Statistical Machine Translation

Part 1: Morning Session

Philipp Koehn

10 September 2007

Before we begin...

- What is about to happen?
  - a journey through the methods of SMT systems
  - focus mostly on the (very) current
  - there will be some maths
- What is **not** about to happen?
  - a guide on how to use statistical machine translation
  - an introduction to tools used in statistical machine translation

Topics

- Philipp Koehn (morning)
  - Introduction
  - Word-based models and the EM algorithm
  - Decoding
  - Phrase-based models
- Kevin Knight (afternoon)
  - Syntax-based statistical MT
  - Learning syntax models from data
  - Decoding for syntax models
  - Tree Automata
- This will take a while...

Fair warning

- Quotes:

  It must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term.

  Noam Chomsky, 1969

  Whenever I fire a linguist our system performance improves.

  Frederick Jelinek, 1988

- Warning: We may agree more with Jelinek than Chomsky (well, at least we know people who do)

Machine translation

- Task: make sense of foreign text like

毒品
本册子爲家長們提供實際和有用的關于毒品
的信息，包括如何減少使用非法毒品的危險.
它有助於您和您的家人討論有關毒品的問題.
這本小册子的主要内容已錄在磁帶上. 如果您
想索取一盒免費的磁帶(中文), 請在下面的

- One of the oldest problems in Artificial Intelligence

- AI-hard: reasoning and world knowledge required

The Rosetta stone

- Egyptian language was a mystery for centuries
- 1799 a stone with Egyptian text and its translation into Greek was found

⇒ Humans *could learn* how to translated Egyptian

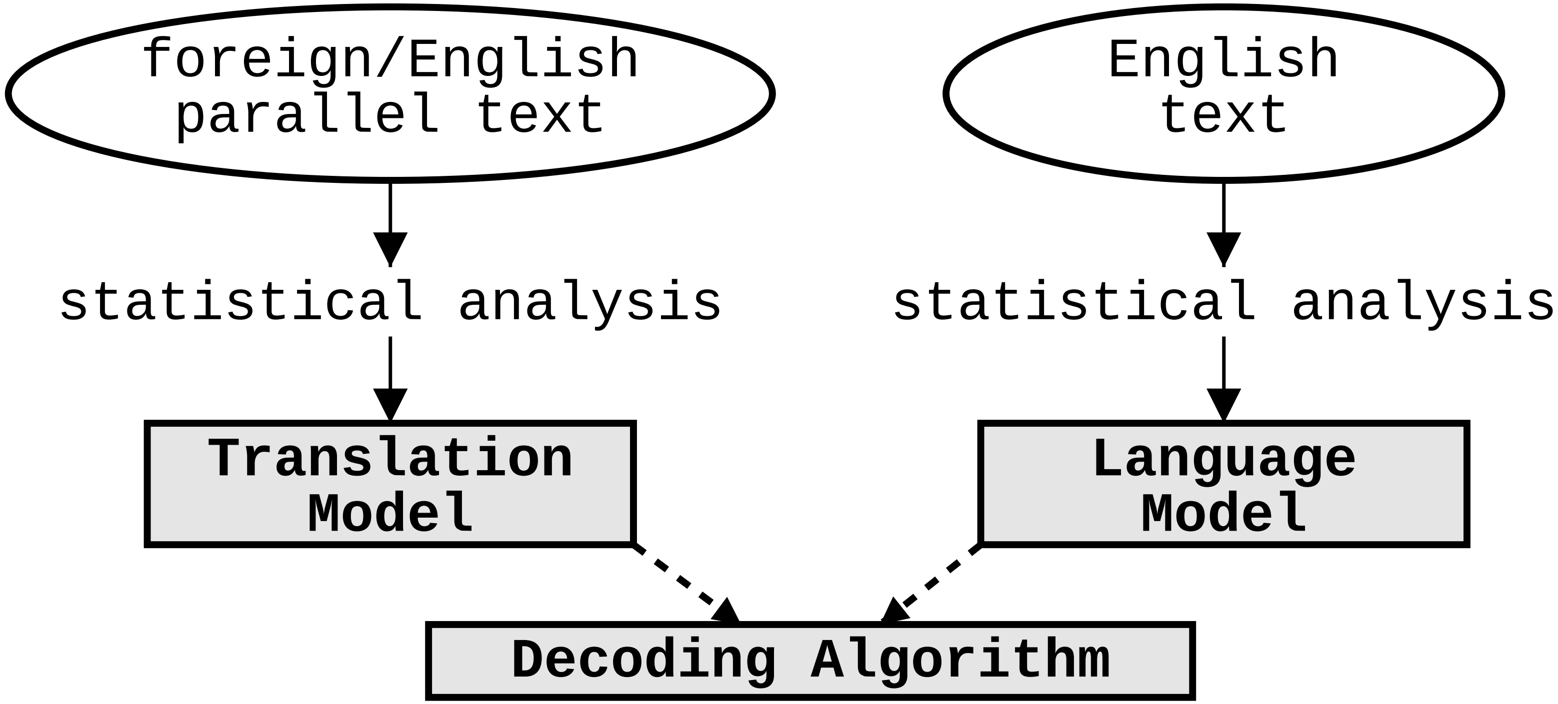
Parallel data

- Lots of translated text available: 100s of million words of translated text for some language pairs
  - a book has a few 100,000s words
  - an educated person may read 10,000 words a day
  - → 3.5 million words a year
  - → *300 million a lifetime*
  - → soon computers will be able to see more translated text than humans read in a lifetime

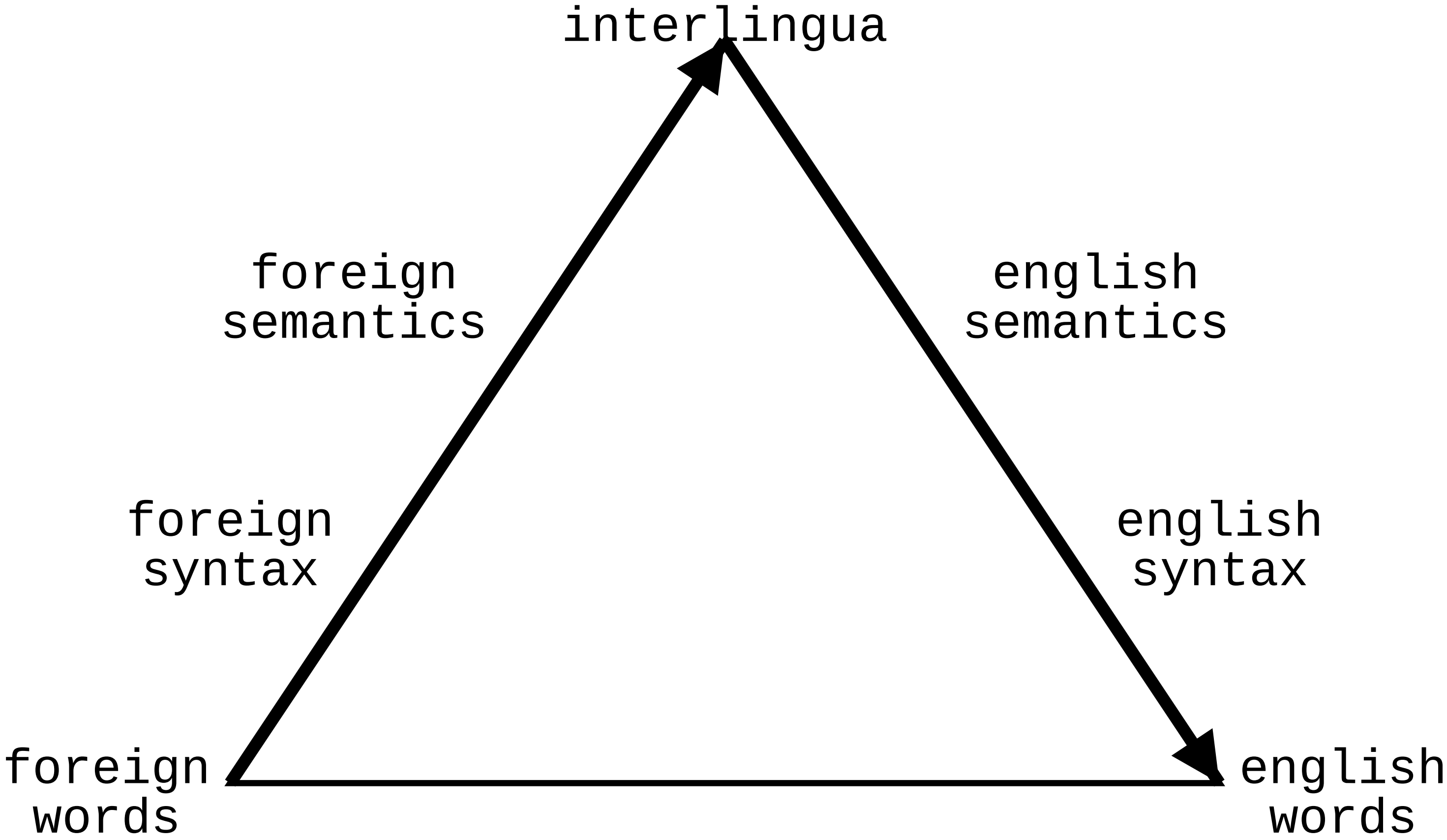
⇒ Machine *can learn* how to translated foreign languages

Statistical machine translation

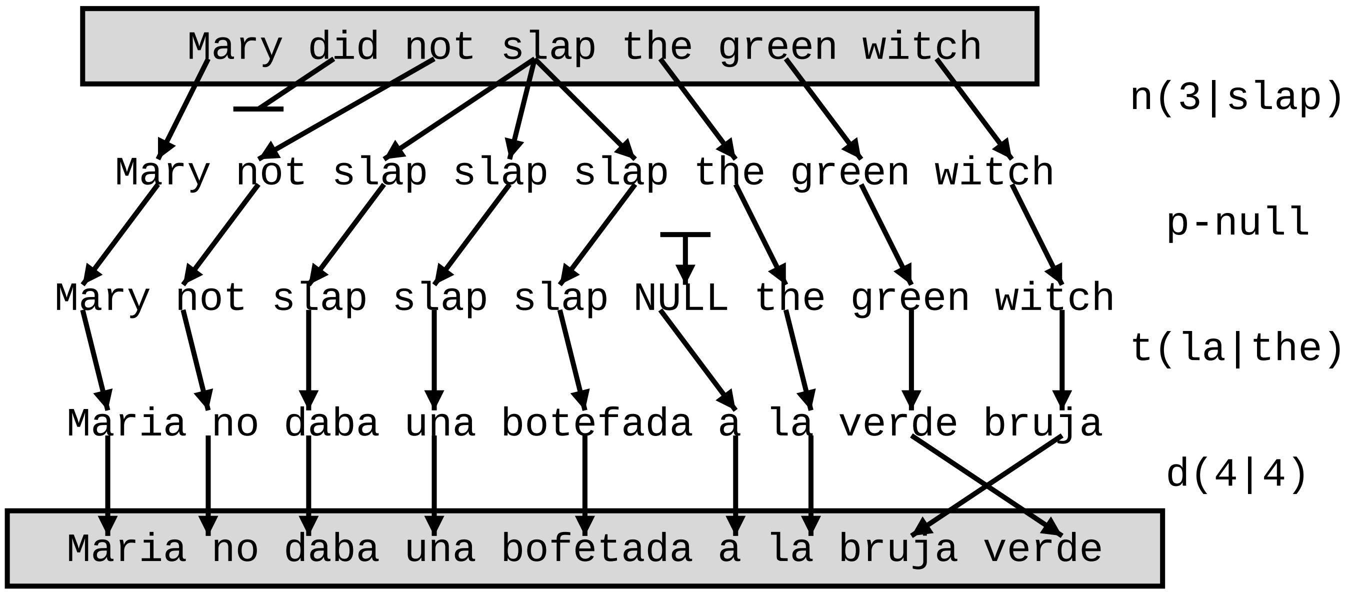
- Components: **Translation model**, **language model**, **decoder**



The machine translation pyramid



Word-based models

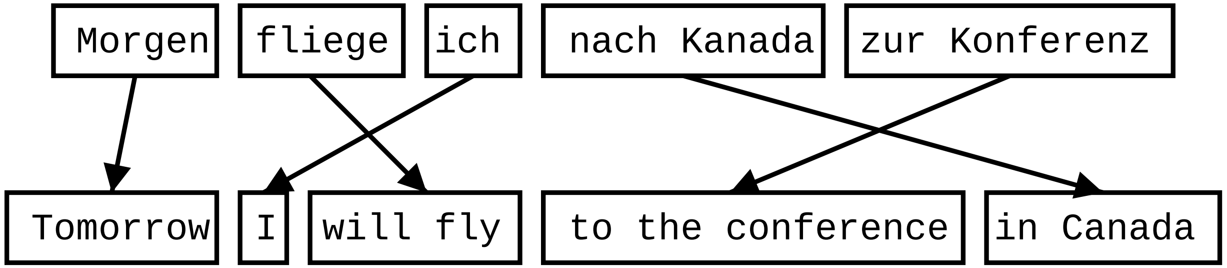


[from Knight, 1997]

- Translation process is *decomposed into smaller steps*, each is tied to words

- Original models for statistical machine translation [Brown et al., 1993]

