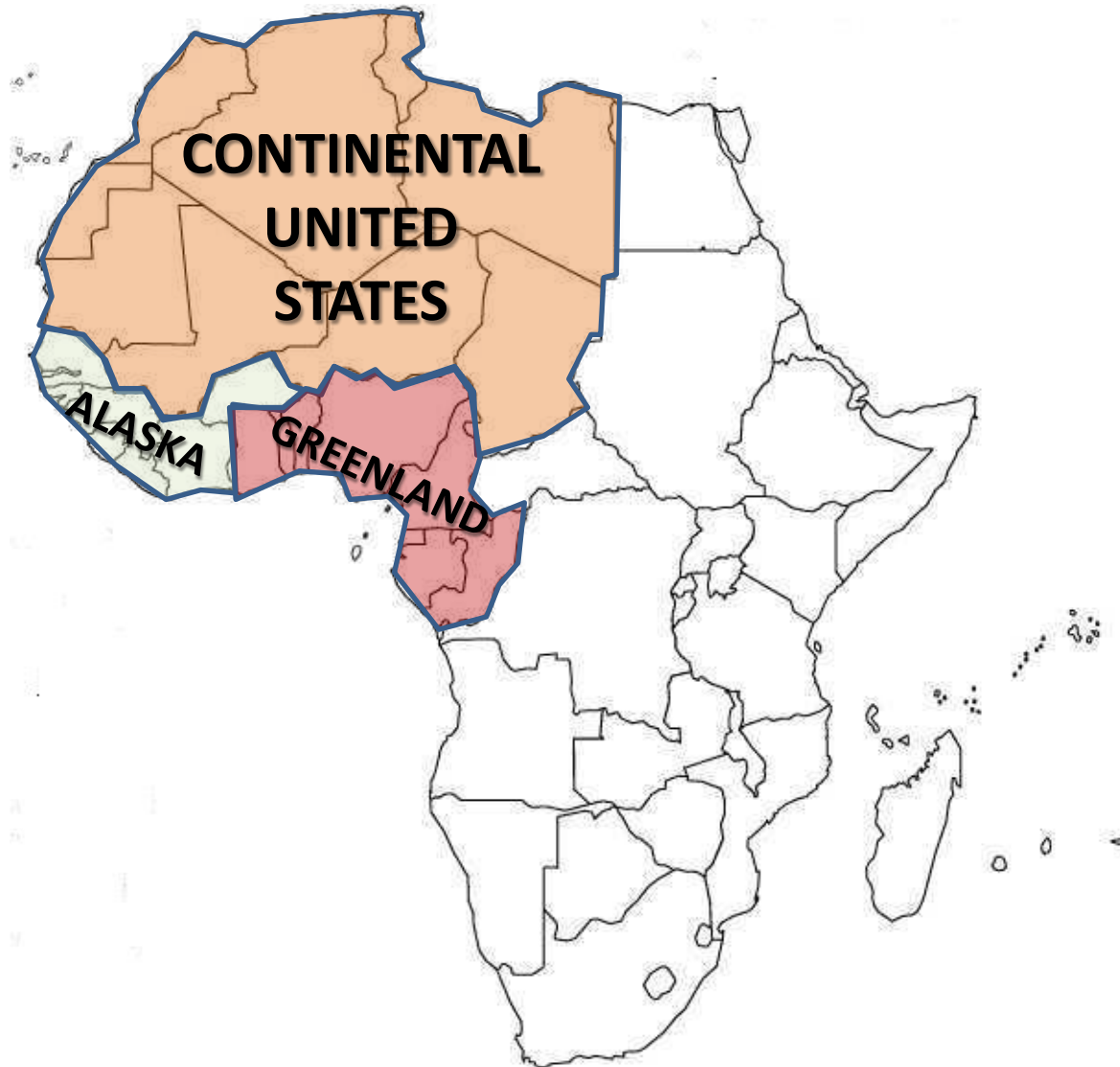
African Languages



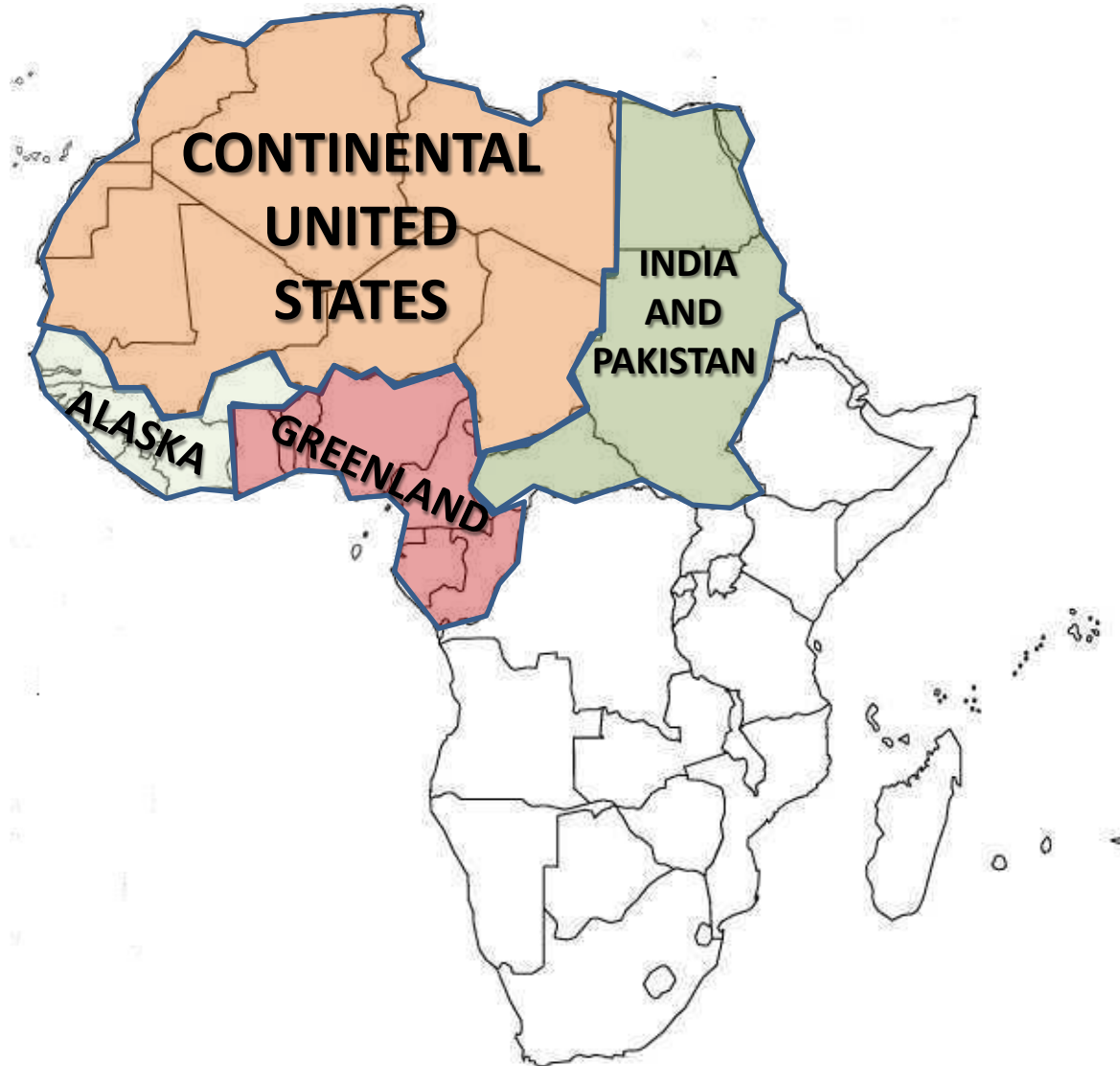
African Languages



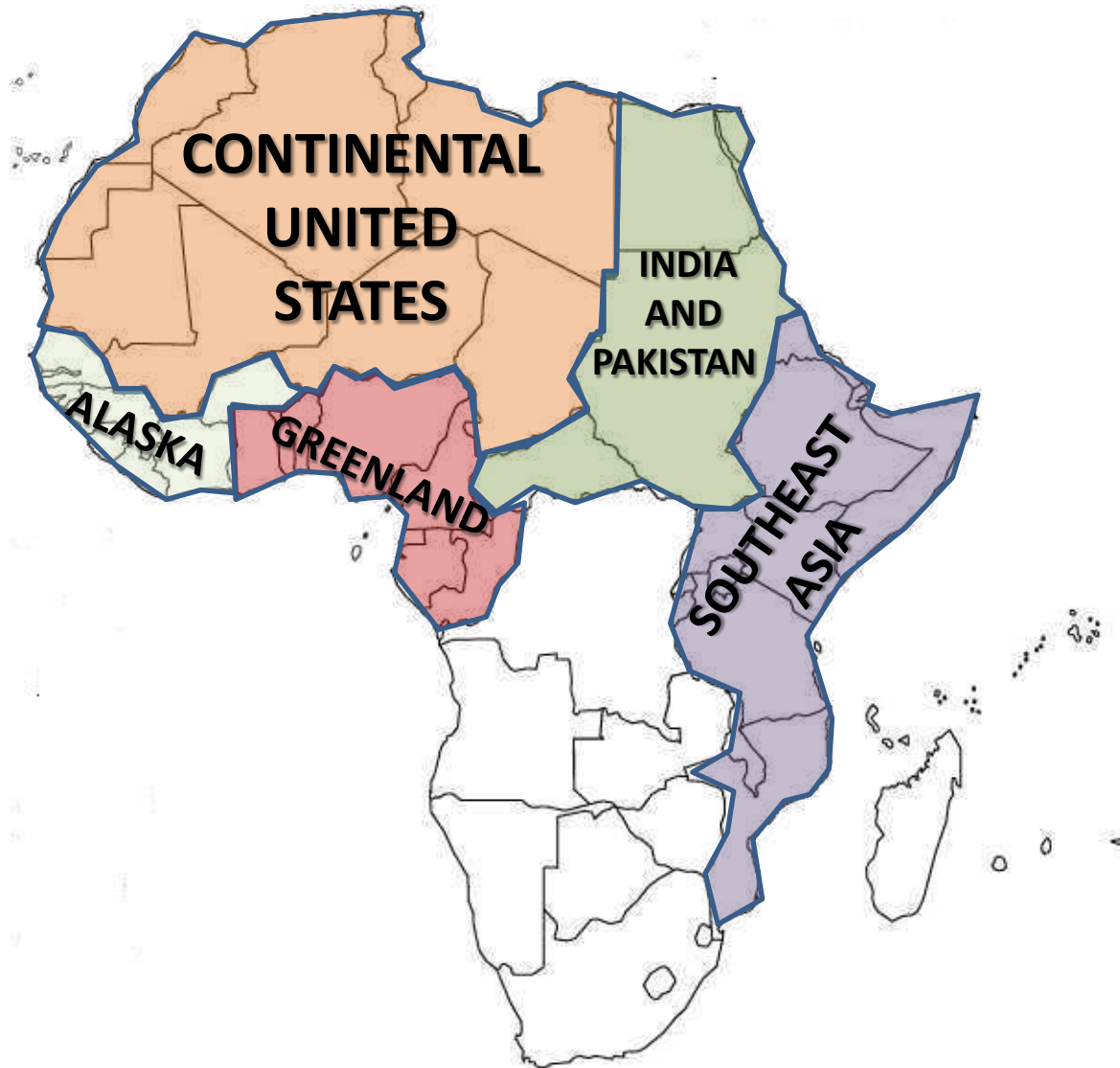
African Languages



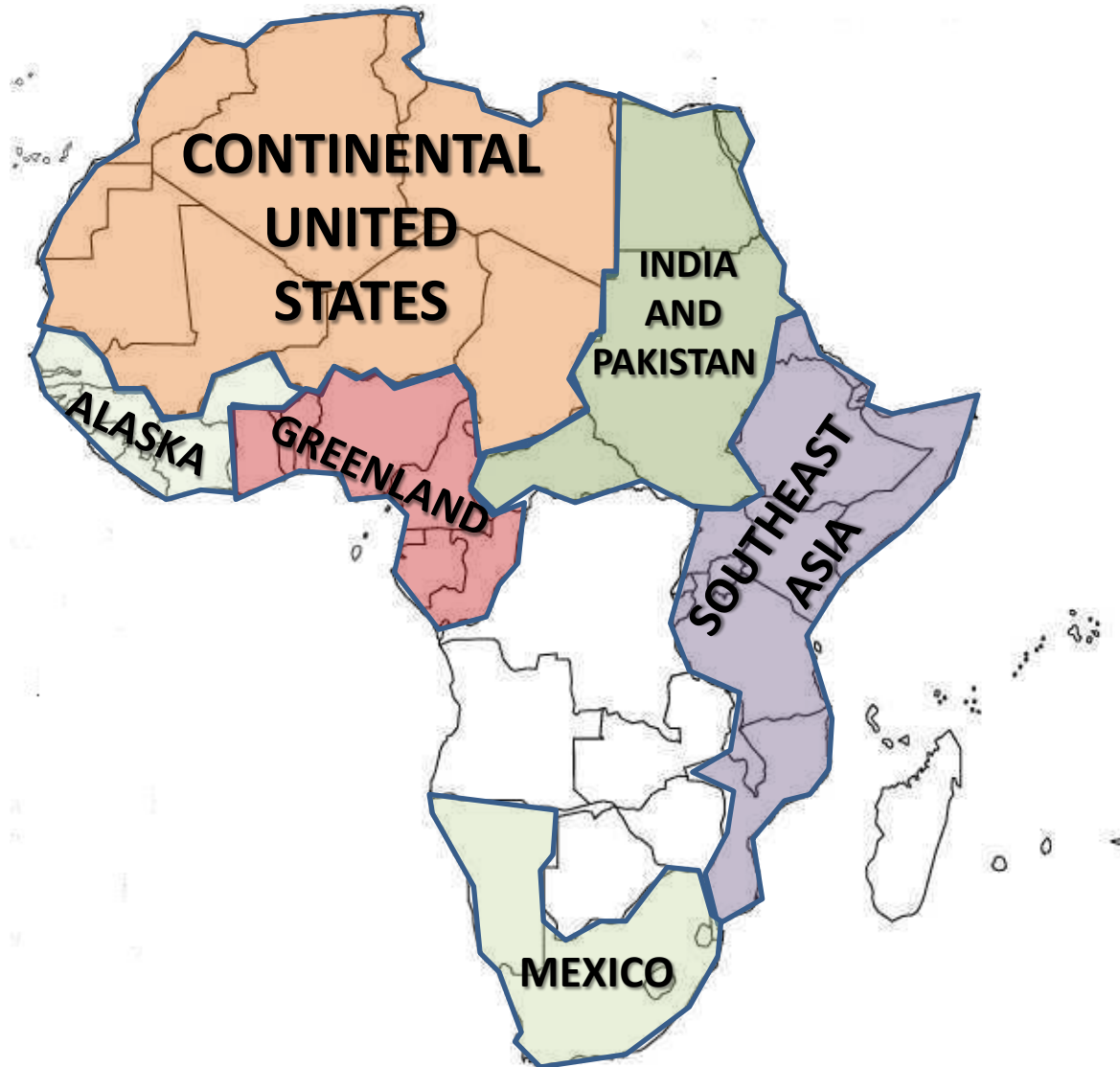
African Languages



African Languages



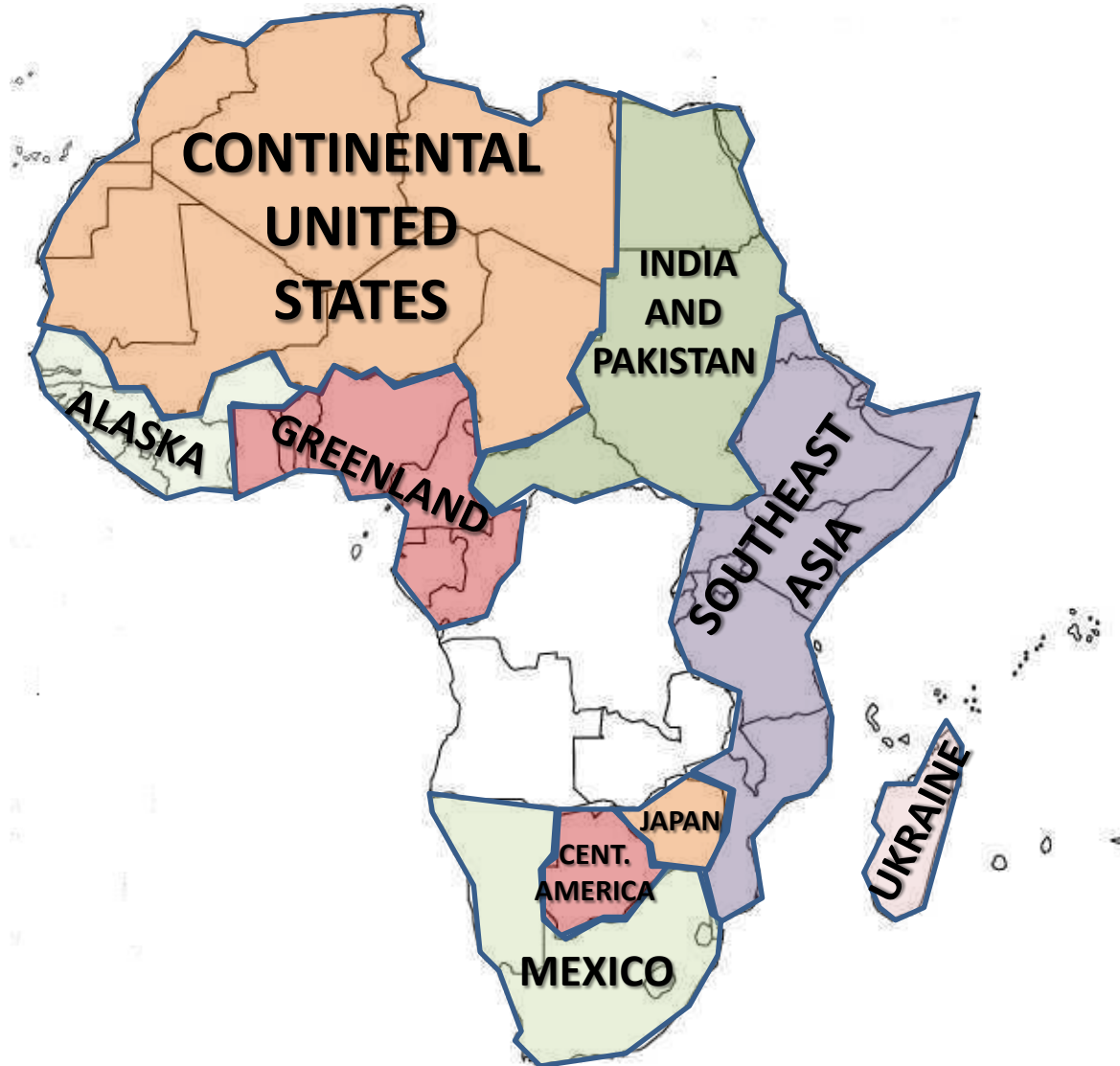
African Languages



African Languages



African Languages



African Languages



Africa



**almost
as big as
the Moon**

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