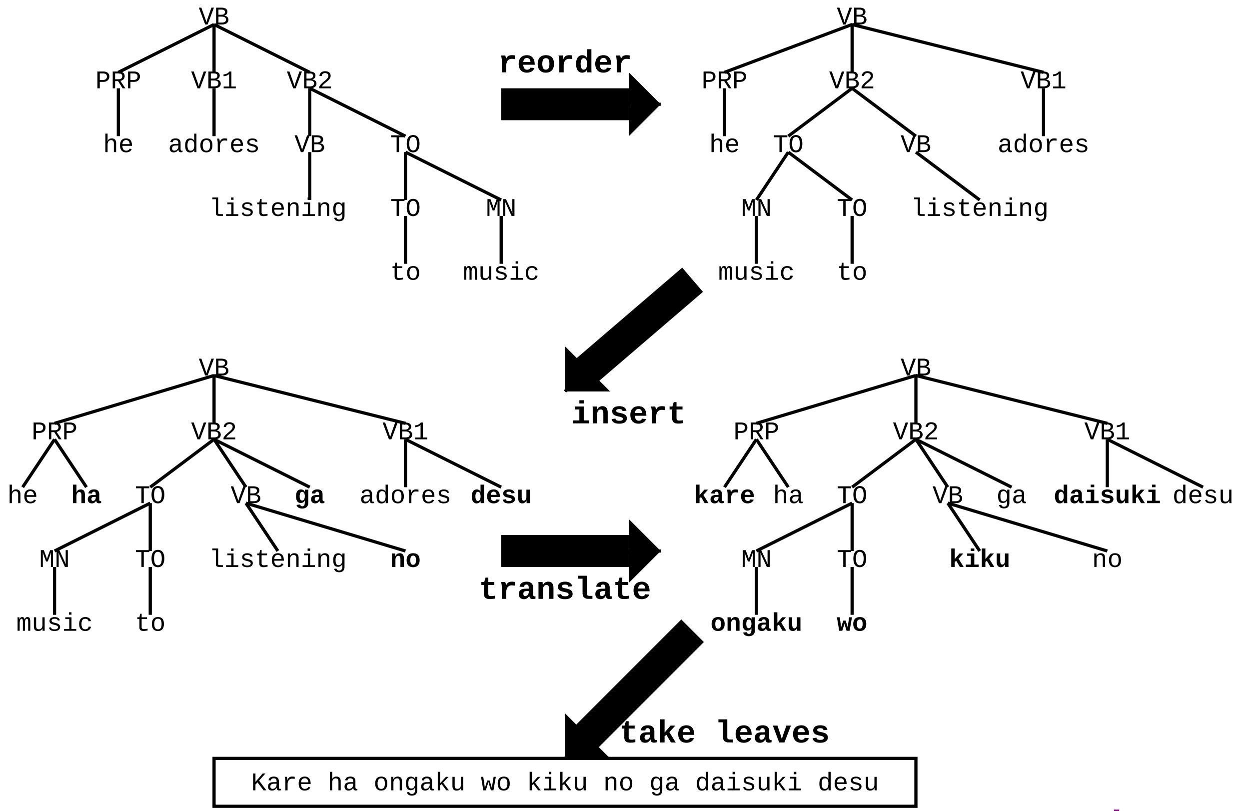
Phrase-based models



[from Koehn et al., 2003, NAACL]

- Foreign input is segmented in **phrases**
  - *any sequence of words*, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered

Syntax-based models



[from Yamada and Knight, 2001]

Automatic evaluation

- Why **automatic evaluation** metrics?
  - Manual evaluation is *too slow*
  - Evaluation on large test sets *reveals minor improvements*
  - **Automatic tuning** to improve machine translation performance

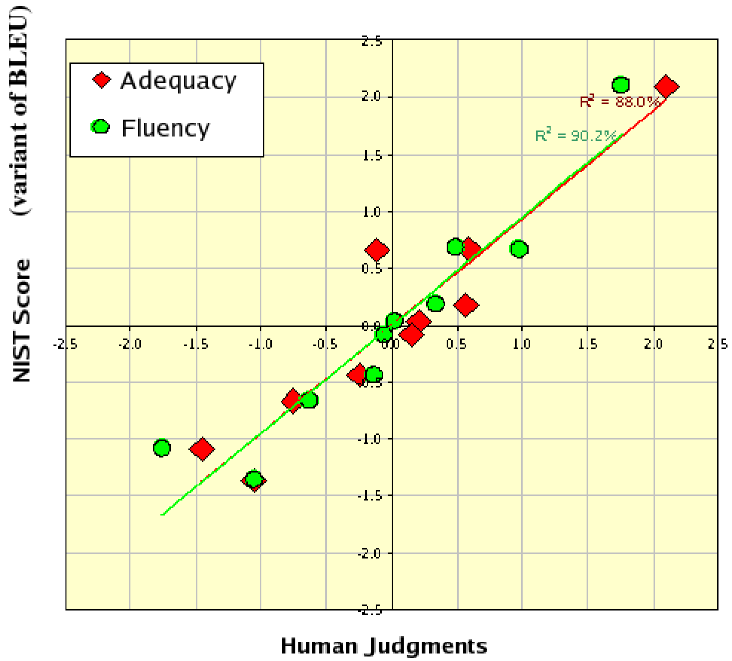
- History
  - Word Error Rate
  - **BLEU** since 2002

- BLEU in short: *Overlap with reference* translations

Automatic evaluation

- Reference Translation
  - the gunman was shot to death by the police .
- System Translations
  - **the gunman was police kill .**
  - **wounded police jaya of**
  - **the gunman was shot dead by the police .**
  - **the gunman arrested by police kill .**
  - **the gunmen were killed .**
  - **the gunman was shot to death by the police .**
  - **gunmen were killed by police ?SUB>0 ?SUB>0**
  - **al by the police .**
  - **the ringer is killed by the police .**
  - **police killed the gunman .**
- Matches
  - green = 4 gram match (good!)
  - red = word not matched (bad!)

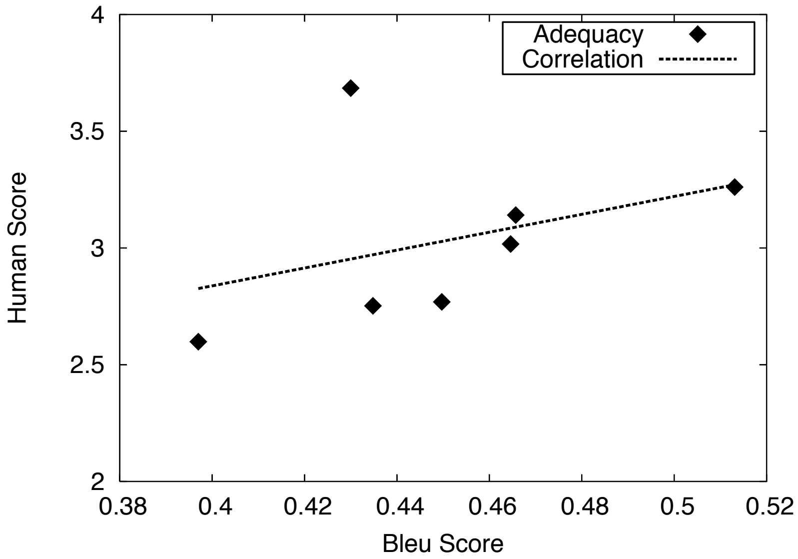
Automatic evaluation

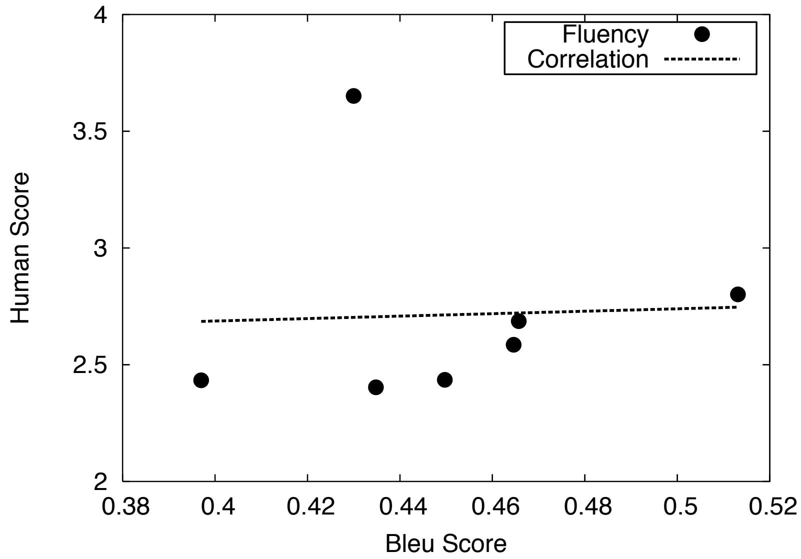


[from George Doddington, NIST]

- BLEU **correlates** with human judgement
  - **multiple reference translations** may be used

Correlation? [Callison-Burch et al., 2006]



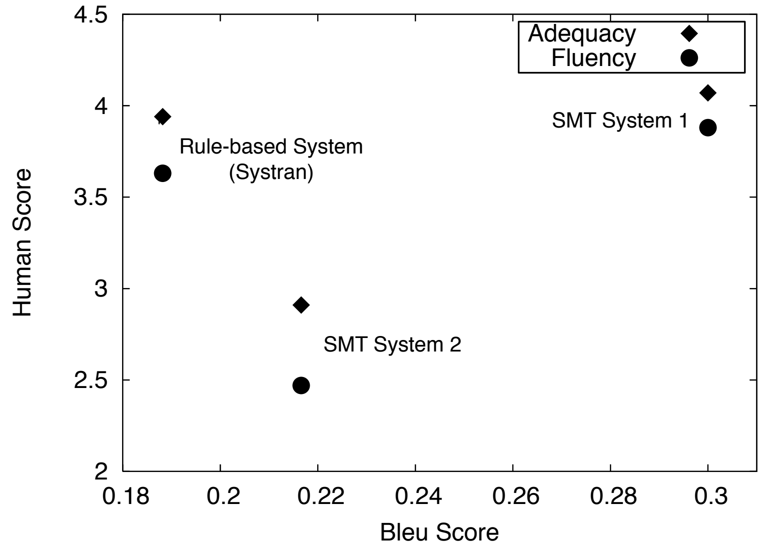


[from Callison-Burch et al., 2006, EACL]

- DARPA/NIST MT Eval 2005
  - Mostly statistical systems (all but one in graphs)
  - One submission **manual post-edit** of statistical system's output

  $\rightarrow$ Good adequacy/fluency scores *not reflected* by BLEU

Correlation? [Callison-Burch et al., 2006]



[from Callison-Burch et al., 2006, EACL]

- Comparison of
  - *good statistical* system: high BLEU, high adequacy/fluency
  - *bad statistical* sys. (trained on less data): low BLEU, low adequacy/fluency
  - *Systran*: lowest BLEU score, but high adequacy/fluency

Automatic evaluation: outlook

- Research questions
  - why does BLEU *fail* Systran and manual post-edits?
  - how can this *overcome* with novel evaluation metrics?

- Future of automatic methods
  - automatic metrics too *useful* to be abandoned
  - evidence still supports that during *system development*, a better BLEU indicates a better system
  - *final assessment* has to be human judgement

Competitions

- Progress driven by **MT Competitions**
  - **NIST/DARPA**: Yearly campaigns for Arabic-English, Chinese-English, newstexts, since 2001
  - **IWSLT**: Yearly competitions for Asian languages and Arabic into English, speech travel domain, since 2003
  - **WPT/WMT**: Yearly competitions for European languages, European Parliament proceedings, since 2005

- Increasing number of statistical MT groups participate

Euromatrix

- Proceedings of the European Parliament
  - translated into *11 official languages*
  - entry of new members in May 2004: more to come...
- Europarl corpus
  - collected 20-30 million words per language
  - → *110 language pairs*
- 110 Translation systems
  - 3 weeks on 16-node cluster computer
  - → *110 translation systems*

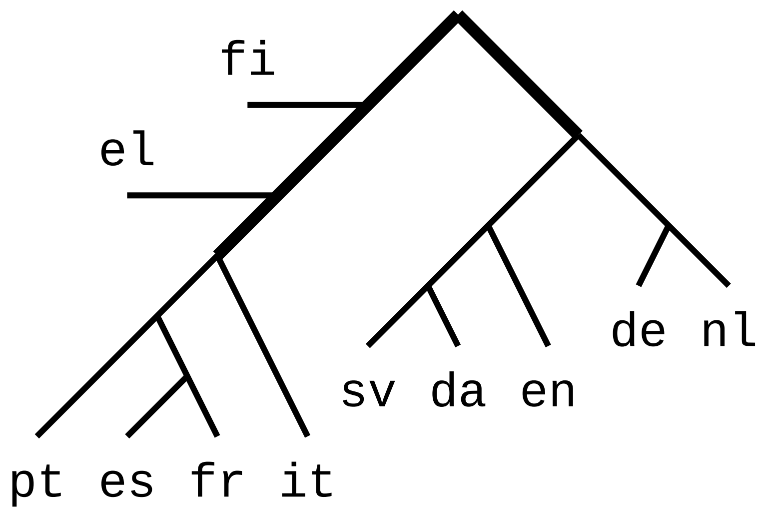
Quality of translation systems

- *Scores* for all 110 systems `http://www.statmt.org/matrix/`

	da	de	el	en	es	fr	fi	it	nl	pt	sv
da	-	**18.4**	**21.1**	**28.5**	**26.4**	**28.7**	**14.2**	**22.2**	**21.4**	**24.3**	**28.3**
de	**22.3**	-	**20.7**	**25.3**	**25.4**	**27.7**	**11.8**	**21.3**	**23.4**	**23.2**	**20.5**
el	**22.7**	**17.4**	-	**27.2**	**31.2**	**32.1**	**11.4**	**26.8**	**20.0**	**27.6**	**21.2**
en	**25.2**	**17.6**	**23.2**	-	**30.1**	**31.1**	**13.0**	**25.3**	**21.0**	**27.1**	**24.8**
es	**24.1**	**18.2**	**28.3**	**30.5**	-	**40.2**	**12.5**	**32.3**	**21.4**	**35.9**	**23.9**
fr	**23.7**	**18.5**	**26.1**	**30.0**	**38.4**	-	**12.6**	**32.4**	**21.1**	**35.3**	**22.6**
fi	**20.0**	**14.5**	**18.2**	**21.8**	**21.1**	**22.4**	-	**18.3**	**17.0**	**19.1**	**18.8**
it	**21.4**	**16.9**	**24.8**	**27.8**	**34.0**	**36.0**	**11.0**	-	**20.0**	**31.2**	**20.2**
nl	**20.5**	**18.3**	**17.4**	**23.0**	**22.9**	**24.6**	**10.3**	**20.0**	-	**20.7**	**19.0**
pt	**23.2**	**18.2**	**26.4**	**30.1**	**37.9**	**39.0**	**11.9**	**32.0**	**20.2**	-	**21.9**
sv	**30.3**	**18.9**	**22.8**	**30.2**	**28.6**	**29.7**	**15.3**	**23.9**	**21.9**	**25.9**	-

[from Koehn, 2005: Europarl]

Clustering languages



[from Koehn, 2005, MT Summit]

- **Clustering** languages based on how easy they translate into each other

⇒ Approximation of language families

Translate into vs. out of a language

- Some languages are *easier* to translate into that out of

Language	From	Into	Diff
da	23.4	23.3	0.0
de	**22.2**	**17.7**	**-4.5**
el	23.8	22.9	-0.9
en	**23.8**	**27.4**	**+3.6**
es	26.7	29.6	+2.9
fr	26.1	31.1	+5.1
fi	19.1	12.4	-6.7
it	24.3	25.4	+1.1
nl	19.7	20.7	+1.1
pt	26.1	27.0	+0.9
sv	24.8	22.1	-2.6

[from Koehn, 2005: Europarl]

- *Morphologically rich languages* harder to generate (German, Finnish)

Backtranslations

- Checking translation quality by **back-transliteration**
- *The spirit is willing, but the flesh is weak*
- English → Russian → English
- *The vodka is good but the meat is rotten*

Backtranslations II

- *Does not correlate* with unidirectional performance

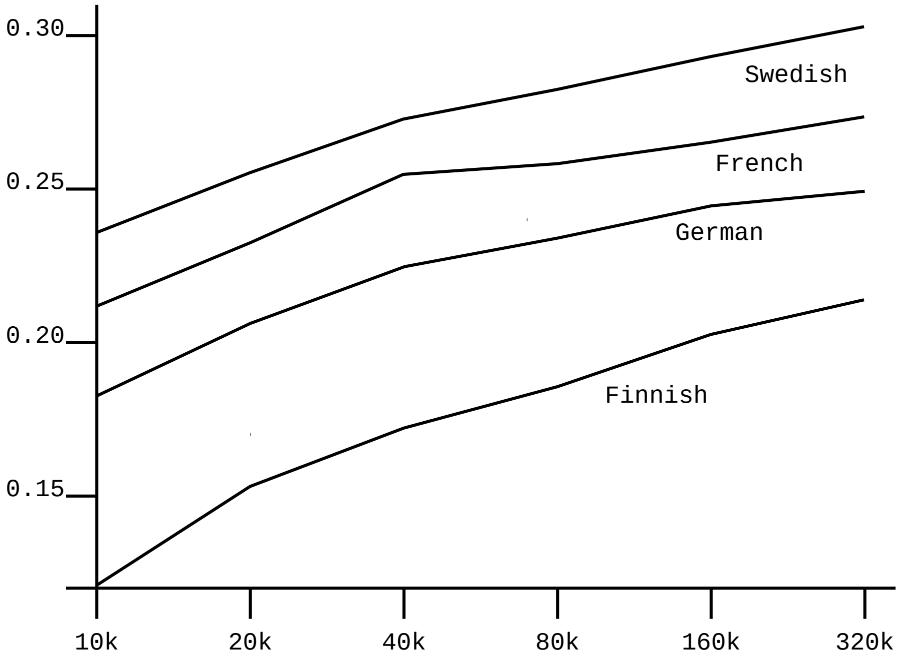
Language	From	Into	Back
da	28.5	25.2	56.6
de	25.3	17.6	48.8
el	27.2	23.2	**56.5**
es	30.5	30.1	52.6
fi	21.8	13.0	44.4
it	27.8	25.3	49.9
nl	23.0	21.0	46.0
pt	30.1	27.1	**53.6**
sv	30.2	24.8	54.4

[from Koehn, 2005: Europarl]

Available data

- Available *parallel text*
  - **Europarl**: *30 million words* in 11 languages http://www.statmt.org/europarl/
  - **Acquis Communitaire**: *8-50 million words* in 20 EU languages
  - **Canadian Hansards**: *20 million words* from Ulrich Germann, ISI
  - Chinese/Arabic to English: *over 100 million words* from **LDC**
  - lots more French/English, Spanish/French/English from **LDC**
- Available monolingual text (for language modeling)
  - *2.8 billion words* of English from **LDC**
  - *100s of billions, trillions* on the web

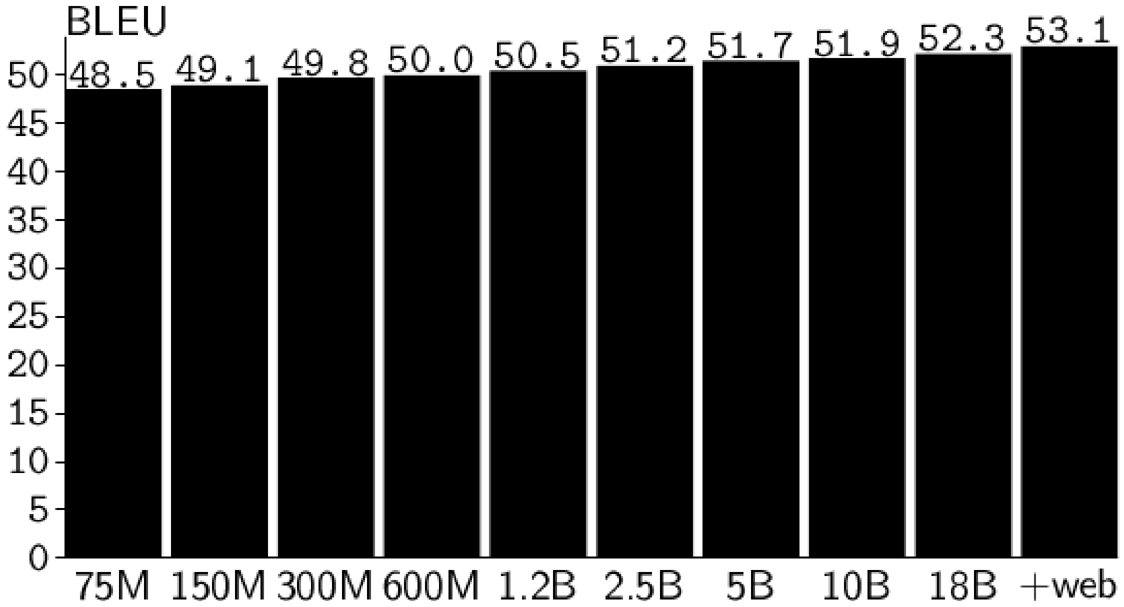
More data, better translations



[from Koehn, 2003: Europarl]

- **Log-scale improvements** on BLEU:
  Doubling the training data gives constant improvement *(+1 %BLEU)*

More LM data, better translations



[from Och, 2005: MT Eval presentation]

- Also **log-scale improvements** on BLEU:
  doubling the training data gives constant improvement *(+0.5 %BLEU)*
  (last addition is 218 billion words out-of-domain web data)

Output of Chinese-English system

In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong's export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province's total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong's export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province's total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion us dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Partially excellent translations

In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong's export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province's total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong's export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province's total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion US dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Mangled grammar

In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong's export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province's total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong's export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province's total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion us dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Word-based models and the EM algorithm

Lexical translation

- How to translate a word → look up in dictionary

  Haus — *house, building, home, household, shell.*

- *Multiple translations*
  - some more frequent than others
  - for instance: *house*, and *building* most common
  - special cases: *Haus* of a *snail* is its *shell*

- Note: During all the lectures, we will translate from a foreign language into English

Collect statistics

- Look at a *parallel corpus* (German text along with English translation)

Translation of $\boldsymbol{Haus}$	**Count**
house	8,000
building	1,600
home	200
household	150
shell	50

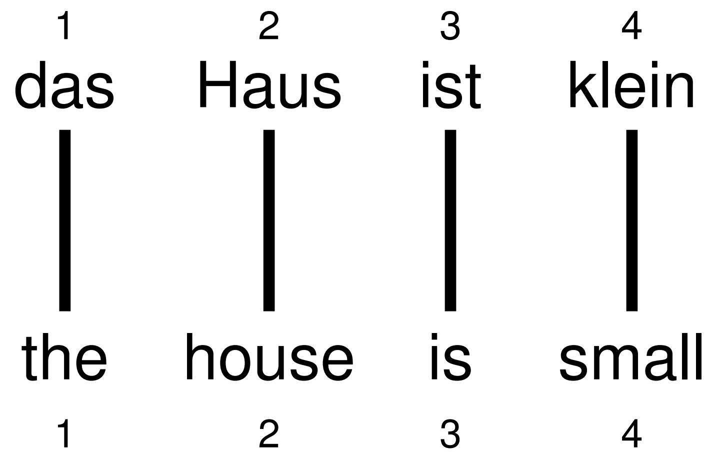
Estimate translation probabilities

- *Maximum likelihood estimation*

$$p_f(e) = \begin{cases} 0.8 & \text{if } e = \textit{house}, \\ 0.16 & \text{if } e = \textit{building}, \\ 0.02 & \text{if } e = \textit{home}, \\ 0.015 & \text{if } e = \textit{household}, \\ 0.005 & \text{if } e = \textit{shell}. \end{cases}$$

Alignment

- In a parallel text (or when we translate), we **align** words in one language with the words in the other



- Word *positions* are numbered 1–4

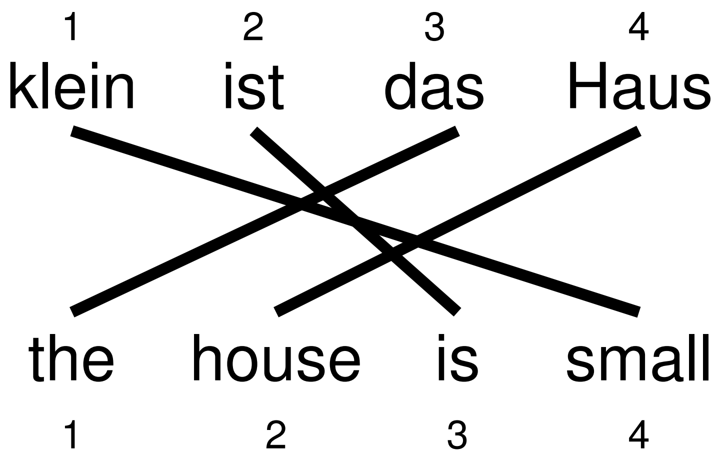
Alignment function

- Formalizing *alignment* with an **alignment function**
- Mapping an English target word at position i to a German source word at position j with a function $a : i \rightarrow j$
- Example