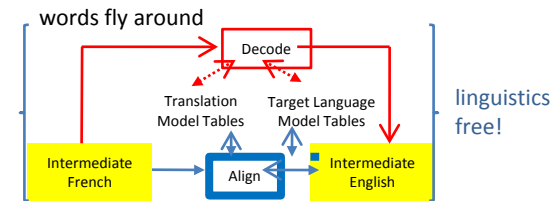
Recent Research

- 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

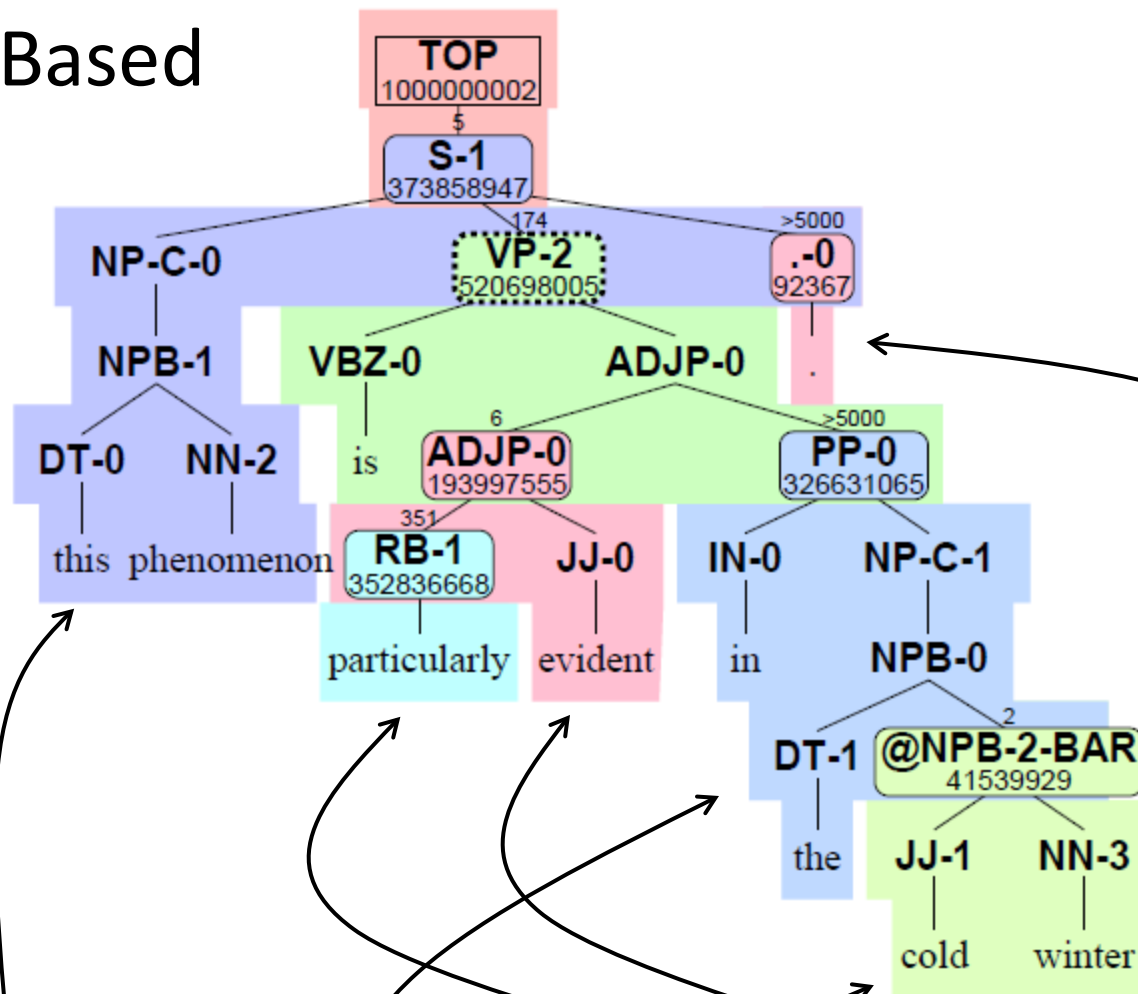
Not Super-Satisfying?

- 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

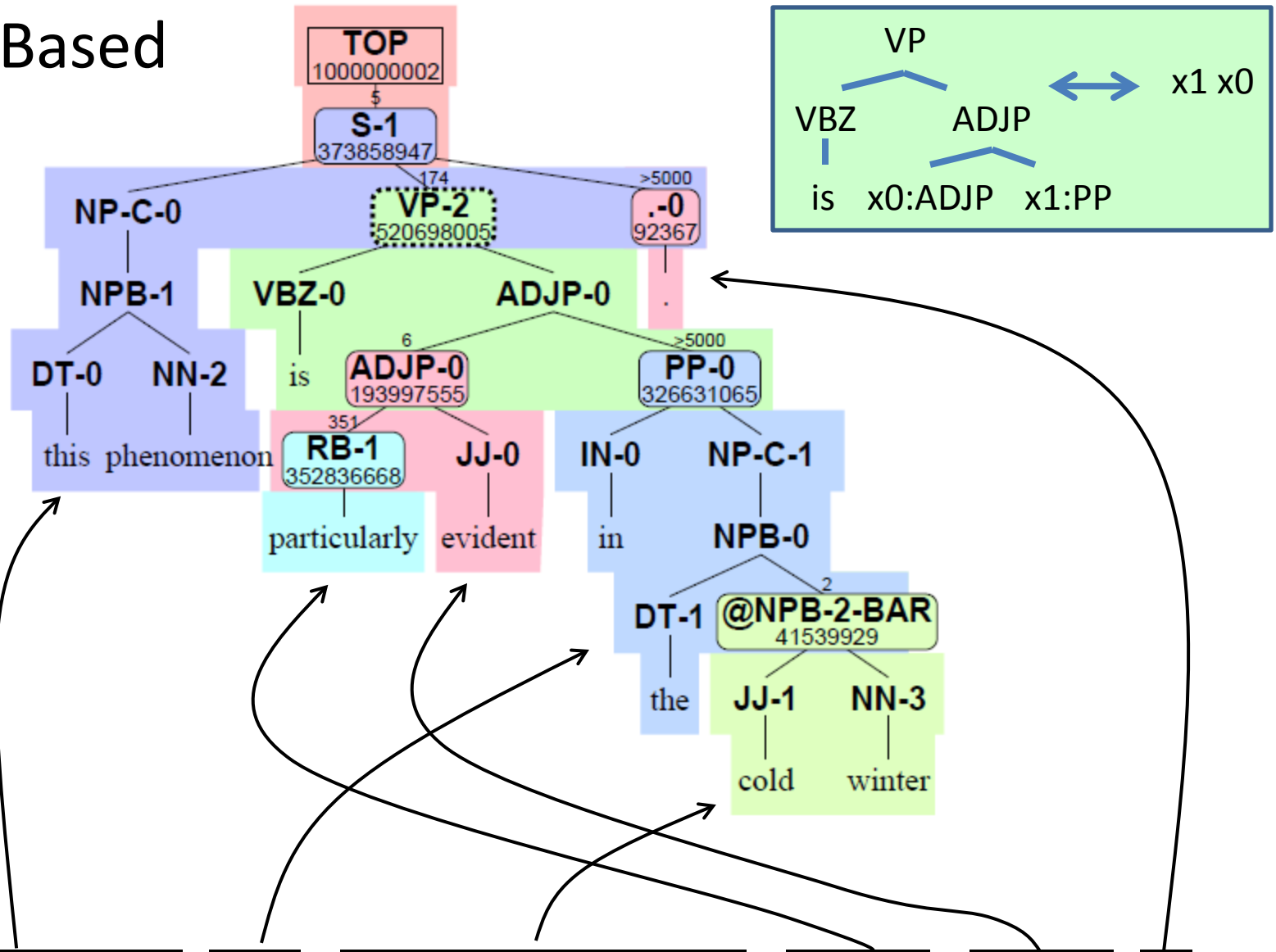
这种现象在寒冷的冬季尤其明显。

Syntax-Based SMT



这种现象在寒冷的冬季尤其明显。

Syntax-Based SMT



这种现象在寒冷的冬季尤其明显。

Tree-Based Output

枪手 被 警方 击毙 .

Tree-Based Output

枪手 被 警方 击毙 .

The gunman killed by police .