$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4\}$$

Reordering

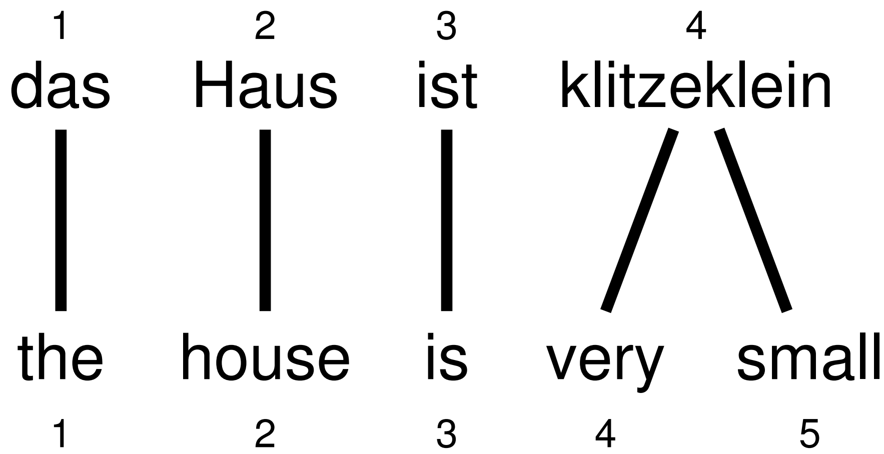
- Words may be **reordered** during translation



$$a : \{1 \rightarrow 3, 2 \rightarrow 4, 3 \rightarrow 2, 4 \rightarrow 1\}$$

One-to-many translation

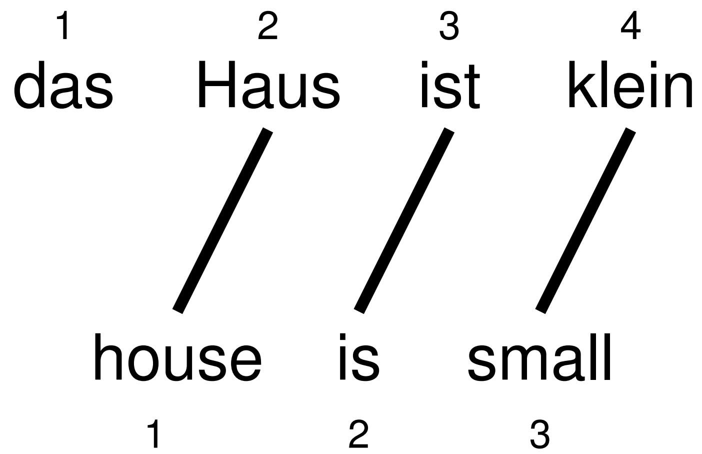
- A source word may translate into **multiple** target words



$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4, 5 \rightarrow 4\}$$

Dropping words

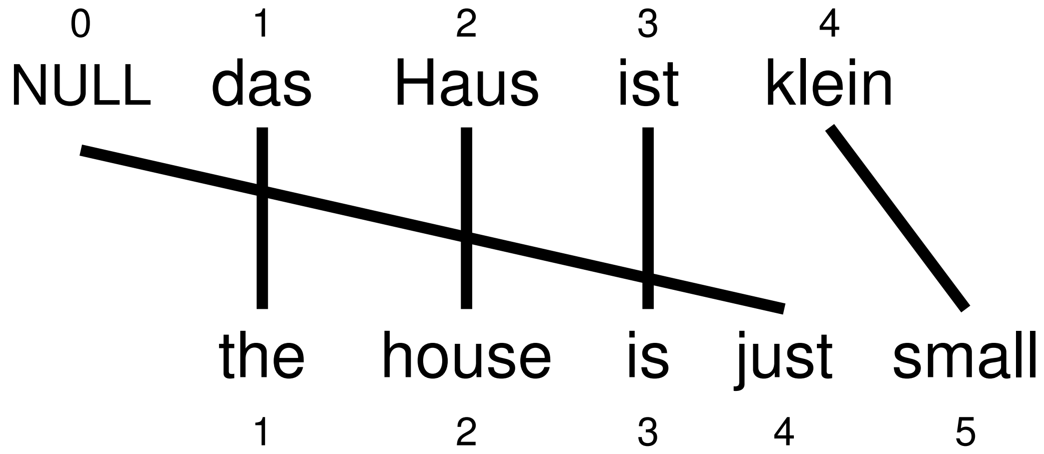
- Words may be **dropped** when translated
  - The German article *das* is dropped



$$a : \{1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 4\}$$

Inserting words

- Words may be **added** during translation
  - The English *just* does not have an equivalent in German
  - We still need to map it to something: special NULL token



$$a : \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 0, 5 \rightarrow 4\}$$

IBM Model 1

- *Generative model*: break up translation process into smaller steps
  - **IBM Model 1** only uses *lexical translation*

- Translation probability
  - for a foreign sentence $\mathbf{f} = (f_1, ..., f_{l_f})$ of length l_f
  - to an English sentence $\mathbf{e} = (e_1, ..., e_{l_e})$ of length l_e
  - with an alignment of each English word e_j to a foreign word f_i according to the alignment function $a : j \rightarrow i$

$$p(\mathbf{e}, a|\mathbf{f}) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

  - parameter ϵ is a *normalization constant*

Example

das

e	$t(e\|f)$
the	0.7
that	0.15
which	0.075
who	0.05
this	0.025

Haus

e	$t(e\|f)$
house	0.8
building	0.16
home	0.02
household	0.015
shell	0.005

ist

e	$t(e\|f)$
is	0.8
's	0.16
exists	0.02
has	0.015
are	0.005

klein

e	$t(e\|f)$
small	0.4
little	0.4
short	0.1
minor	0.06
petty	0.04

$$\begin{aligned} p(e,a|f) &= \frac{\epsilon}{4^3} \times t(\text{the}|\text{das}) \times t(\text{house}|\text{Haus}) \times t(\text{is}|\text{ist}) \times t(\text{small}|\text{klein}) \\ &= \frac{\epsilon}{4^3} \times 0.7 \times 0.8 \times 0.8 \times 0.4 \\ &= 0.0028\epsilon \end{aligned}$$

Learning lexical translation models

- We would like to *estimate* the lexical translation probabilities $t(e|f)$ from a parallel corpus

- ... but we do not have the alignments

- **Chicken and egg problem**
  - if we had the *alignments*,
    $\rightarrow$ we could estimate the *parameters* of our generative model
  - if we had the *parameters*,
    $\rightarrow$ we could estimate the *alignments*

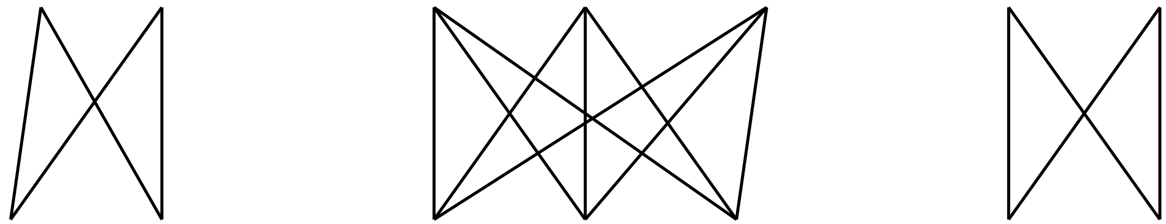
EM algorithm

- **Incomplete data**
  - if we had *complete data*, would could estimate *model*
  - if we had *model*, we could fill in the *gaps in the data*

- **Expectation Maximization (EM)** in a nutshell
  - initialize model parameters (e.g. uniform)
  - assign probabilities to the missing data
  - estimate model parameters from completed data
  - iterate

EM algorithm

... la maison ... la maison blue ... la fleur ...

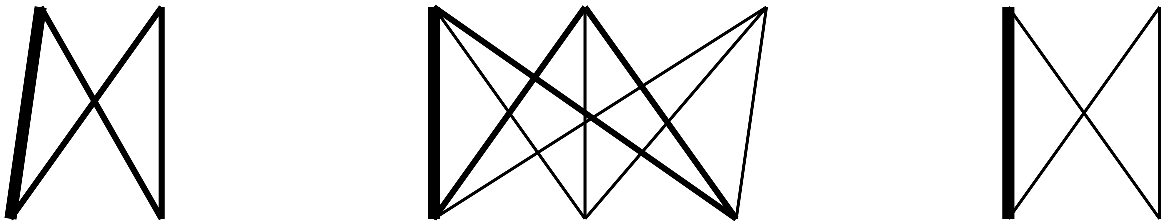


... the house ... the blue house ... the flower ...

- Initial step: all alignments equally likely
- Model learns that, e.g., *la* is often aligned with *the*

EM algorithm

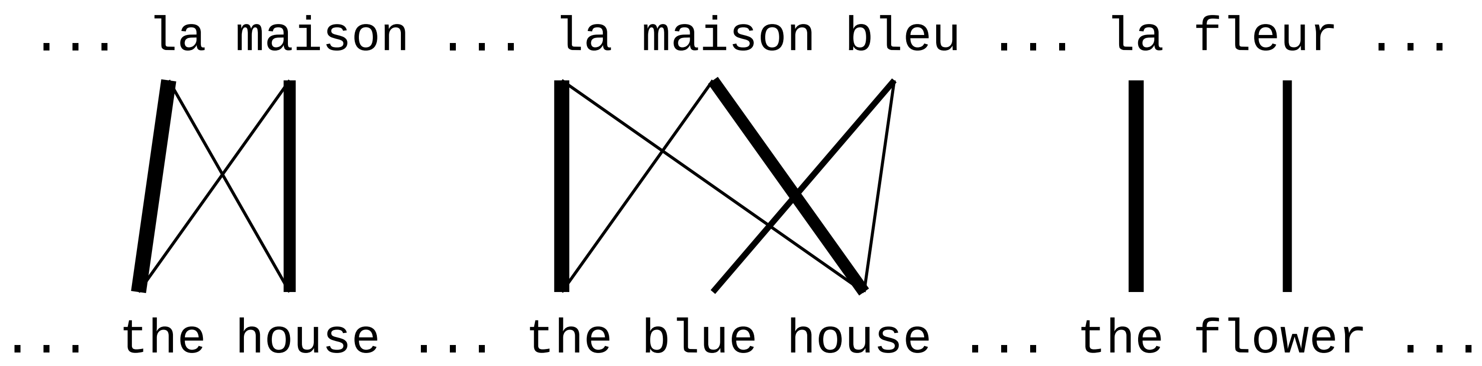
... la maison ... la maison blue ... la fleur ...



... the house ... the blue house ... the flower ...

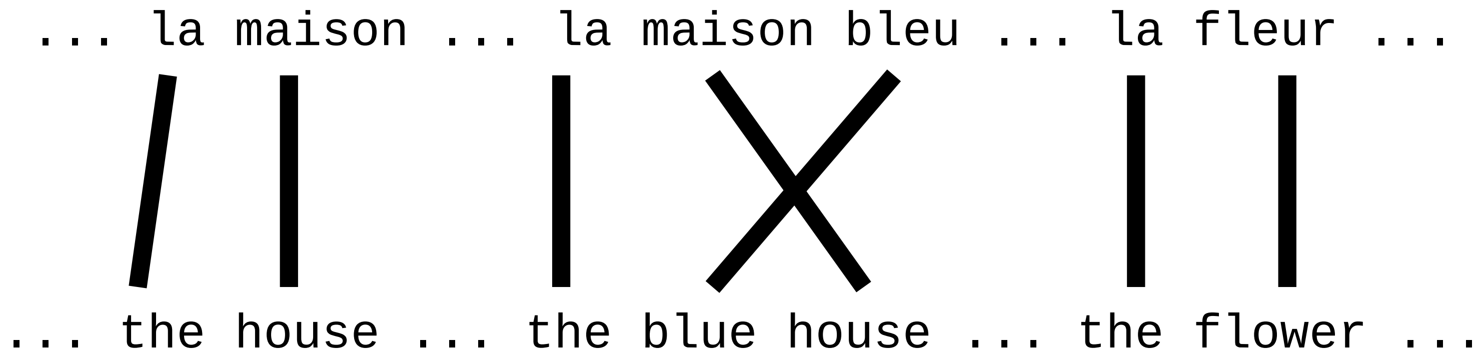
- After one iteration
- Alignments, e.g., between *la* and *the* are more likely

EM algorithm



- After another iteration
- It becomes apparent that alignments, e.g., between *fleur* and *flower* are more likely (**pigeon hole principle**)

EM algorithm



- Convergence
- Inherent hidden structure revealed by EM

EM algorithm

```
... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...
```

```
p(la|the) = 0.453
p(le|the) = 0.334
p(maison|house) = 0.876
p(bleu|blue) = 0.563
...
```

- Parameter estimation from the aligned corpus

IBM Model 1 and EM

- EM Algorithm consists of two steps

- **Expectation-Step**: Apply model to the data
  - parts of the model are hidden (here: alignments)
  - using the model, assign probabilities to possible values

- **Maximization-Step**: Estimate model from data
  - take assign values as fact
  - collect counts (weighted by probabilities)
  - estimate model from counts

- Iterate these steps until **convergence**

IBM Model 1 and EM

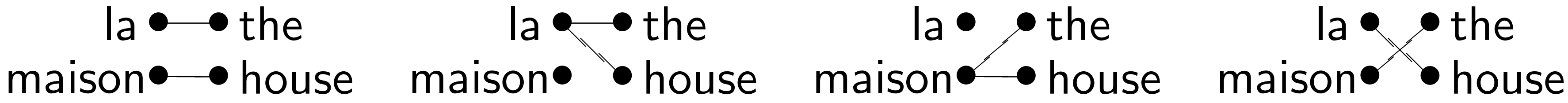
- We need to be able to compute:
  - Expectation-Step: probability of alignments
  - Maximization-Step: count collection

IBM Model 1 and EM

- **Probabilities**

$$p(\text{the}|\text{la}) = 0.7 \qquad p(\text{house}|\text{la}) = 0.05$$
$$p(\text{the}|\text{maison}) = 0.1 \qquad p(\text{house}|\text{maison}) = 0.8$$

- **Alignments**



$$p(\mathbf{e},a|\mathbf{f}) = 0.56 \qquad p(\mathbf{e},a|\mathbf{f}) = 0.035 \qquad p(\mathbf{e},a|\mathbf{f}) = 0.08 \qquad p(\mathbf{e},a|\mathbf{f}) = 0.005$$

$$p(a|\mathbf{e},\mathbf{f}) = 0.824 \qquad p(a|\mathbf{e},\mathbf{f}) = 0.052 \qquad p(a|\mathbf{e},\mathbf{f}) = 0.118 \qquad p(a|\mathbf{e},\mathbf{f}) = 0.007$$

- **Counts**

$$c(\text{the}|\text{la}) = 0.824 + 0.052 \qquad c(\text{house}|\text{la}) = 0.052 + 0.007$$
$$c(\text{the}|\text{maison}) = 0.118 + 0.007 \qquad c(\text{house}|\text{maison}) = 0.824 + 0.118$$

IBM Model 1 and EM: Expectation Step

- We need to compute $p(a|\mathbf{e},\mathbf{f})$

- Applying the *chain rule*:

$$p(a|\mathbf{e},\mathbf{f}) = \frac{p(\mathbf{e},a|\mathbf{f})}{p(\mathbf{e}|\mathbf{f})}$$

- We already have the formula for $p(\mathbf{e},\mathbf{a}|\mathbf{f})$ (definition of Model 1)

IBM Model 1 and EM: Expectation Step

- We need to compute $p(\mathbf{e}|\mathbf{f})$

$$p(\mathbf{e}|\mathbf{f}) = \sum_a p(\mathbf{e}, a|\mathbf{f})$$

$$= \sum_{a(1)=0}^{l_f} \dots \sum_{a(l_e)=0}^{l_f} p(\mathbf{e}, a|\mathbf{f})$$

$$= \sum_{a(1)=0}^{l_f} \dots \sum_{a(l_e)=0}^{l_f} \frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

IBM Model 1 and EM: Expectation Step

$$p(\mathbf{e}|\mathbf{f}) = \sum_{a(1)=0}^{l_f} \dots \sum_{a(l_e)=0}^{l_f} \frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

$$= \frac{\epsilon}{(l_f+1)^{l_e}} \sum_{a(1)=0}^{l_f} \dots \sum_{a(l_e)=0}^{l_f} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

$$= \frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_j|f_i)$$

- Note the trick in the last line
  - removes the need for an *exponential* number of products

  $\rightarrow$ this makes IBM Model 1 estimation **tractable**

IBM Model 1 and EM: Expectation Step

- Combine what we have:

$$p(\mathbf{a}|\mathbf{e},\mathbf{f}) = p(\mathbf{e},\mathbf{a}|\mathbf{f})/p(\mathbf{e}|\mathbf{f})$$

$$= \frac{\frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})}{\frac{\epsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_j|f_i)}$$

$$= \prod_{j=1}^{l_e} \frac{t(e_j|f_{a(j)})}{\sum_{i=0}^{l_f} t(e_j|f_i)}$$

IBM Model 1 and EM: Maximization Step

- Now we have to *collect counts*

- Evidence from a sentence pair **e,f** that word e is a translation of word f:

$$c(e|f;\mathbf{e},\mathbf{f}) = \sum_a p(a|\mathbf{e},\mathbf{f}) \sum_{j=1}^{l_e} \delta(e,e_j)\delta(f,f_{a(j)})$$

- With the same simplication as before:

$$c(e|f;\mathbf{e},\mathbf{f}) = \frac{t(e|f)}{\sum_{j=1}^{l_e} t(e|f_{a(j)})} \sum_{j=1}^{l_e} \delta(e,e_j) \sum_{i=0}^{l_f} \delta(f,f_i)$$

IBM Model 1 and EM: Maximization Step

- After collecting these counts over a corpus, we can estimate the model:

$$t(e|f;\mathbf{e},\mathbf{f}) = \frac{\sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f}))}{\sum_f \sum_{(\mathbf{e},\mathbf{f})} c(e|f;\mathbf{e},\mathbf{f}))}$$

IBM Model 1 and EM: Pseudocode

```
initialize t(e|f) uniformly
do
  set count(e|f) to 0 for all e,f
  set total(f) to 0 for all f
  for all sentence pairs (e_s,f_s)
    for all words e in e_s
      total_s = 0
      for all words f in f_s
        total_s += t(e|f)
    for all words e in e_s
      for all words f in f_s
        count(e|f) += t(e|f) / total_s
        total(f)   += t(e|f) / total_s
  for all f in domain( total(.) )
    for all e in domain( count(.|f) )
      t(e|f) = count(e|f) / total(f)
until convergence
```

Higher IBM Models

IBM Model 1	lexical translation
IBM Model 2	adds absolute **reordering model**
IBM Model 3	adds **fertility model**
IBM Model 4	relative reordering model
IBM Model 5	fixes **deficiency**

- Only IBM Model 1 has *global maximum*
  - training of a higher IBM model builds on previous model
- Compuationally biggest change in Model 3
  - trick to simplify estimation does not work anymore

  $\rightarrow$ *exhaustive* count collection becomes computationally too expensive
  - **sampling** over high probability alignments is used instead

IBM Model 4

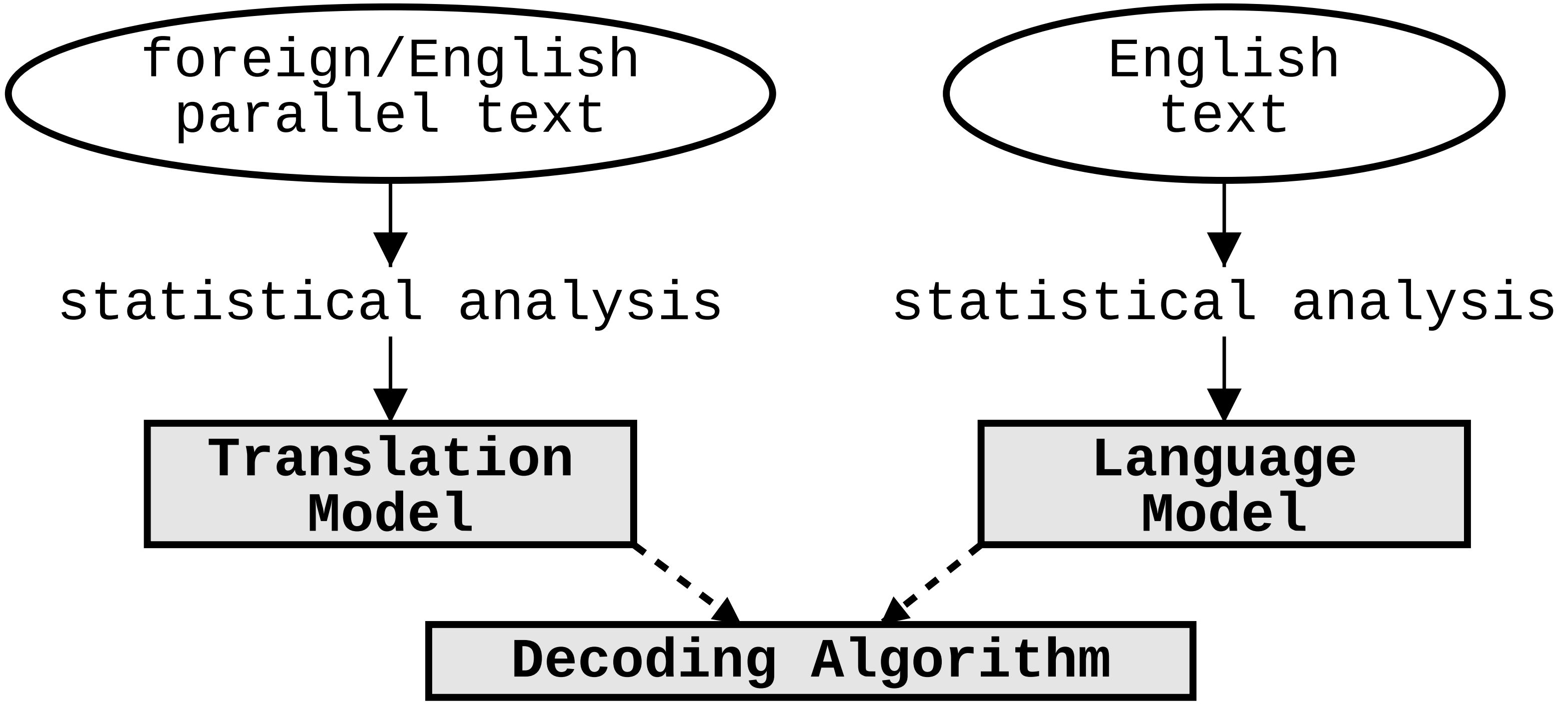

Mary did not slap the green witch
n(3|slap)
Mary not slap slap slap the green witch
p-null
Mary not slap slap slap NULL the green witch
t(la|the)
Maria no daba una botefada a la verde bruja
d(4|4)
Maria no daba una bofetada a la bruja verde

Late morning session

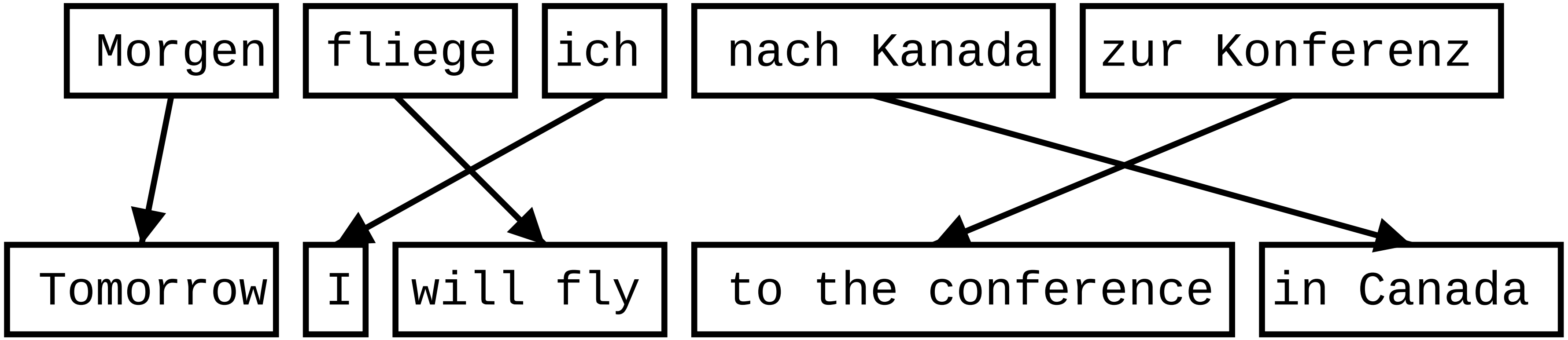
- Decoding
- Phrase-based models

Statistical Machine Translation

- Components: Translation model, language model, decoder



Phrase-Based Translation



- Foreign input is segmented in phrases
  - any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered

Phrase Translation Table

- Phrase Translations for "den Vorschlag":

English	$\phi(\mathbf{e}\vert\mathbf{f})$	**English**	$\phi(\mathbf{e}\vert\mathbf{f})$
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

Decoding Process

Maria	no	dio	una	bofetada	a	la	bruja	verde

- Build translation left to right
  - *select foreign* words to be translated

Decoding Process

- Build translation *left to right*
  - select foreign words to be translated
  - *find English* phrase translation
  - *add English* phrase to end of partial translation

Decoding Process

Maria	no	dio	una	bofetada	a	la	bruja	verde

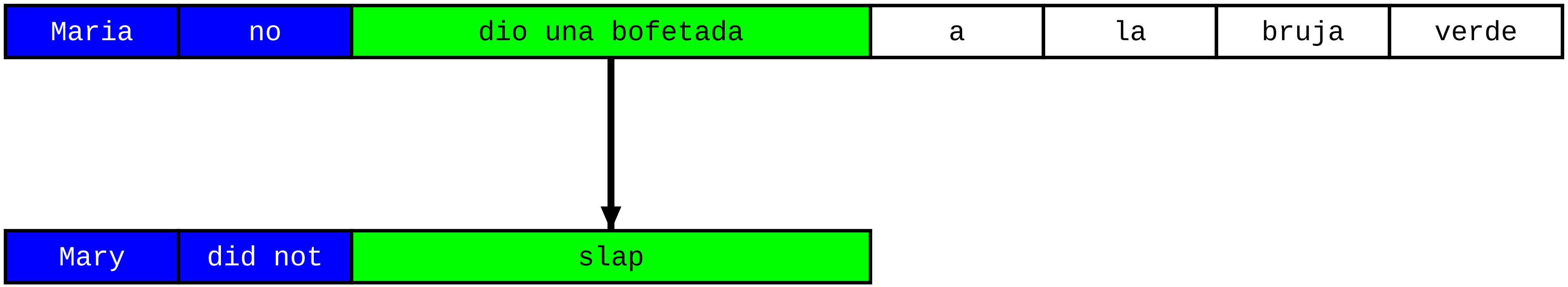
Mary

- Build translation left to right
  - select foreign words to be translated
  - find English phrase translation
  - add English phrase to end of partial translation
  - *mark foreign* words as translated