DT NN VBD IN NN

NPB

PP

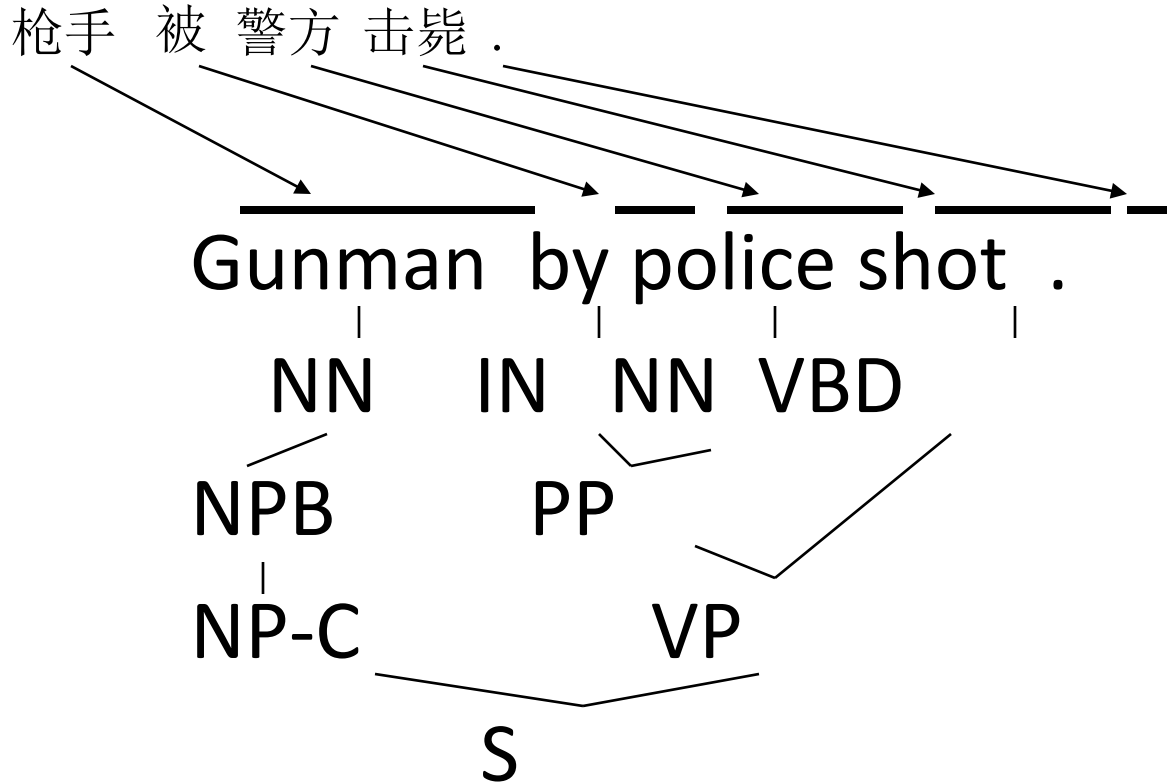
NP-C

VP

S

*Decoder
Hypothesis #1*

Tree-Based Output



*Decoder
Hypothesis #16*

Tree-Based Output

枪手 被 警方 击毙 .

The gunman was killed by police .

*Decoder
Hypothesis #1923*

DT NN AUX VBN IN NN

NPB

PP

NP-C

VP

S

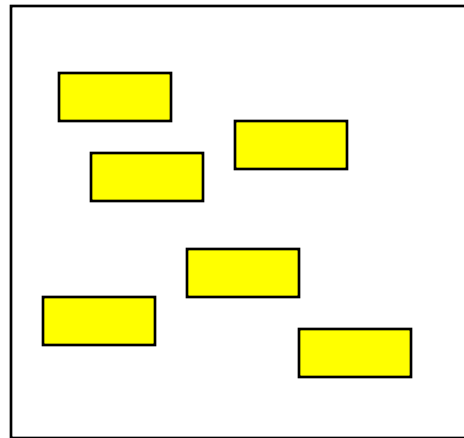
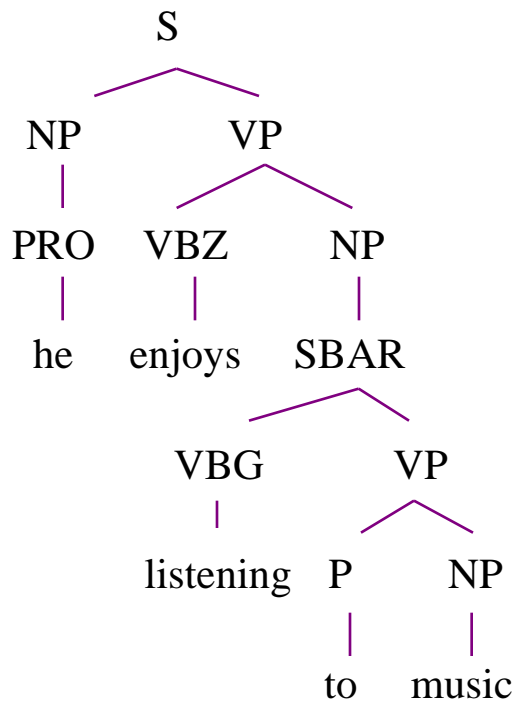
highest scoring
output, syntax-
based model

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

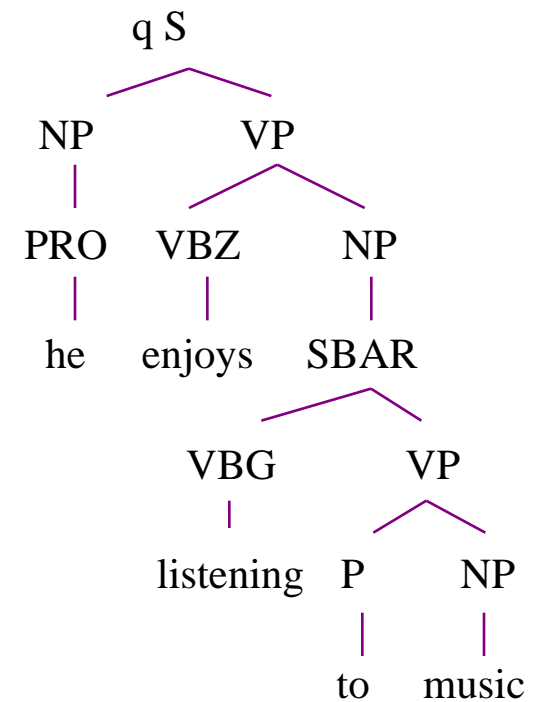
Top-Down Tree Transducer

(W. Rounds 1970; J. Thatcher 1970)

Original input:



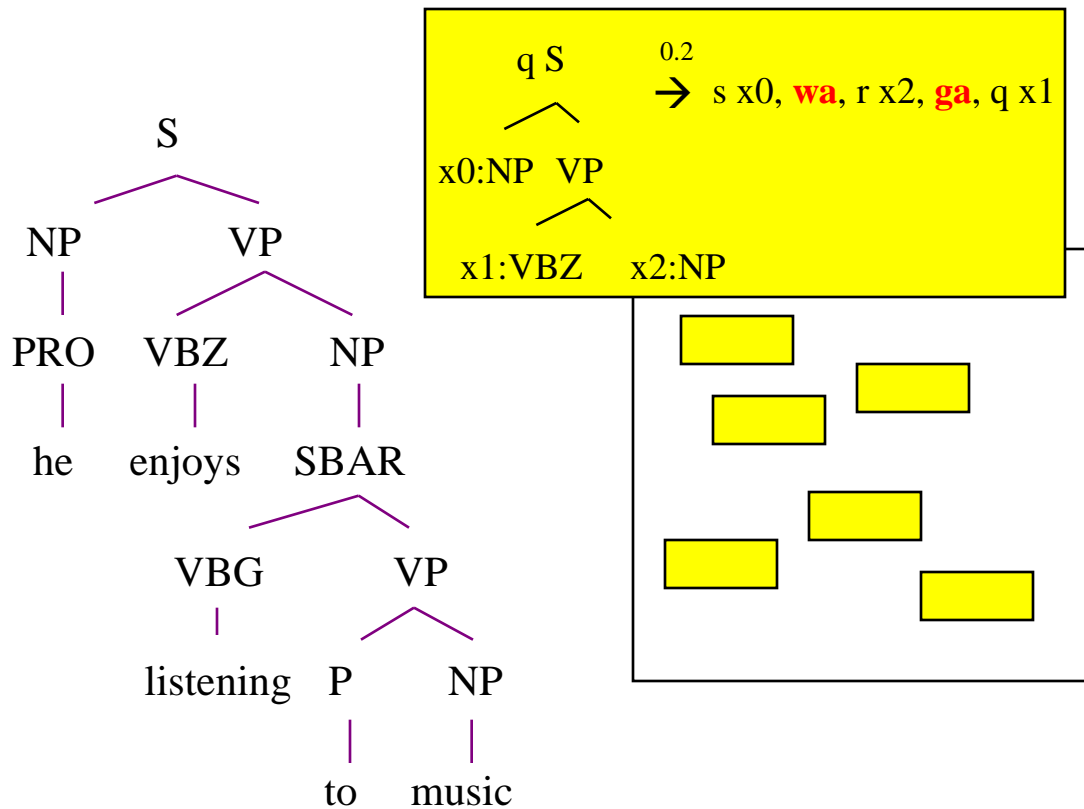
Transformation:



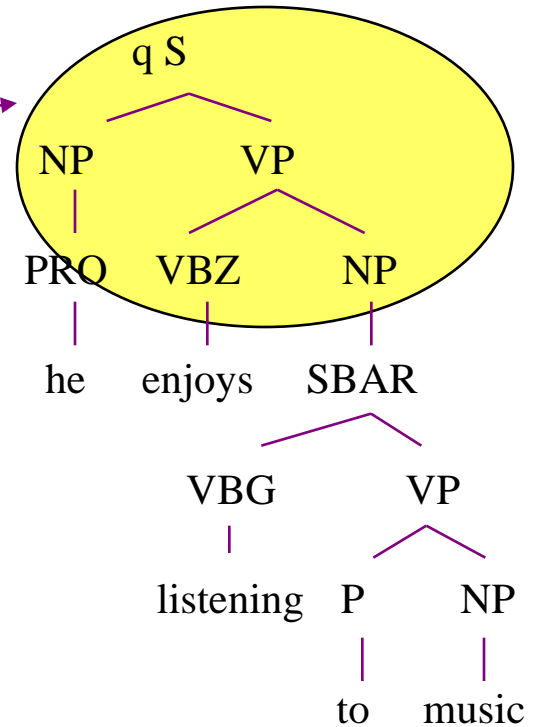
Top-Down Tree Transducer

(W. Rounds 1970; J. Thatcher 1970)

Original input:



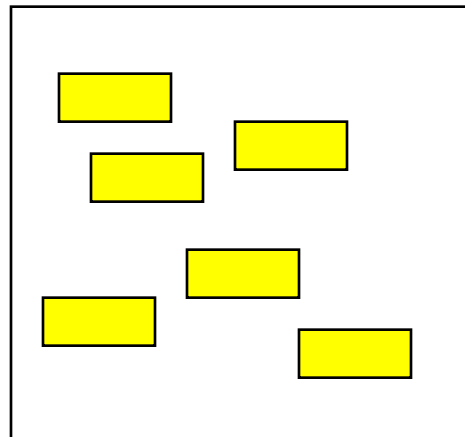
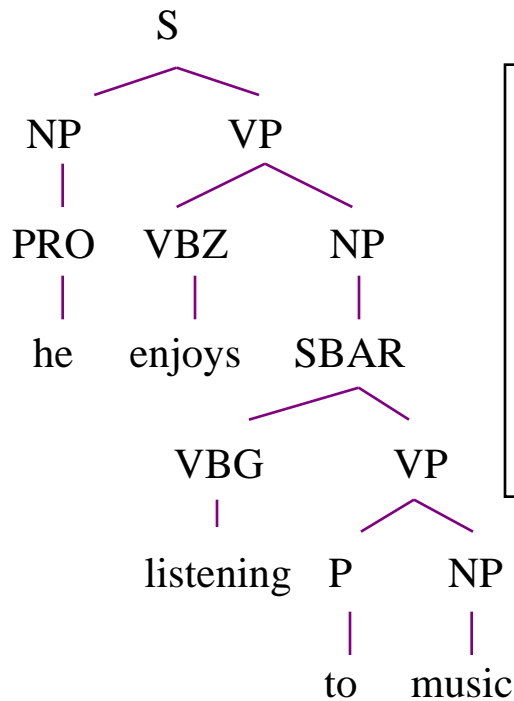
Transformation:



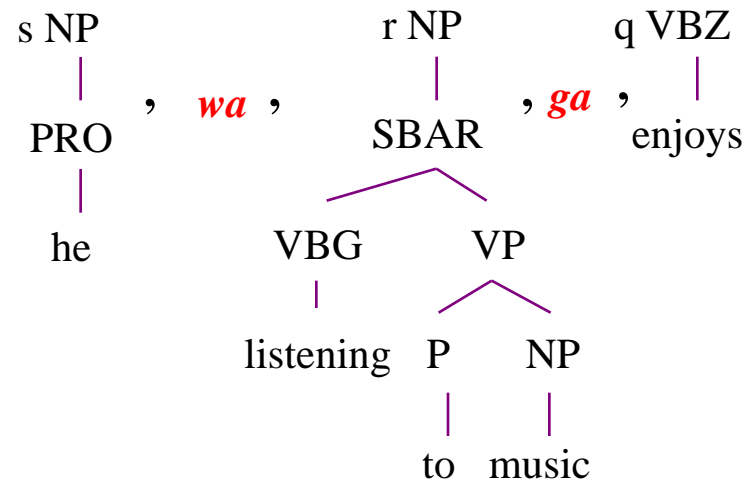
Top-Down Tree Transducer

(W. Rounds 1970; J. Thatcher 1970)

Original input:



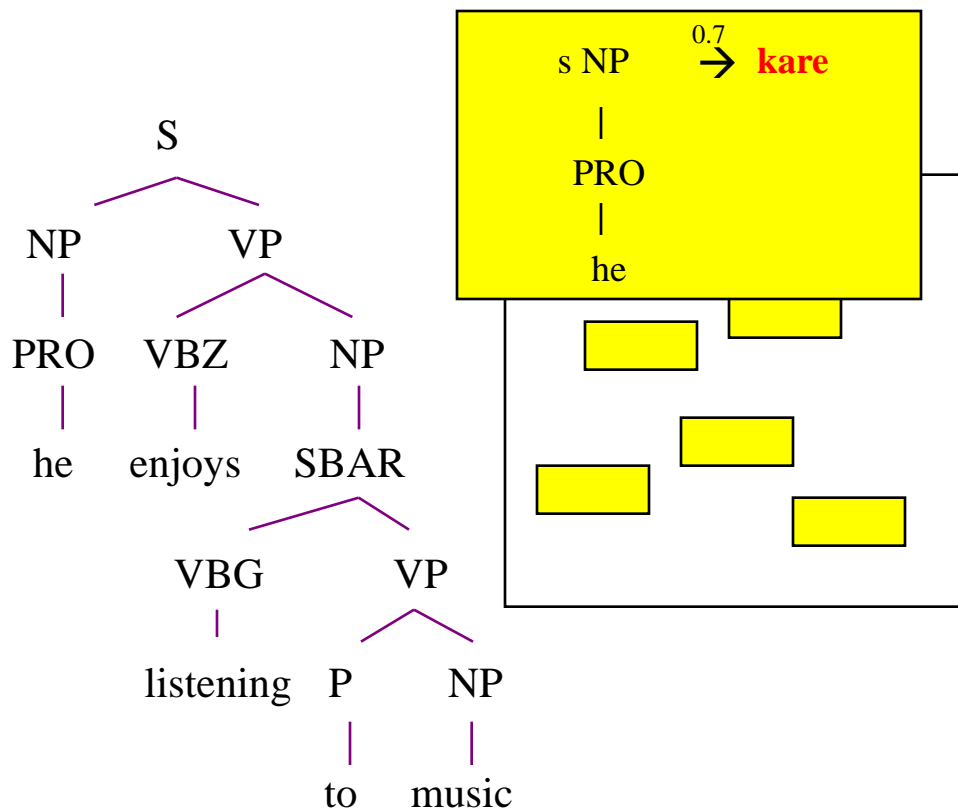
Transformation:



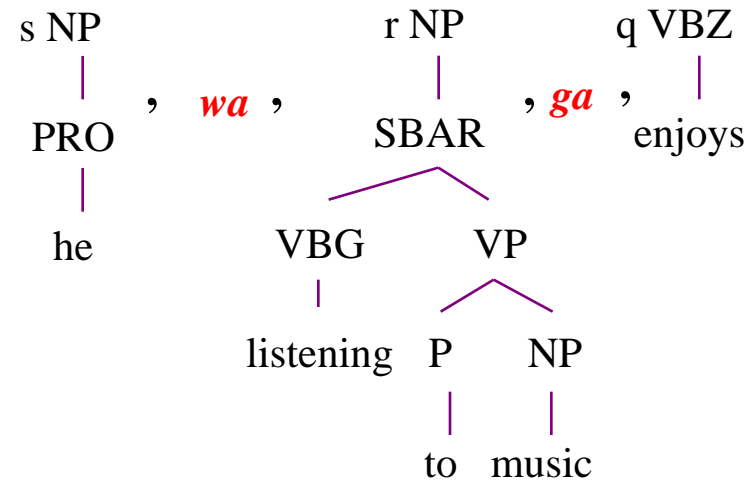
Top-Down Tree Transducer

(W. Rounds 1970; J. Thatcher 1970)

Original input:



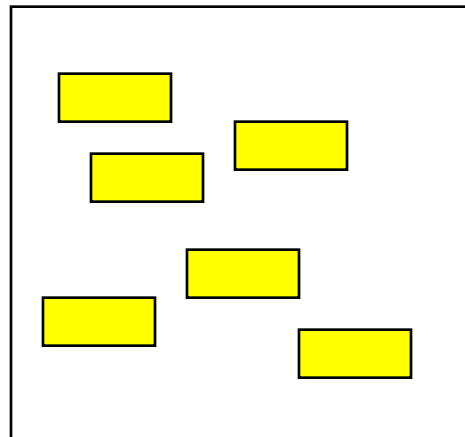
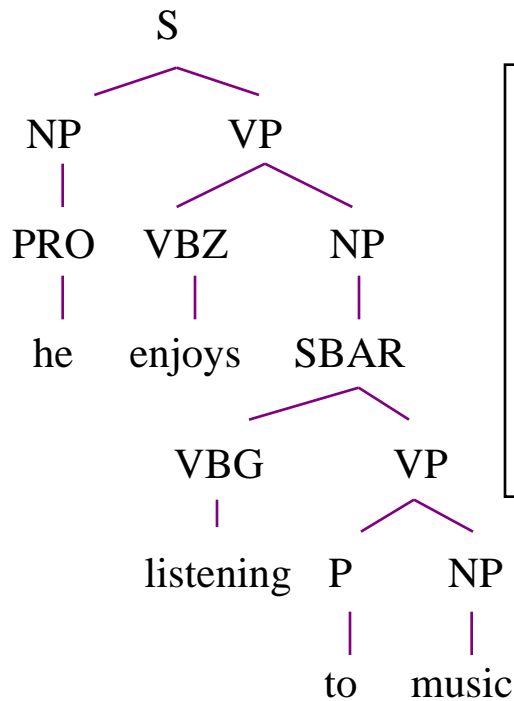
Transformation:



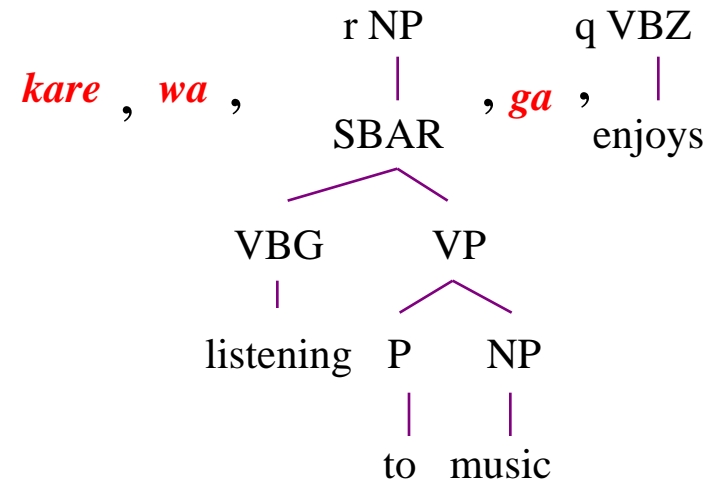
Top-Down Tree Transducer

(W. Rounds 1970; J. Thatcher 1970)

Original input:



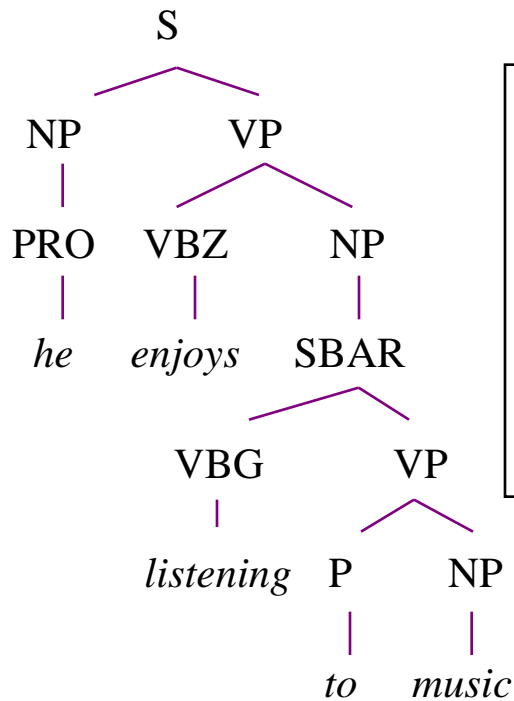
Transformation:



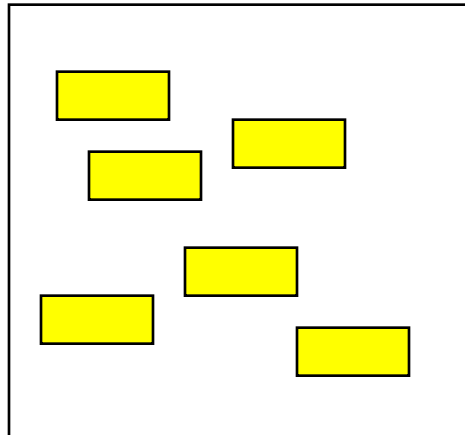
Top-Down Tree Transducer

(W. Rounds 1970; J. Thatcher 1970)

Original input:



Final output:



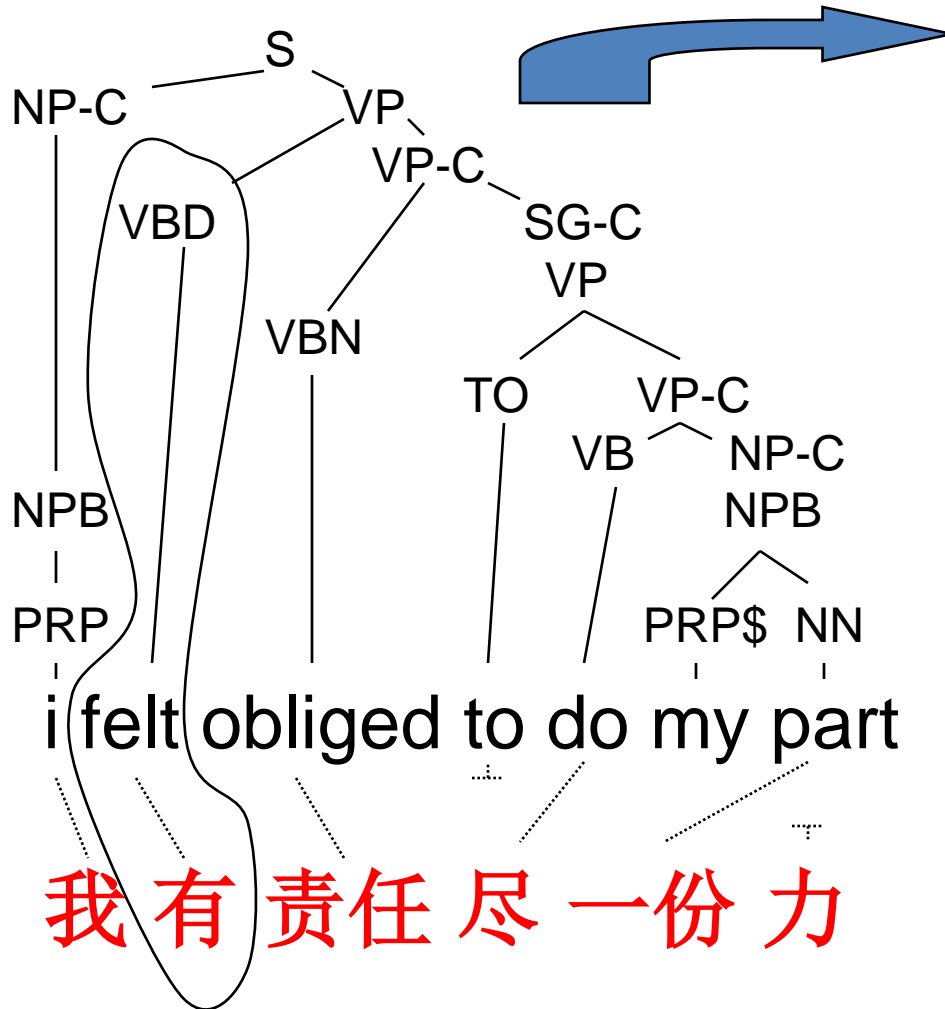
kare, wa, ongaku, o, kiku, no, ga, daisuki, desu

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

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.

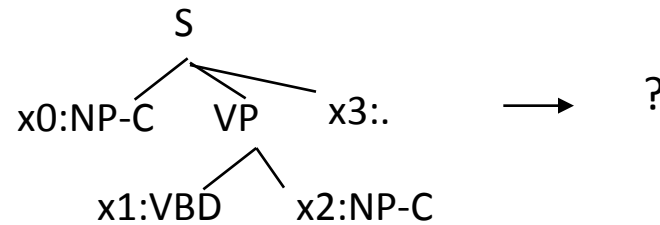
Tree Transducers Can be Extracted from Bilingual Data (Galley, Hopkins, Knight, Marcu, 2004)



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

- $\operatorname{argmax}_{\text{etree}} P(\text{etree} \mid \text{cstring})$
- 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.

“these”

RULE 1:
DT(these)
→ 这

“include”

RULE 2:
VBP(include)
→ 中包括

“France”

RULE 4:
NNP(France)
→ 法国

“and”

RULE 5:
CC(and)
→ 和

“Russia”

RULE 6:
NNP(Russia)
→ 俄罗斯

“astronauts”

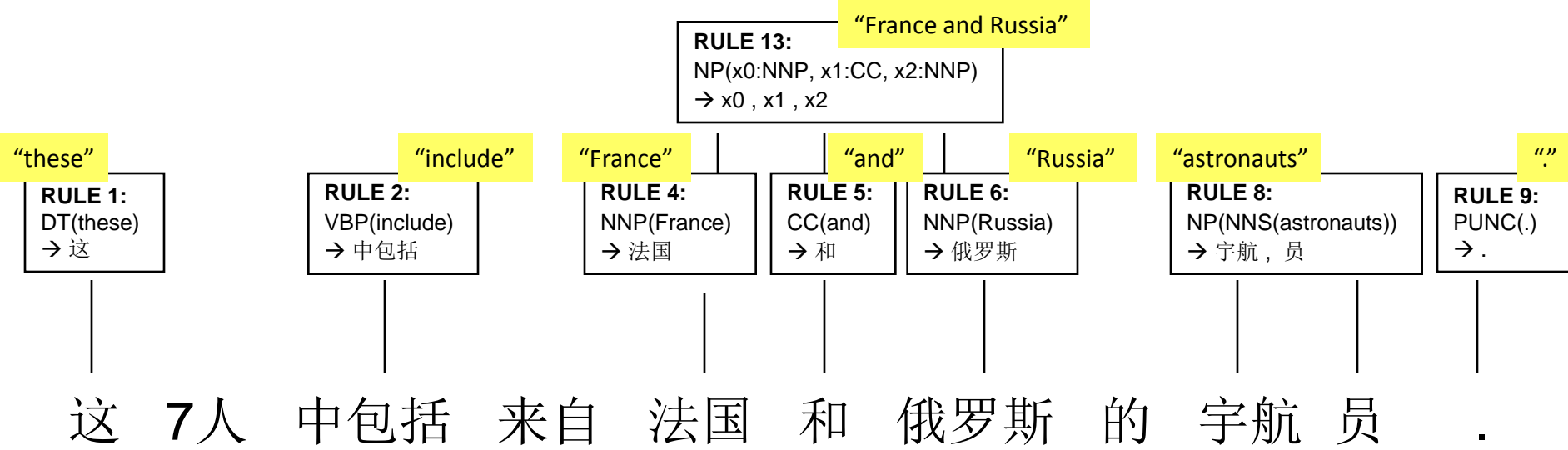
RULE 8:
NP(NNS(astronauts))
→ 宇航, 员

“.”