Decoding Process

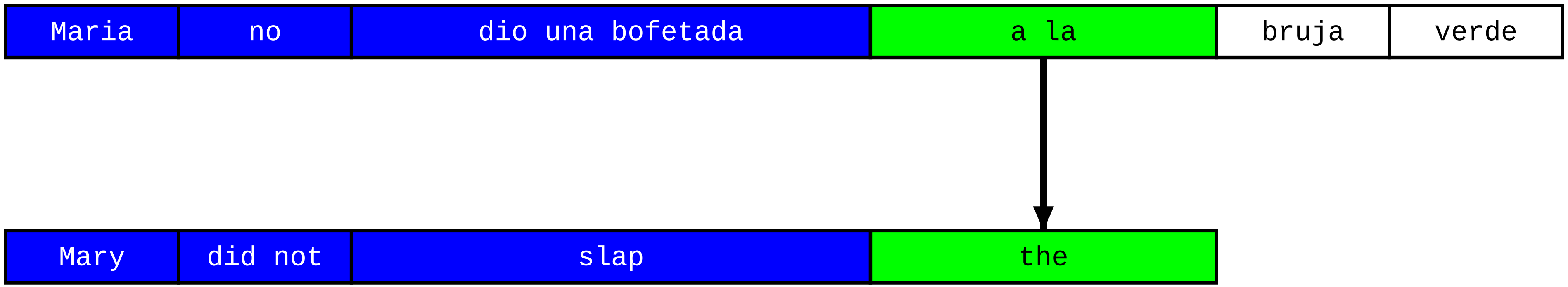
- *One to many* translation

Decoding Process



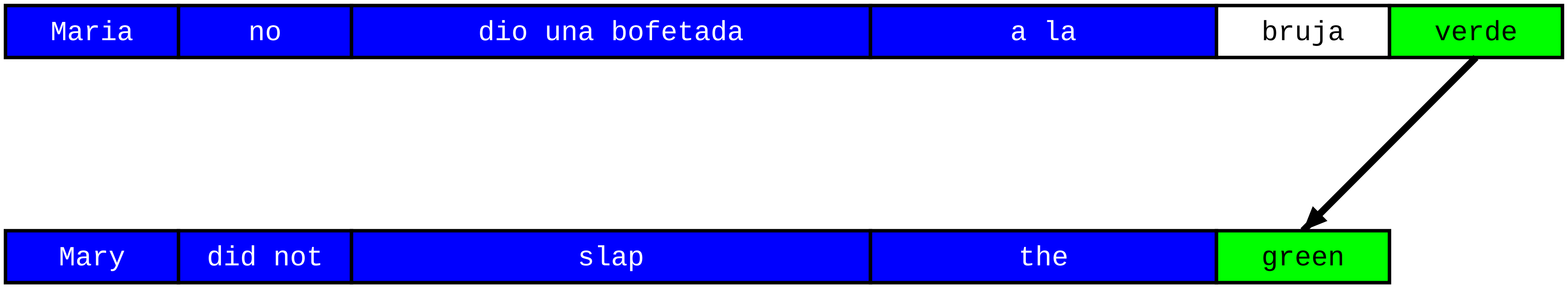
- Many to one translation

Decoding Process



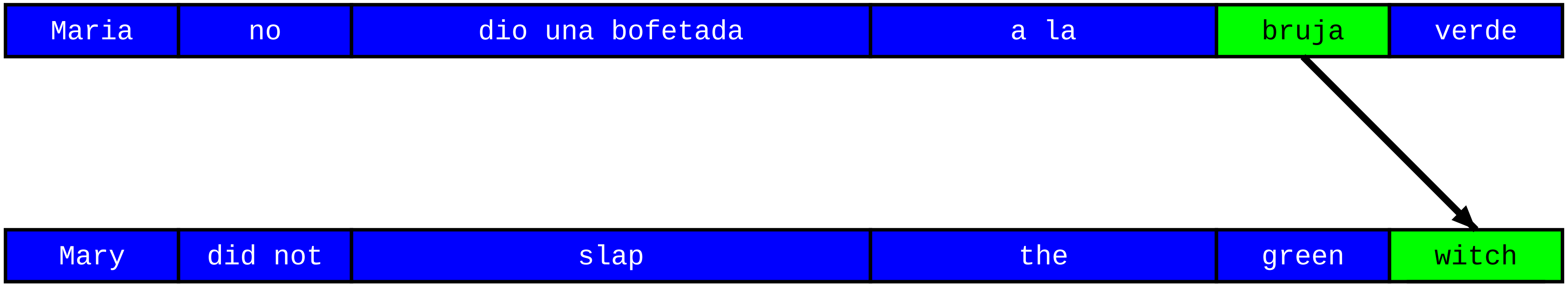
- *Many to one* translation

Decoding Process



- *Reordering*

Decoding Process



- Translation *finished*

Translation Options

- Look up *possible phrase translations*
  - many different ways to *segment* words into phrases
  - many different ways to *translate* each phrase

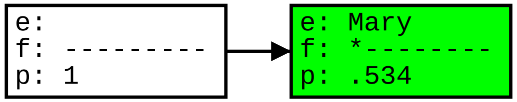
Hypothesis Expansion

Maria	no	dio	una	bofetada	a	la	bruja	verde
Mary	not	give	a	slap	to	the	witch	green
	did not		a slap		by		green witch	
	no	slap			to the			
	did not give				to			
					the			
		slap				the witch		

```
e:
f: ---------
p: 1
```

- Start with **empty hypothesis**
  - e: no English words
  - f: no foreign words covered
  - p: probability 1

Hypothesis Expansion

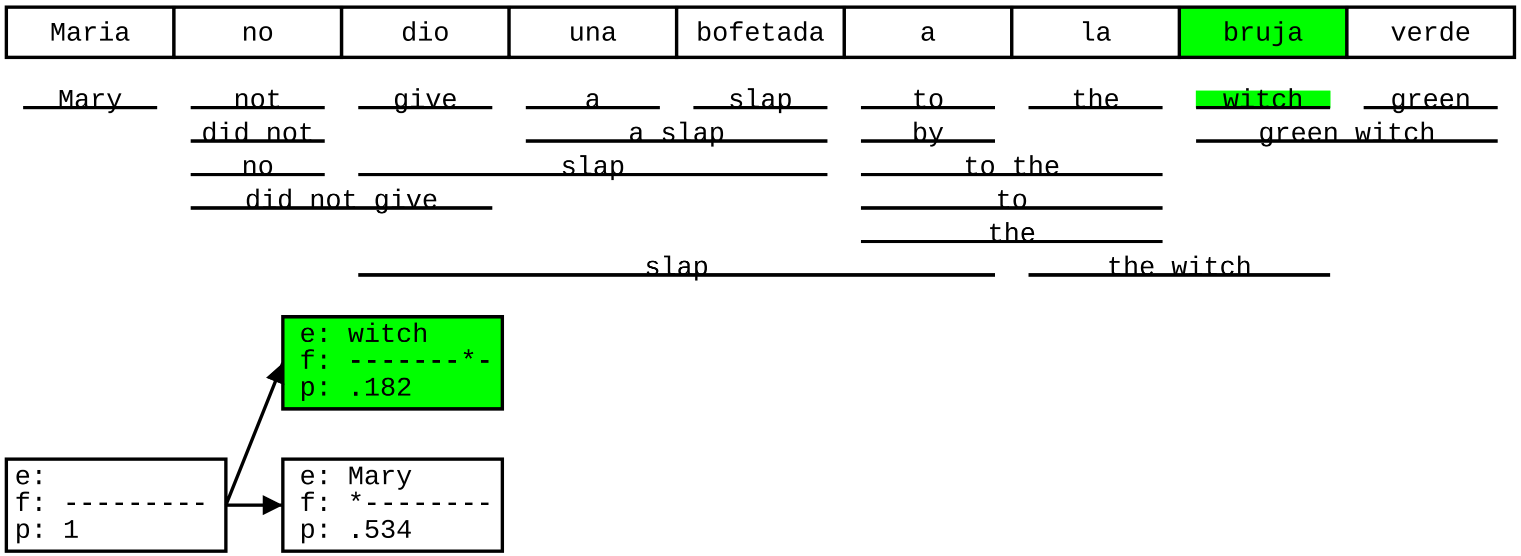


- Pick *translation option*
- Create *hypothesis*
  - e: add English phrase `Mary`
  - f: first foreign word covered
  - p: probability 0.534

A Quick Word on Probabilities

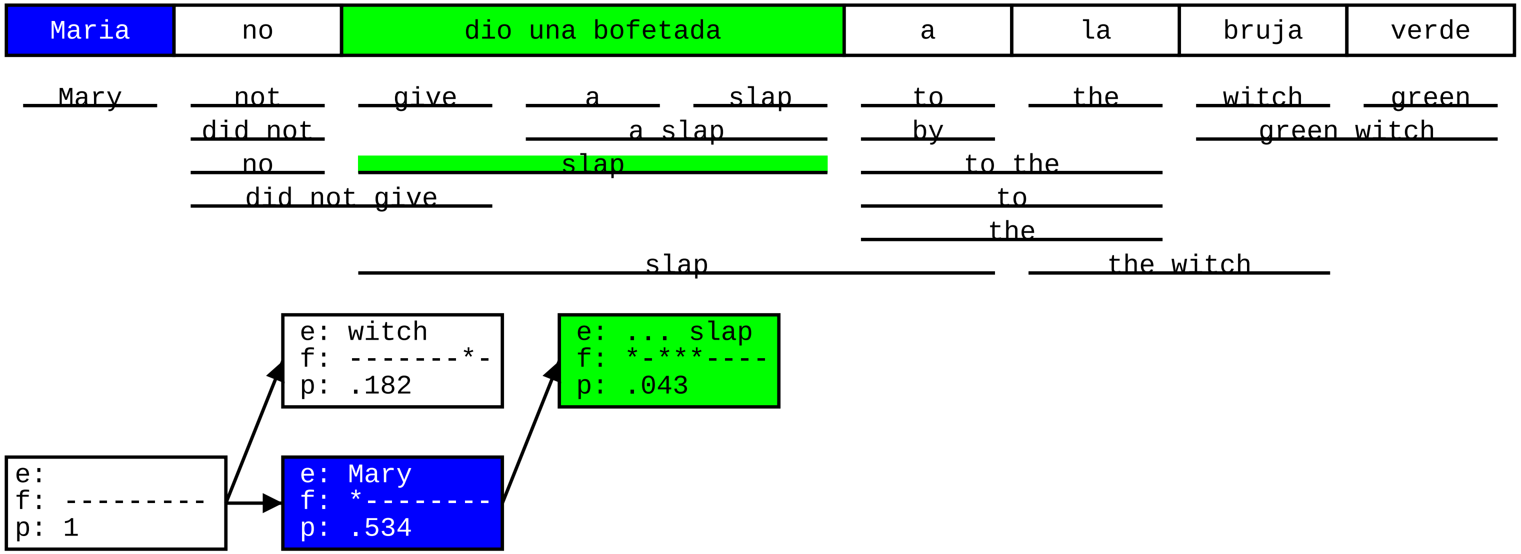
- Not going into detail here, but...
- *Translation Model*
  - phrase translation probability p(Mary|Maria)
  - reordering costs
  - phrase/word count costs
  - ...
- *Language Model*
  - uses trigrams:
  - p(Mary did not) $=$
    p(Mary|START) $\times p$(did|Mary,START) $\times$ p(not|Mary did)

Hypothesis Expansion



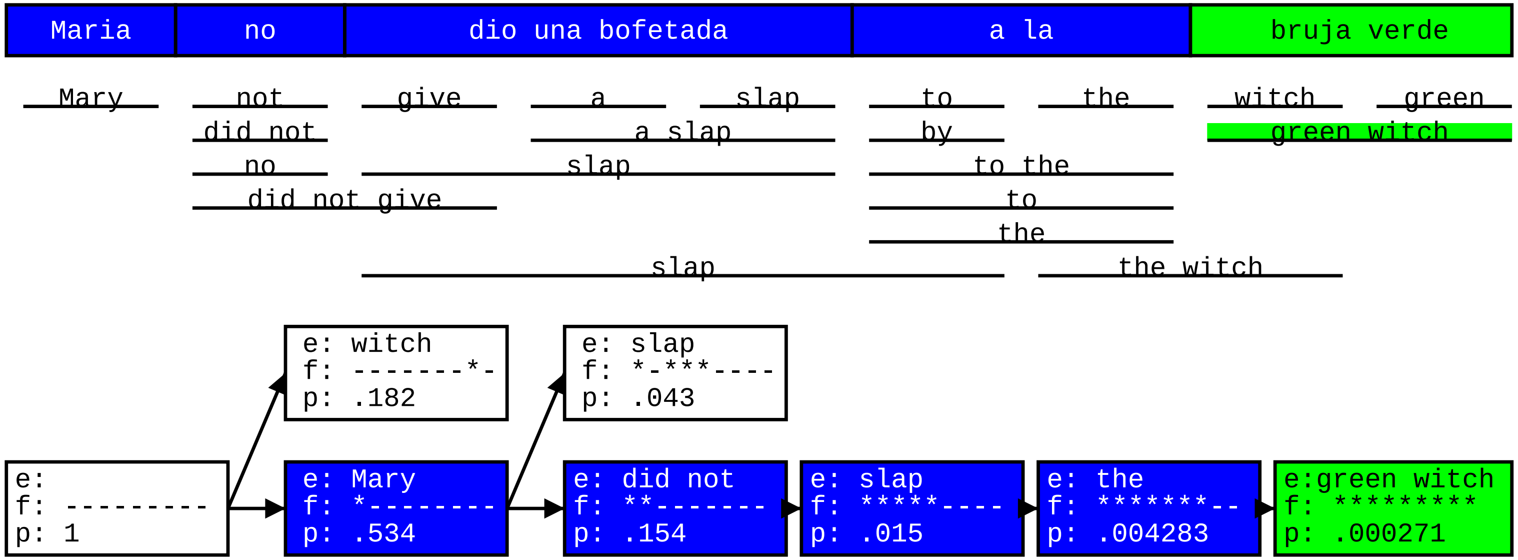
- Add another *hypothesis*

Hypothesis Expansion



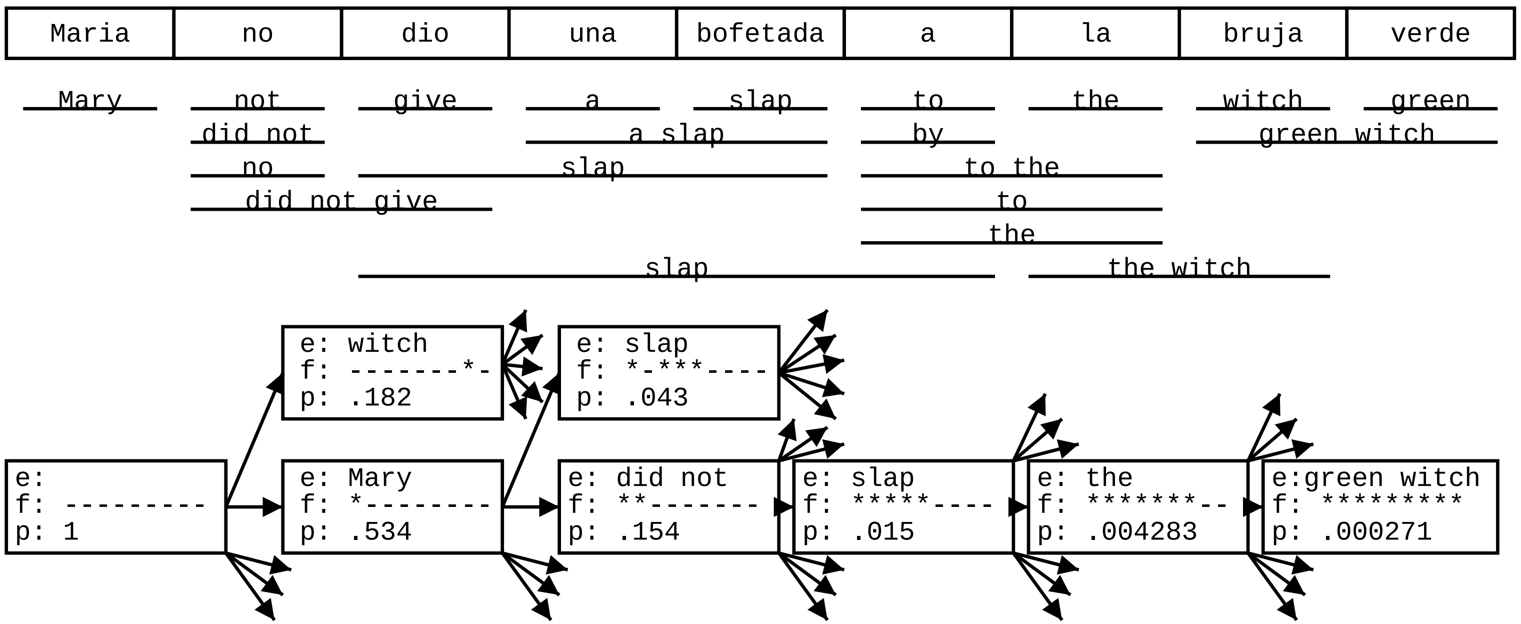
- Further *hypothesis expansion*

Hypothesis Expansion



- ... until all foreign words *covered*
  - find *best hypothesis* that covers all foreign words
  - *backtrack* to read off translation

Hypothesis Expansion



- Adding more hypothesis

⇒ *Explosion* of search space

Explosion of Search Space

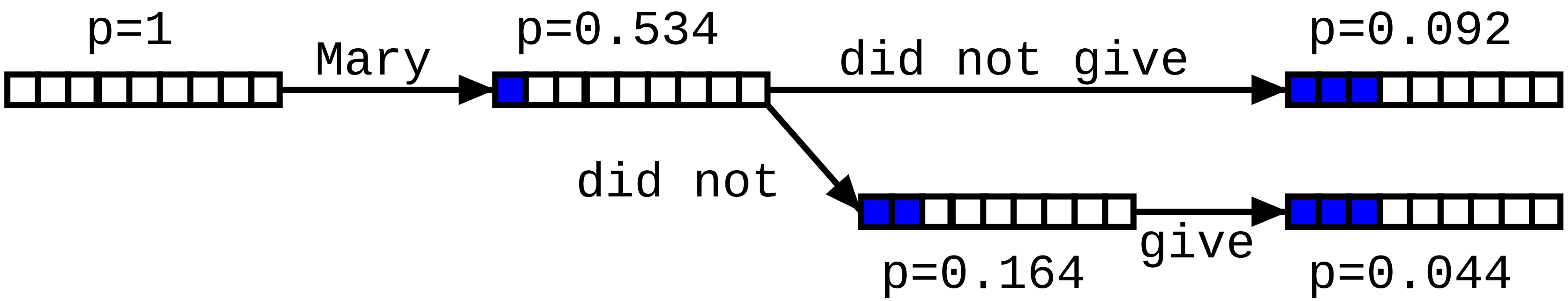
- Number of hypotheses is *exponential* with respect to sentence length

⇒ Decoding is NP-complete [Knight, 1999]

⇒ Need to *reduce search space*

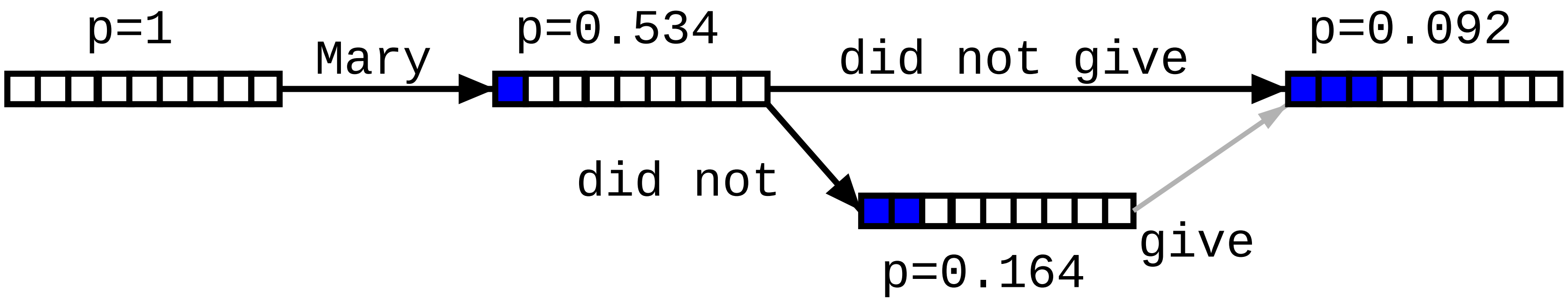
  - risk free: hypothesis **recombination**
  - risky: **histogram/threshold pruning**

Hypothesis Recombination



- Different paths to the *same* partial translation

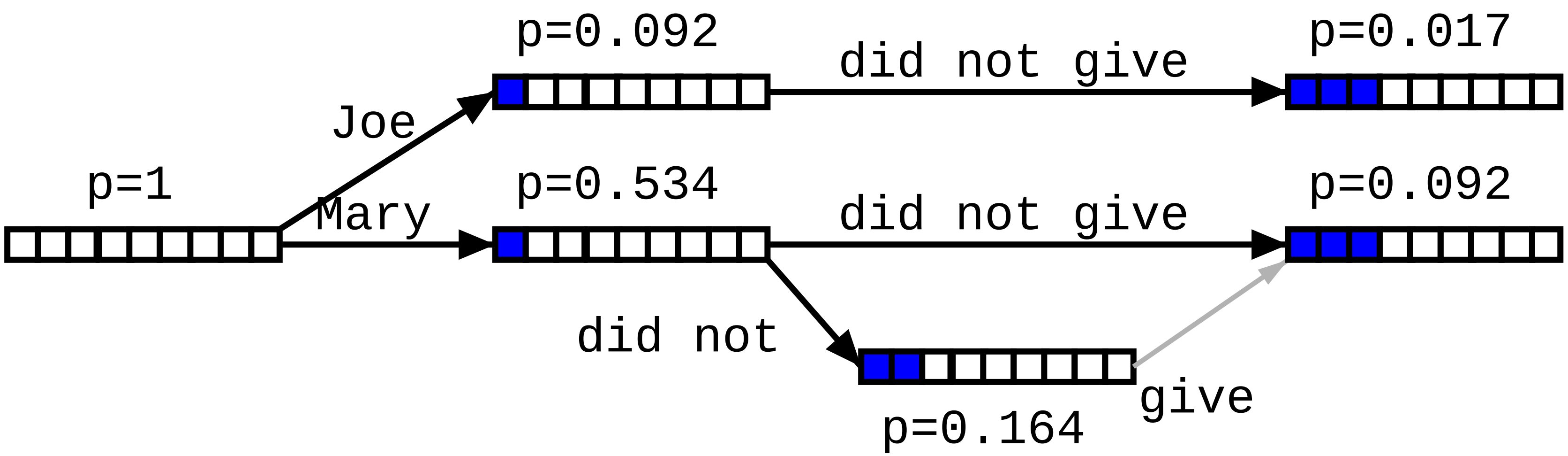
Hypothesis Recombination



- Different paths to the same partial translation

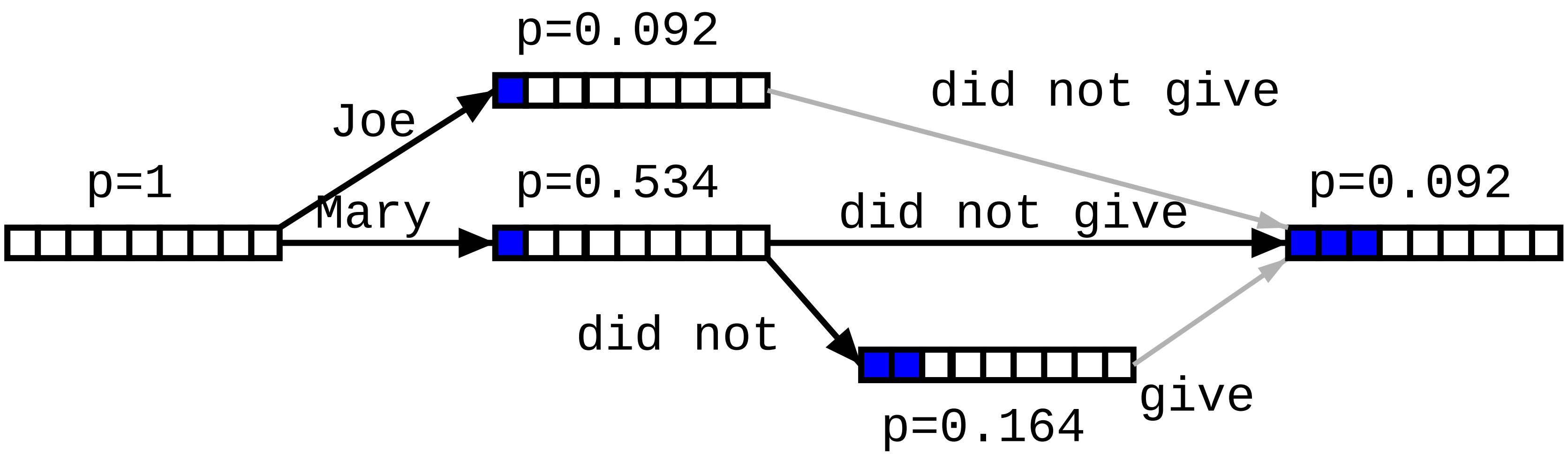
⇒ *Combine paths*
  - *drop weaker* path
  - keep pointer from weaker path (for lattice generation)

Hypothesis Recombination



- Recombined hypotheses do *not* have to *match completely*
- No matter what is added, weaker path can be dropped, if:
  - *last two English words* match (matters for language model)
  - *foreign word coverage* vectors match (effects future path)

Hypothesis Recombination



- Recombined hypotheses do not have to match completely
- No matter what is added, weaker path can be dropped, if:
  - last two English words match (matters for language model)
  - foreign word coverage vectors match (effects future path)