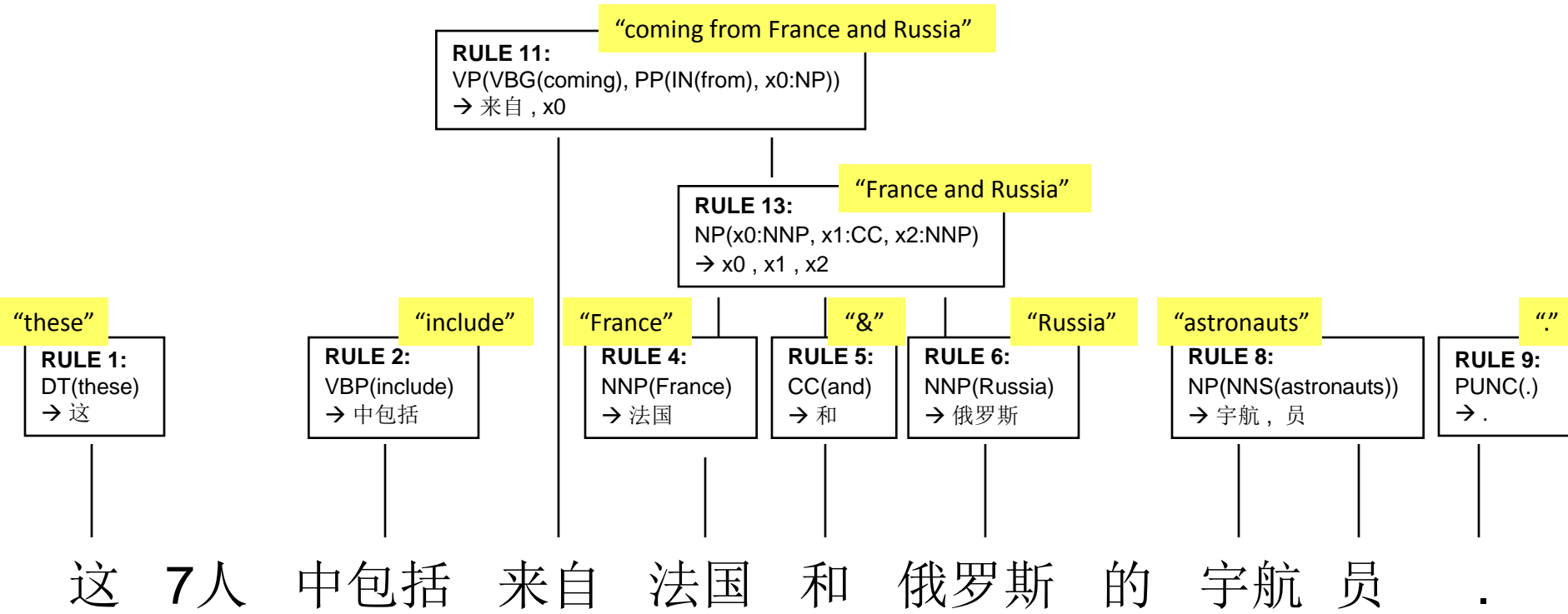
RULE 9:
PUNC(.)
→ .

这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

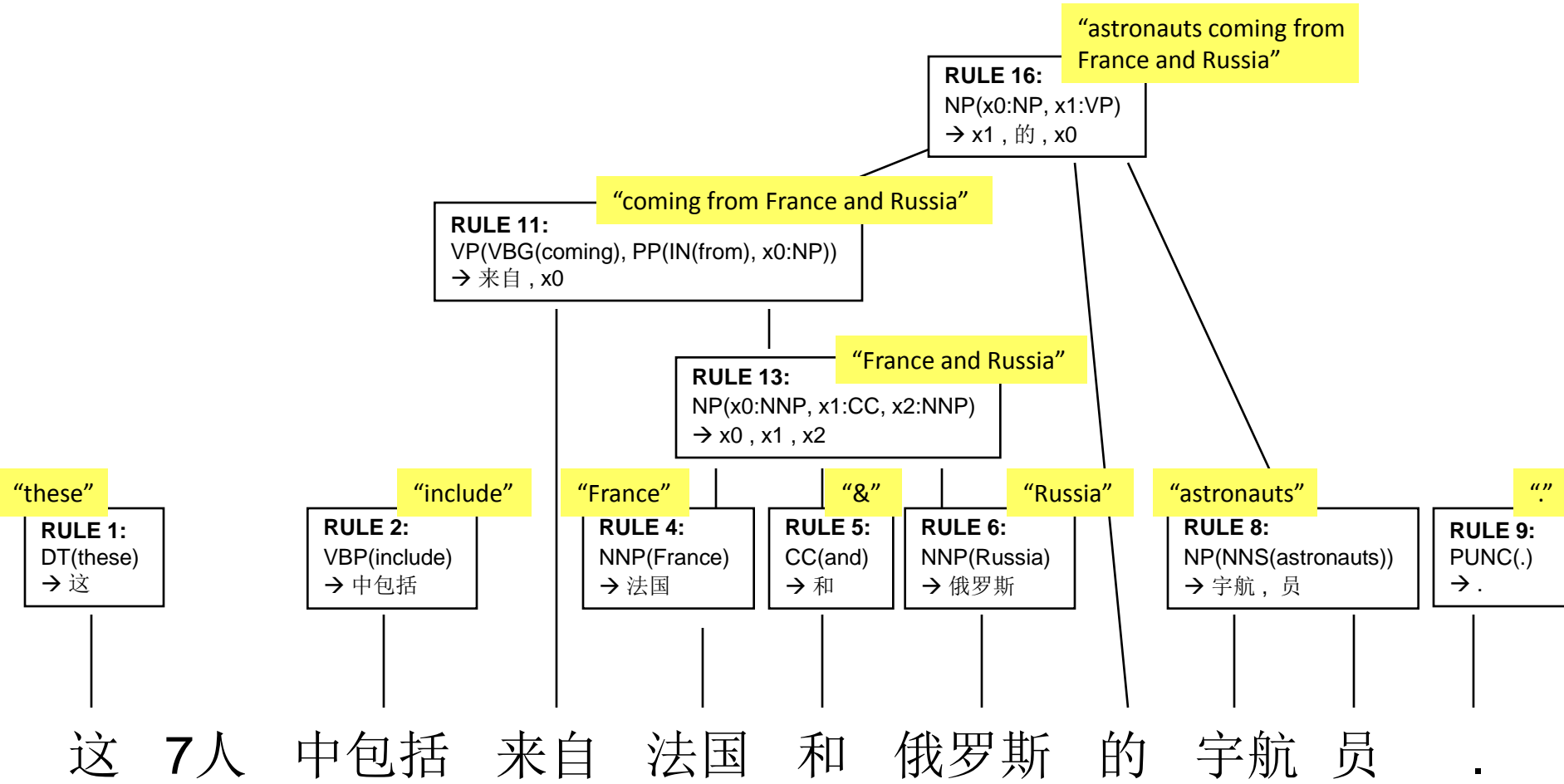
Syntax-Based Decoding

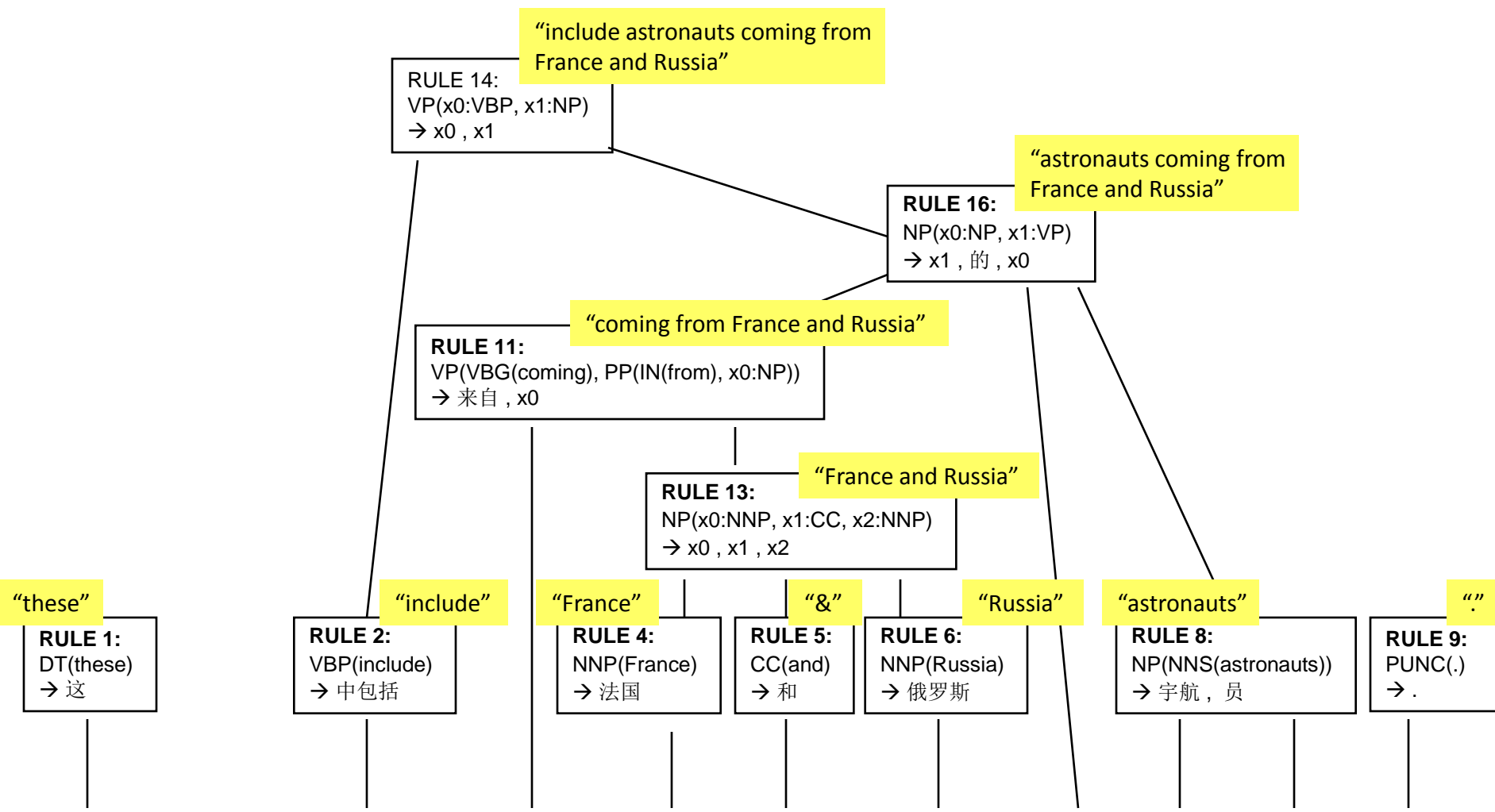


Syntax-Based Decoding



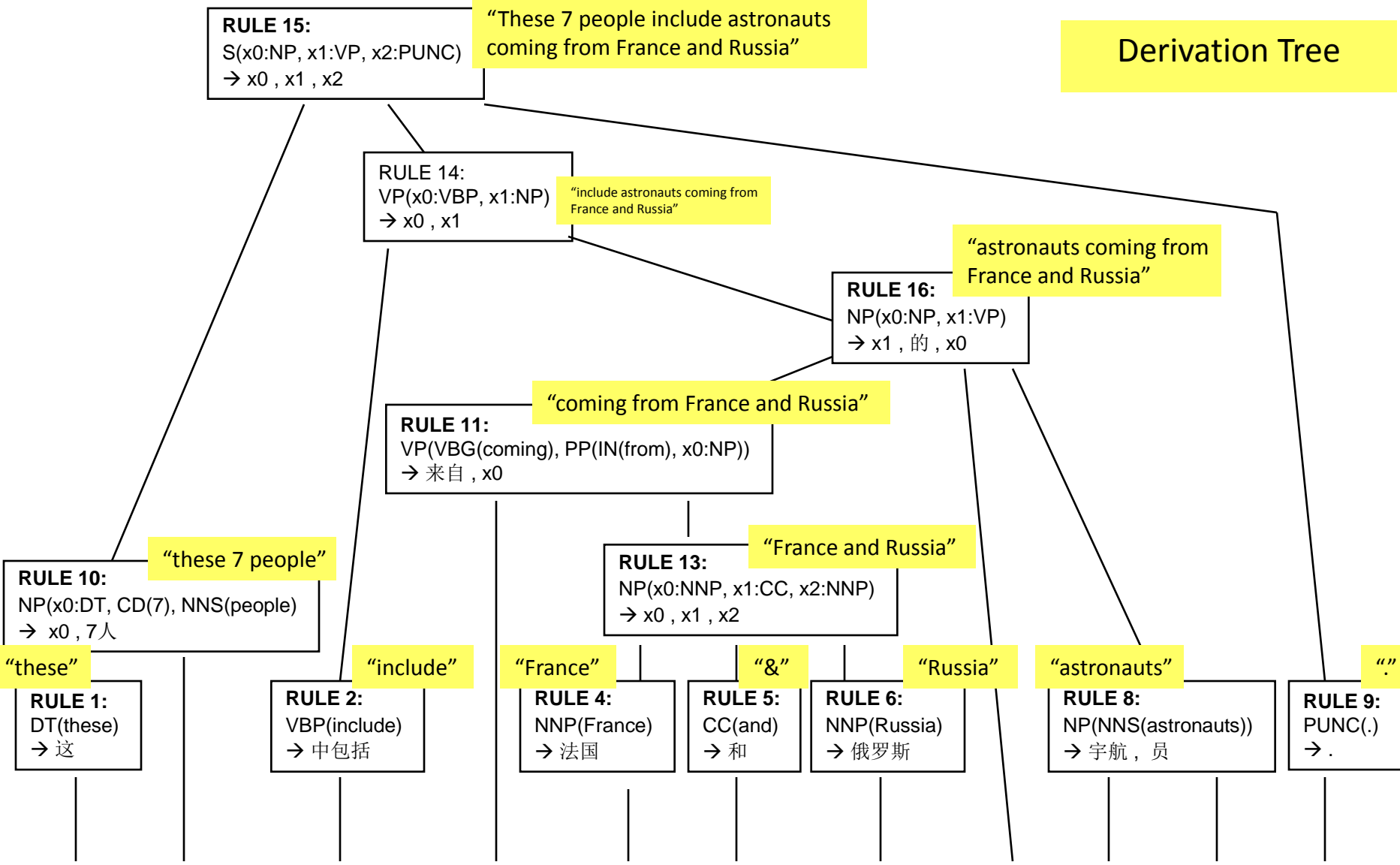
Syntax-Based Decoding





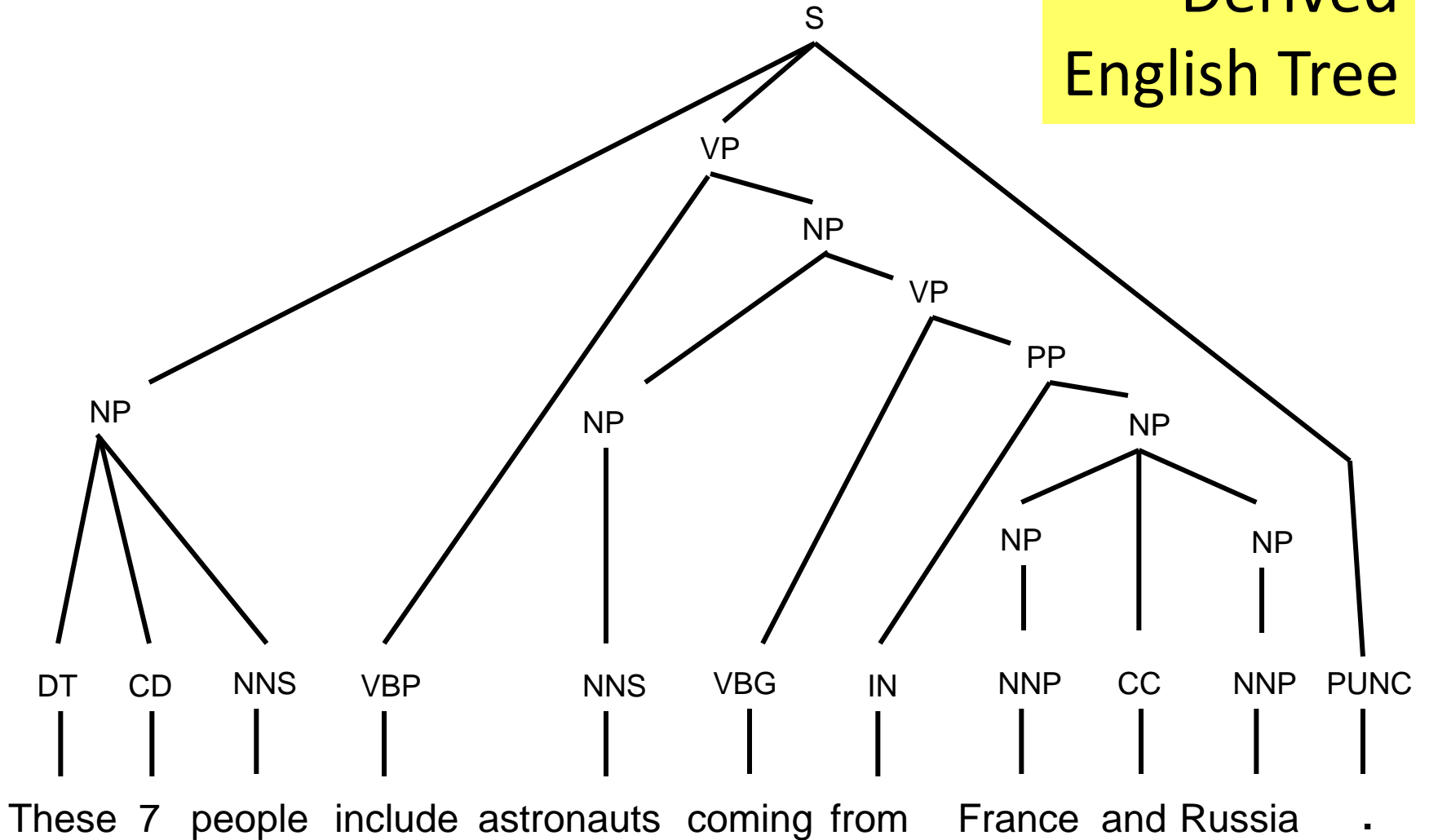
这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

Derivation Tree

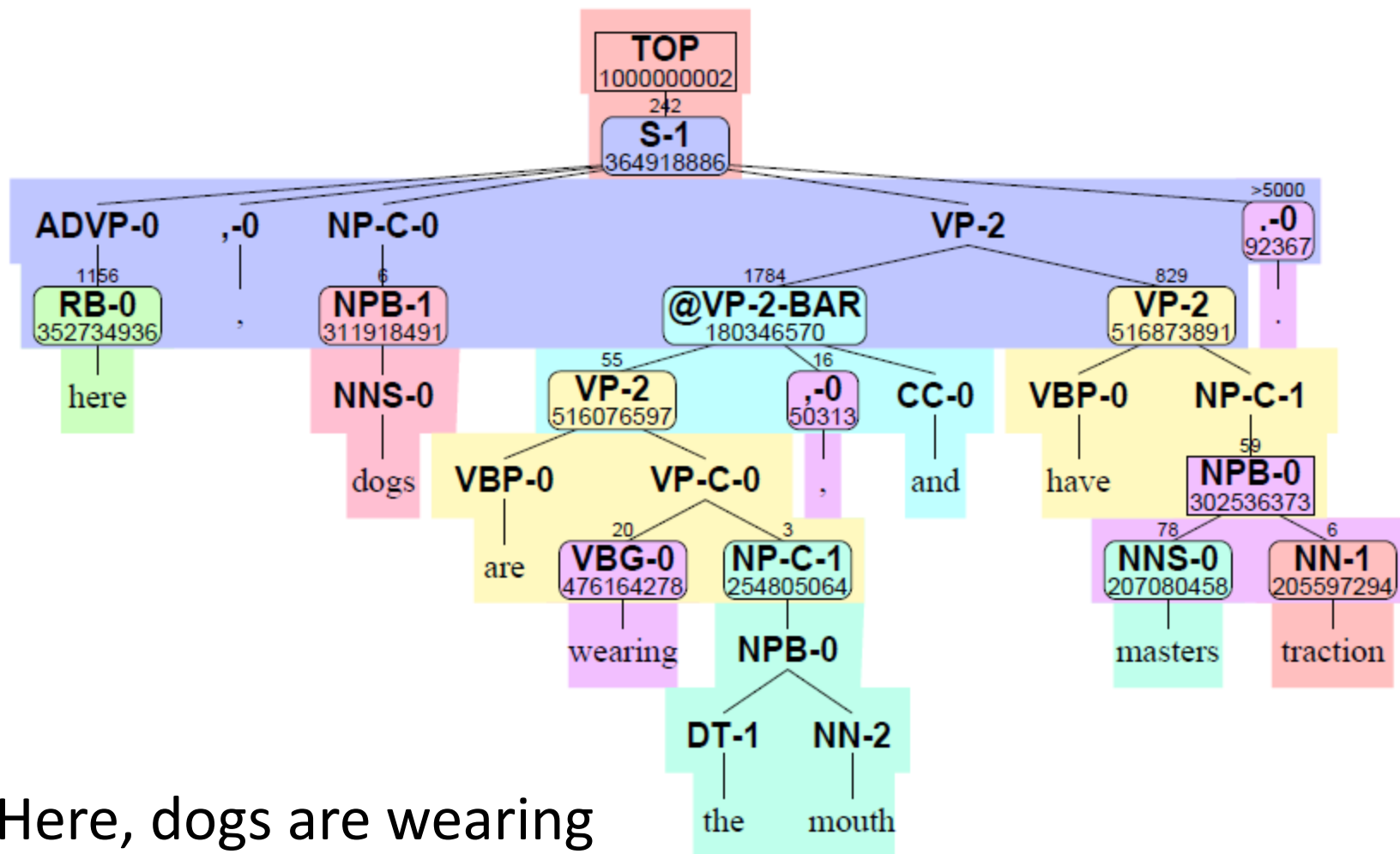


这 7人中包括来自法国和俄罗斯的宇航员。

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

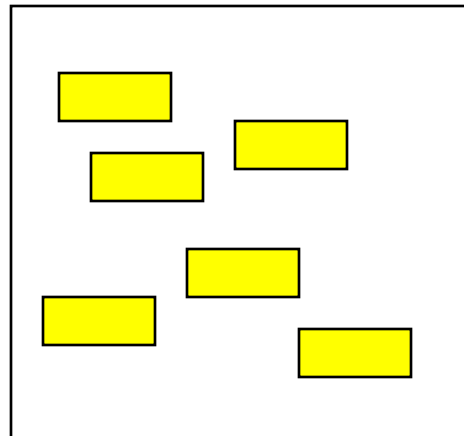
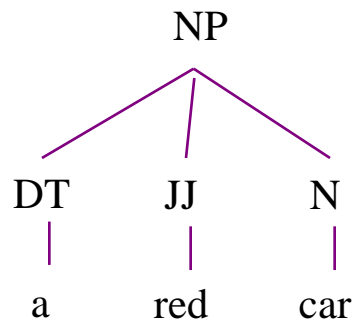
q.NP(x0:DT x1:JJ x2:N) <-> q.x0 q.x2 q.x1

} Very large
wordform-to-wordform
dictionary

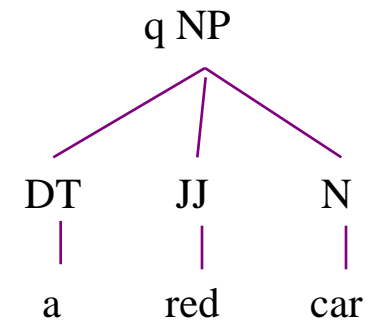
} Simple syntactic
combination

Current Syntax-Based SMT

Original input:

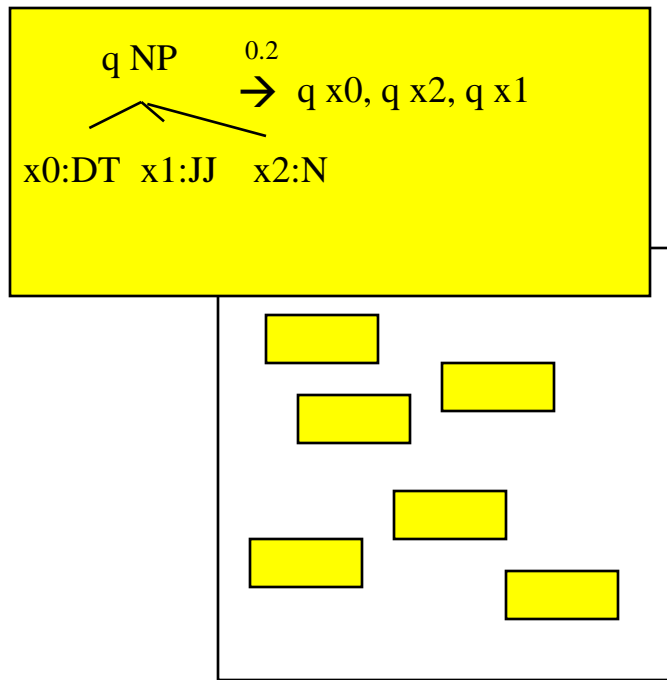
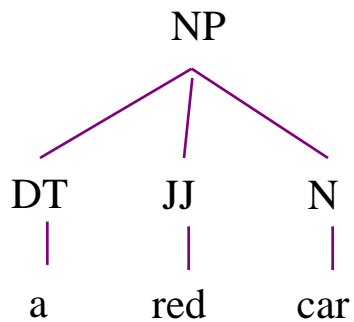


Transformation:

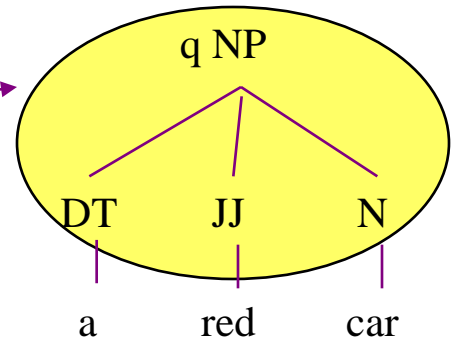


Current Syntax-Based SMT

Original input:

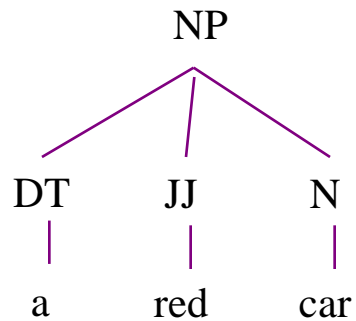


Transformation:

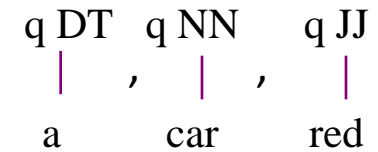
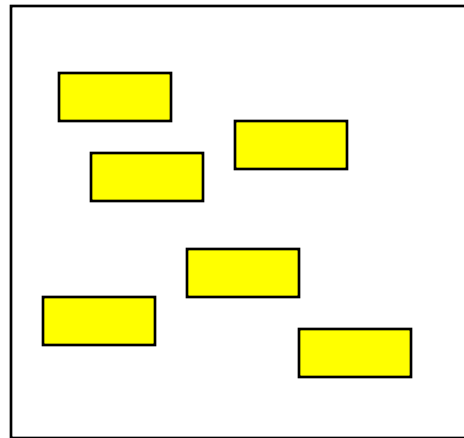


Current Syntax-Based SMT

Original input:

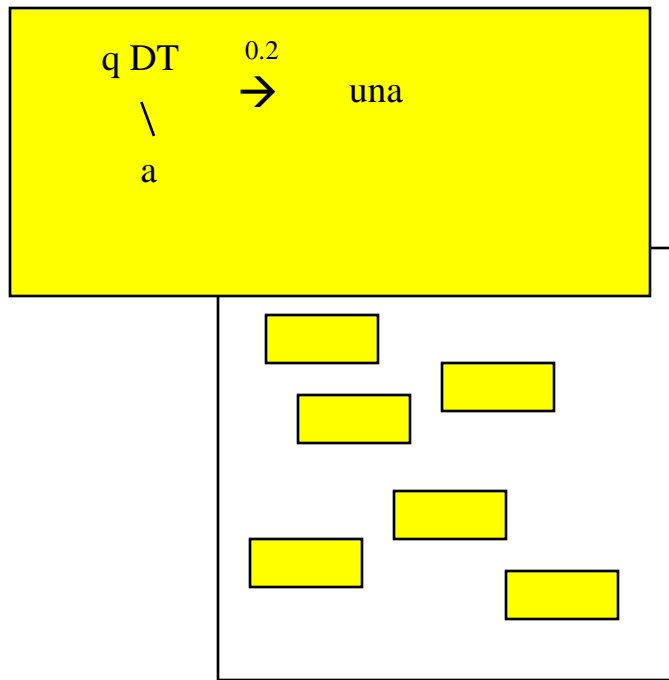
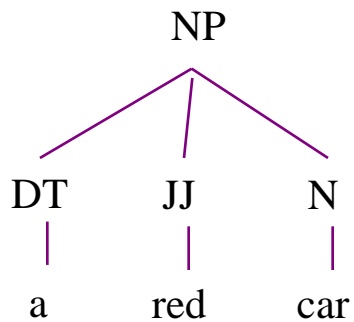


Transformation:

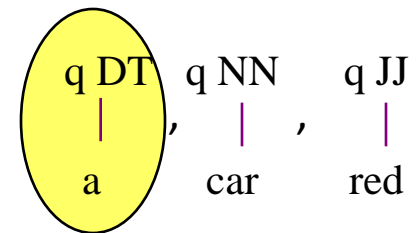


Current Syntax-Based SMT

Original input:

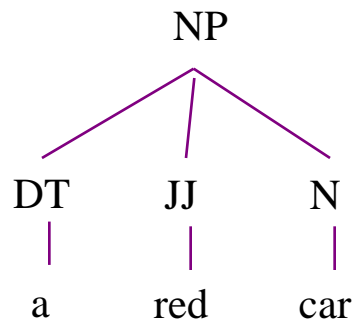


Transformation:

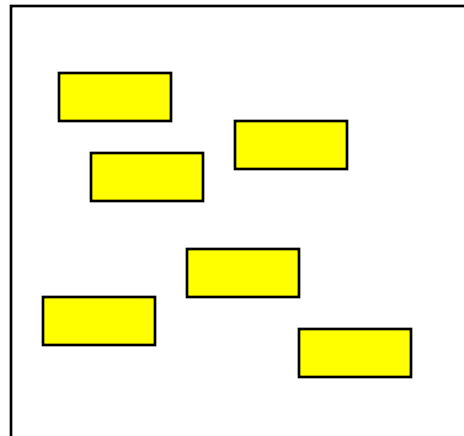


Current Syntax-Based SMT

Original input:



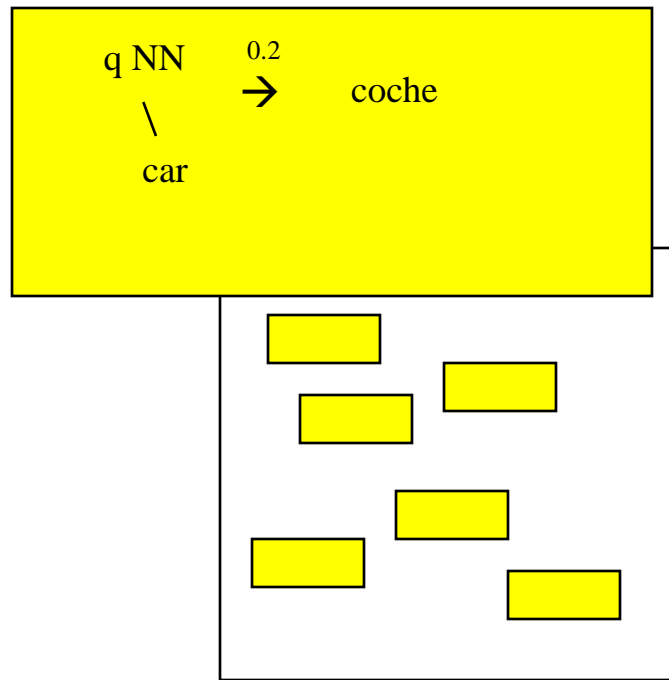
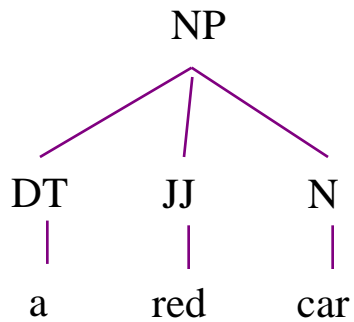
Transformation:



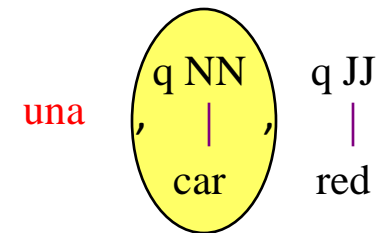
una , q NN , q JJ
car red

Current Syntax-Based SMT

Original input:

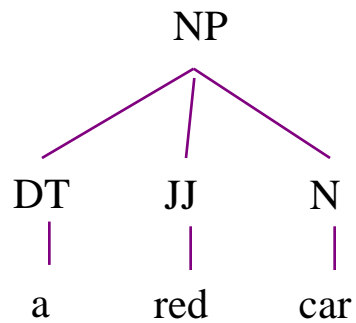


Transformation:

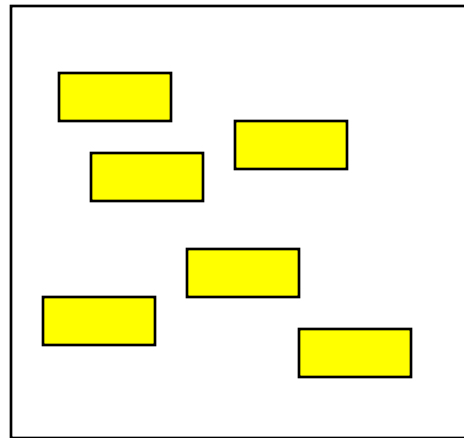


Current Syntax-Based SMT

Original input:



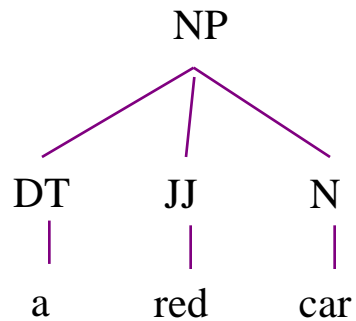
Transformation:



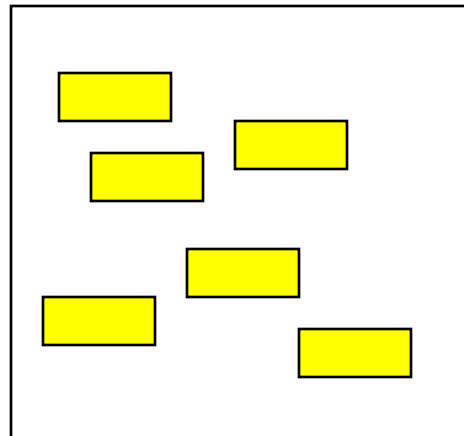
una , coche, q JJ
|
red

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
qnsmasc.car <-> c o c h e
qnsmasc.cat <-> g a t o
qnsmasc.light <-> l u z
qdfem.a <-> u n a
```

Compact
root-to-root
dictionary

```
qmasc.JJ(x0:) <-> qjo.x0 o
qmasc.JJ(x0:) <-> qje.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
```

Morpho-
syntax

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

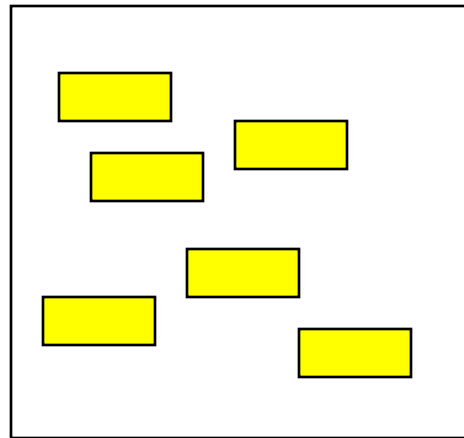
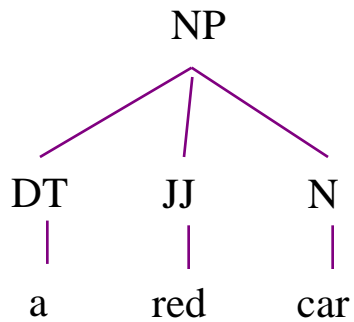
OUTPUTS:

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

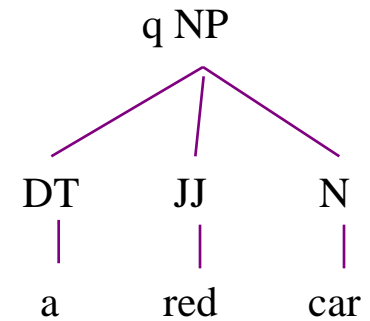
Translation

Possible Morpho-Syntactic SMT?

Original input:

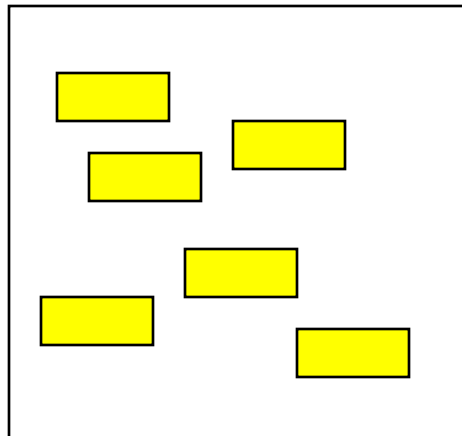
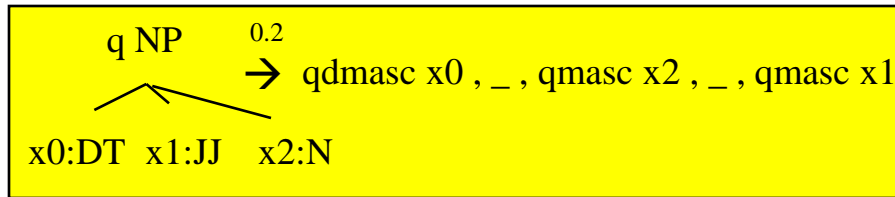
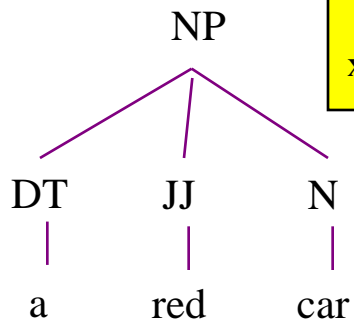


Transformation:

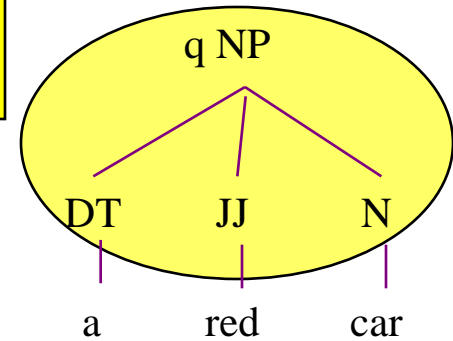


Possible Morpho-Syntactic SMT?

Original input:

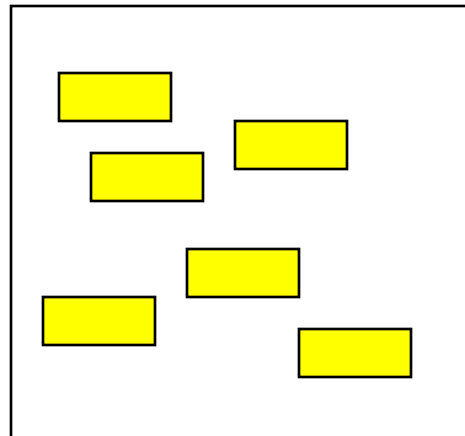
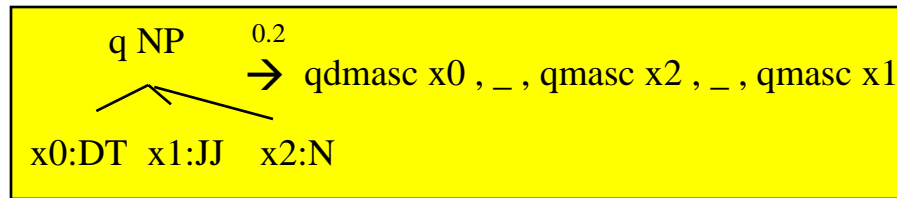
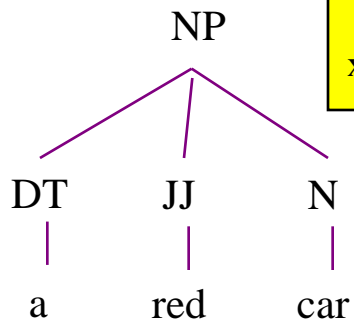


Transformation:

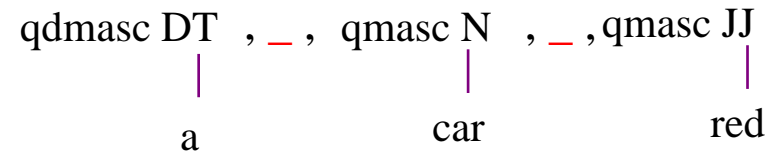


Possible Morpho-Syntactic SMT?

Original input:

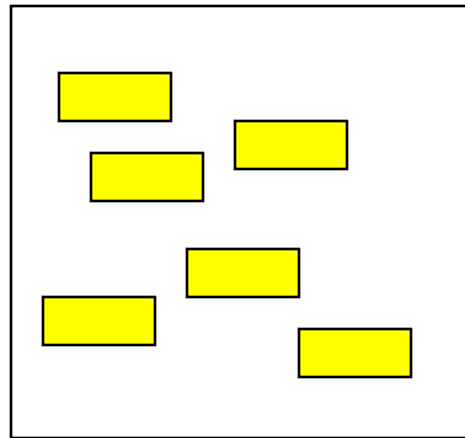
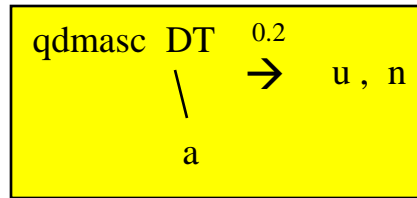
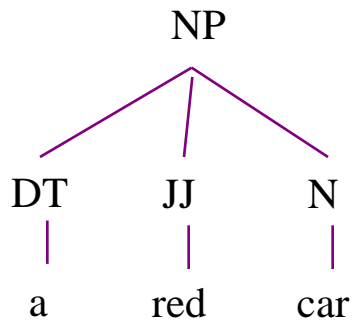


Transformation:

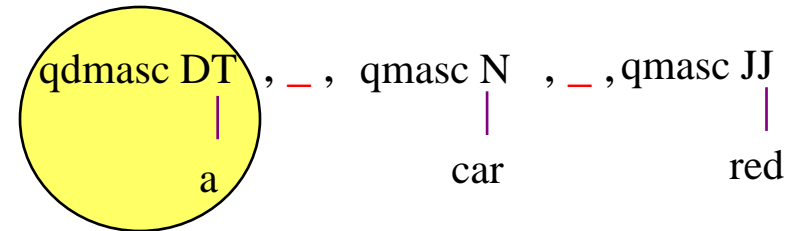


Possible Morpho-Syntactic SMT?

Original input:

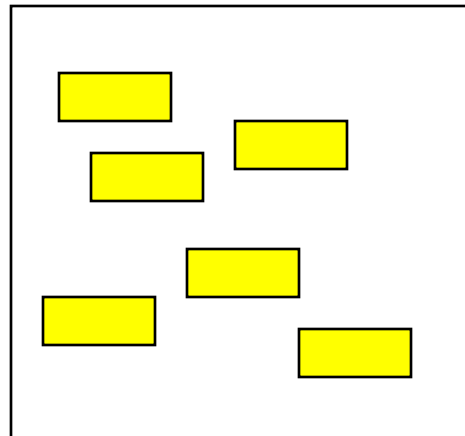
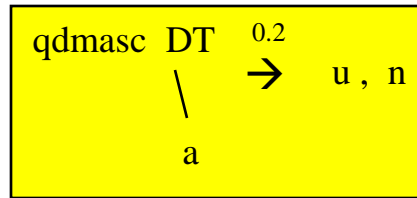
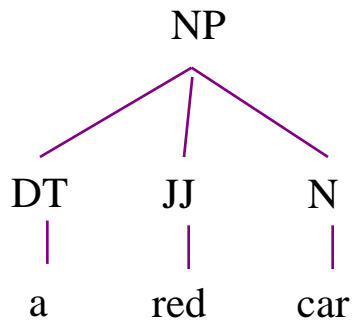


Transformation:

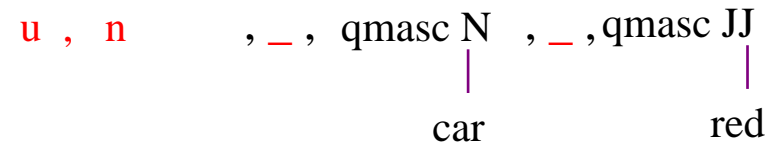


Possible Morpho-Syntactic SMT?

Original input:

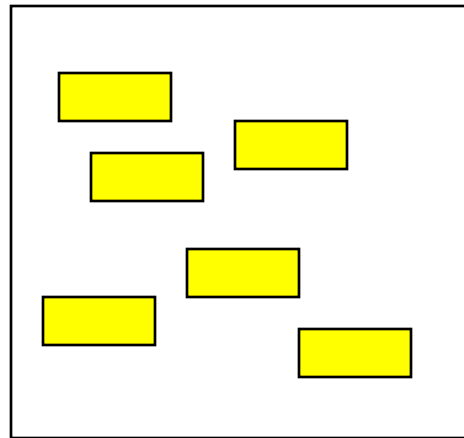
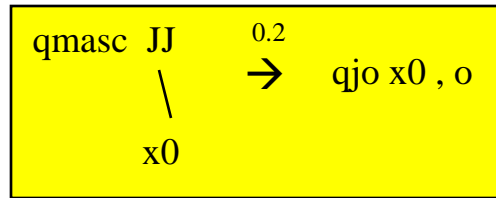
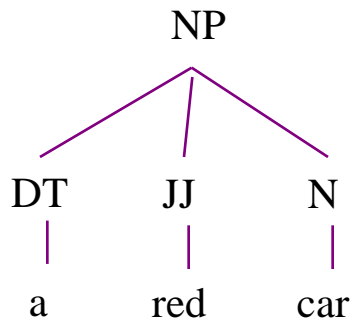


Transformation:



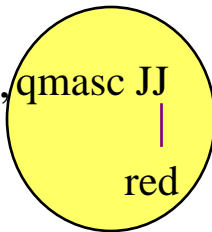
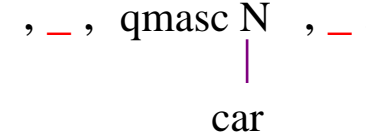
Possible Morpho-Syntactic SMT?

Original input:



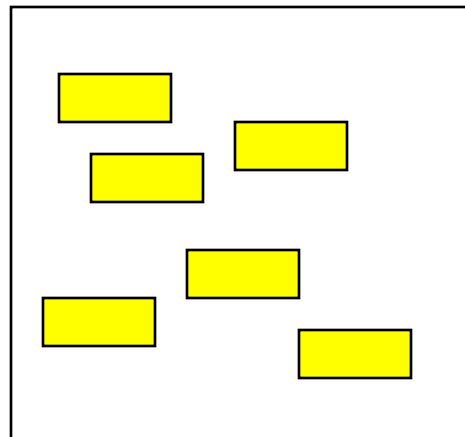
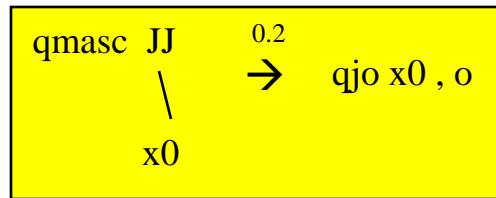
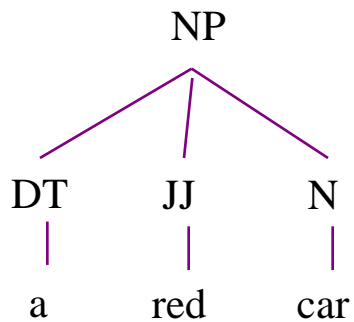
Transformation:

u , n

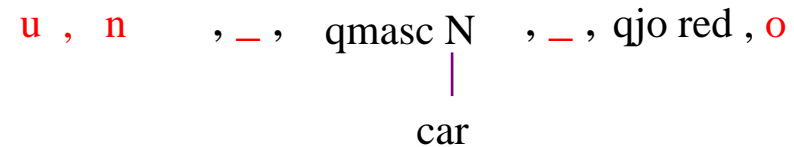


Possible Morpho-Syntactic SMT?

Original input:

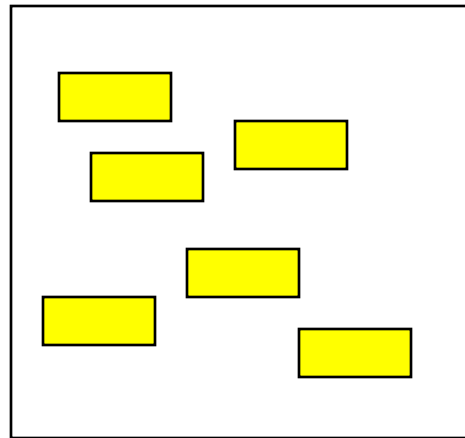
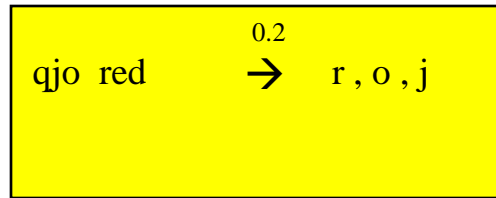
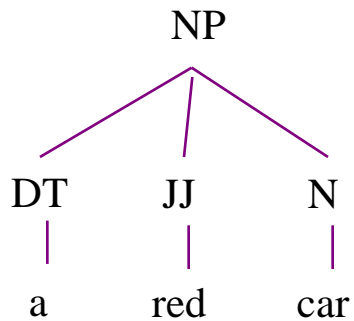


Transformation:

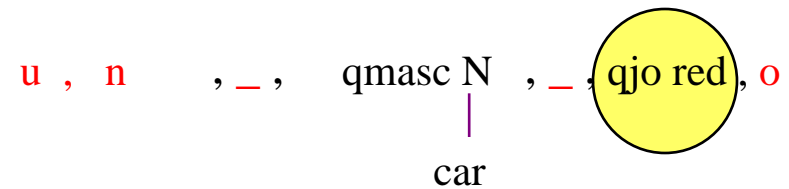


Possible Morpho-Syntactic SMT?

Original input:

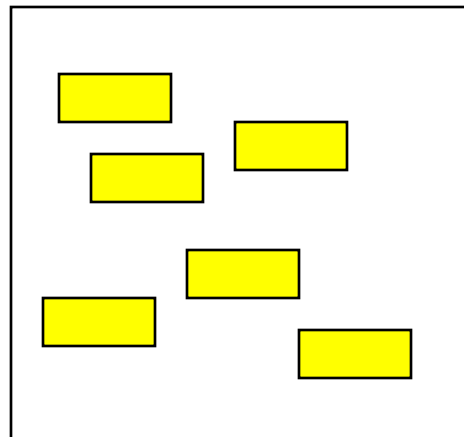
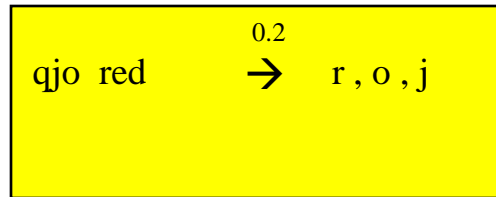
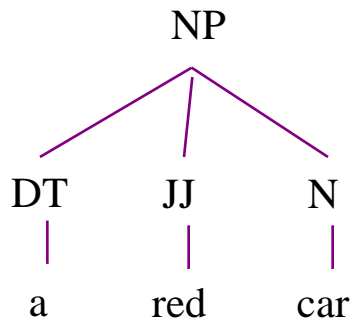


Transformation:

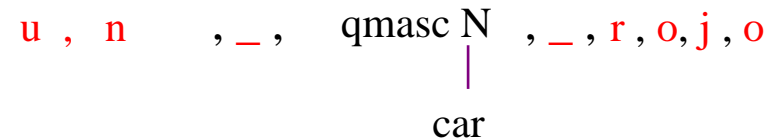


Possible Morpho-Syntactic SMT?

Original input:

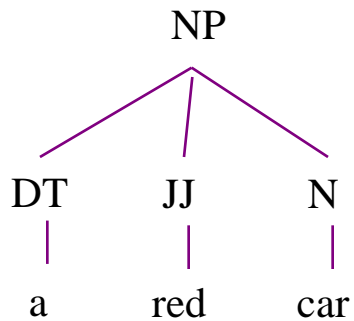


Transformation:

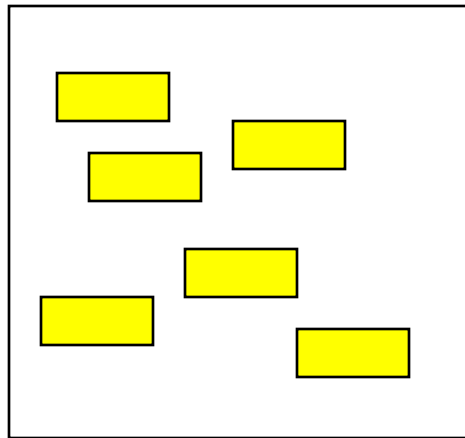


Possible Morpho-Syntactic SMT?

Original input:



Transformation:



u , n , _ , 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!

Suckers!



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.

Suckers!



Summary

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

thanks