⇒ *Combine paths*

Pruning

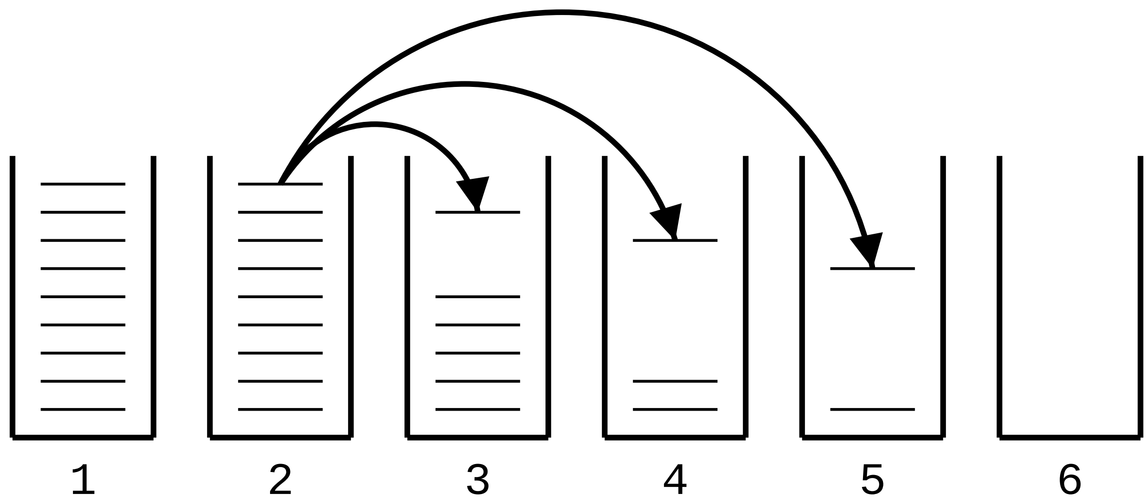
- Hypothesis recombination is *not sufficient*

⇒ Heuristically *discard* weak hypotheses early

- Organize Hypothesis in **stacks**, e.g. by
  - *same* foreign words covered
  - *same number* of foreign words covered
  - *same number* of English words produced

- Compare hypotheses in stacks, discard bad ones
  - **histogram pruning**: keep top n hypotheses in each stack (e.g., n=100)
  - **threshold pruning**: keep hypotheses that are at most α times the cost of best hypothesis in stack (e.g., $\alpha = 0.001$)

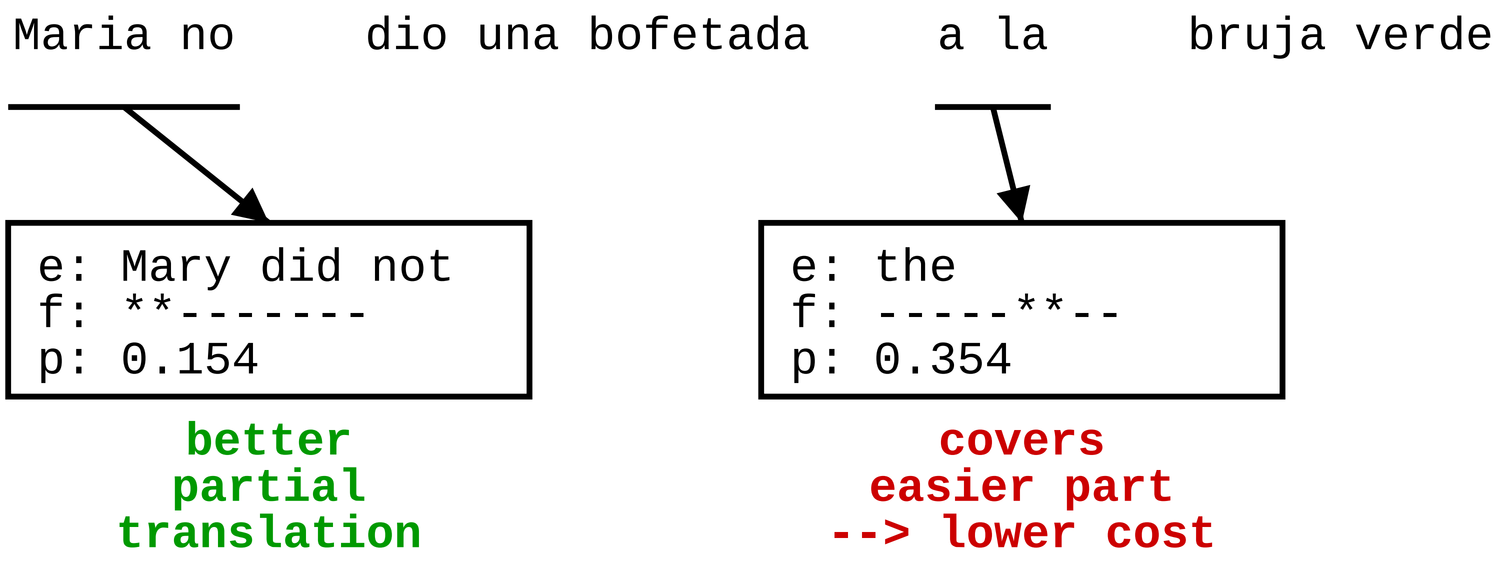
Hypothesis Stacks



- Organization of hypothesis into stacks
  - here: based on *number of foreign words* translated
  - during translation all hypotheses from one stack are expanded
  - expanded Hypotheses are placed into stacks

Comparing Hypotheses

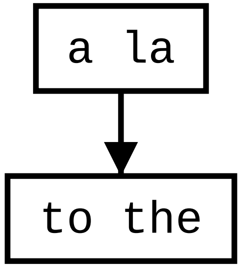
- Comparing hypotheses with *same number of foreign words* covered



- Hypothesis that covers *easy part* of sentence is preferred

⇒ Need to consider **future cost** of uncovered parts

Future Cost Estimation

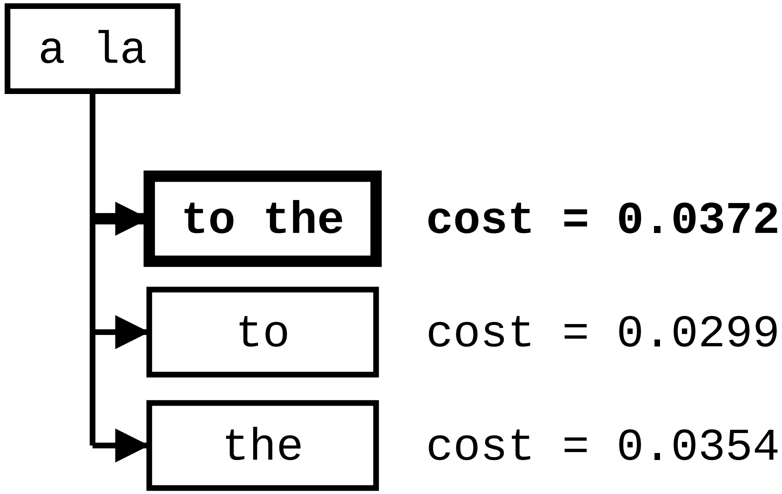


- *Estimate cost* to translate remaining part of input

- Step 1: estimate future cost for each *translation option*
  - look up translation model cost
  - estimate language model cost (no prior context)
  - ignore reordering model cost

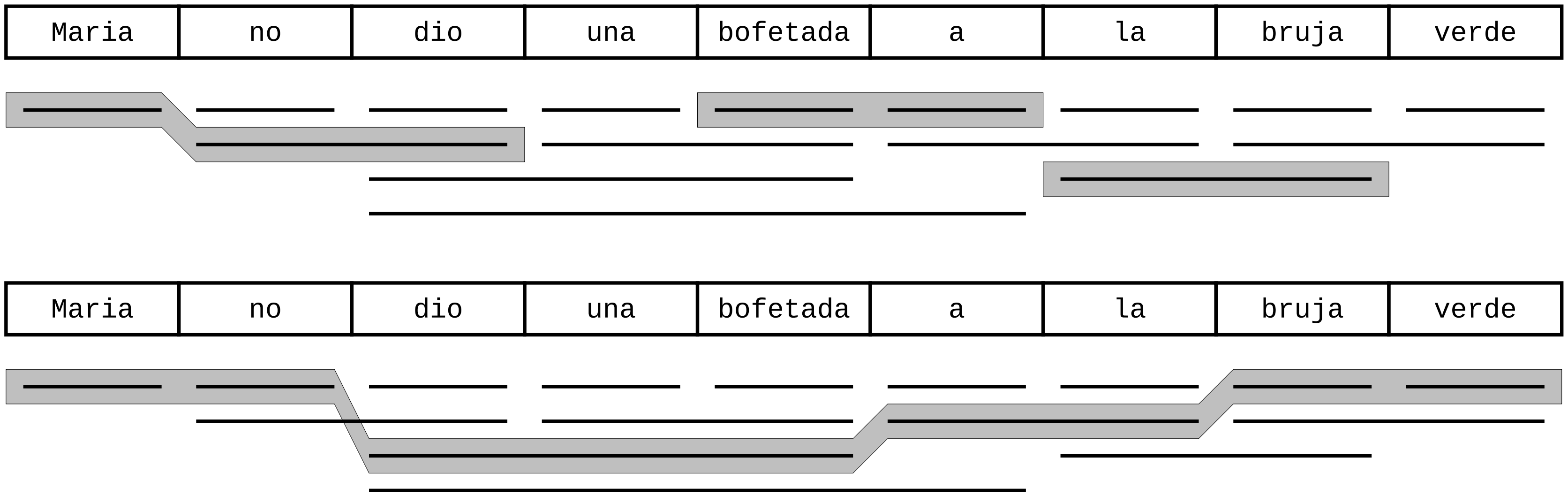
  $\rightarrow$ LM * TM = p(to) * p(the|to) * p(to the|a la)

Future Cost Estimation: Step 2



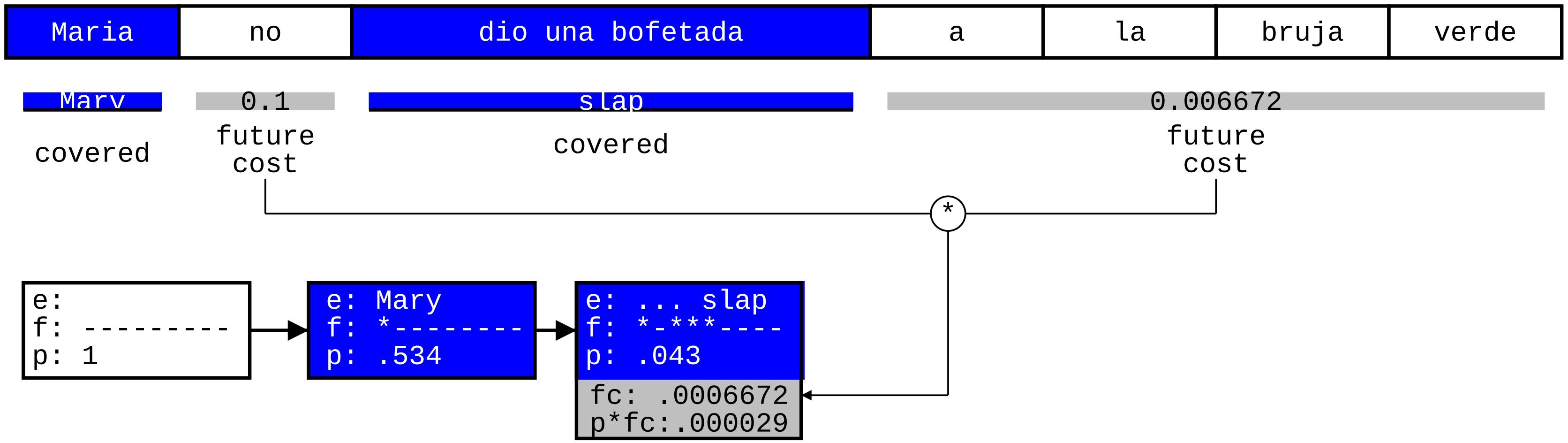
- Step 2: find *cheapest cost* among translation options

Future Cost Estimation: Step 3



- Step 3: find *cheapest future cost path* for each span
  - can be done *efficiently* by dynamic programming
  - future cost for every span can be *pre-computed*

Future Cost Estimation: Application



- Use future cost estimates when *pruning* hypotheses
- For each *uncovered contiguous span*:
  - look up *future costs* for each maximal contiguous uncovered span
  - *add* to actually accumulated cost for translation option for pruning

A* search

- Pruning might drop hypothesis that lead to the best path (**search error**)

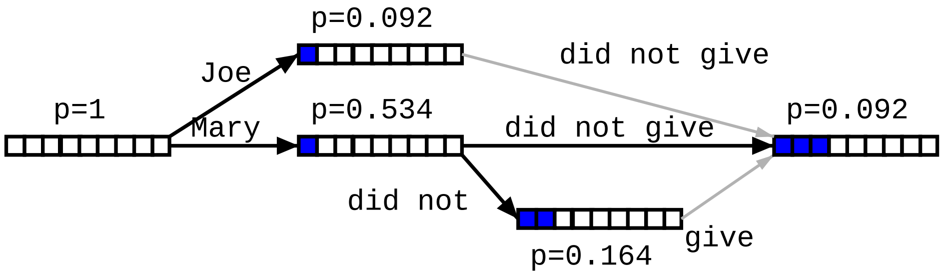
- **A* search**: safe pruning
  - future cost estimates have to be accurate or underestimates
  - **lower bound** for probability is established early by **depth first search**: compute cost for one complete translation
  - if cost-so-far and future cost are worse than *lower bound*, hypothesis can be safely discarded

- Not commonly done, since not aggressive enough

Limits on Reordering

- Reordering may be **limited**
  - **Monotone** Translation: No reordering at all
  - Only phrase movements of at most n words
- Reordering limits *speed* up search (polynomial instead of exponential)
- Current reordering models are weak, so limits *improve* translation quality

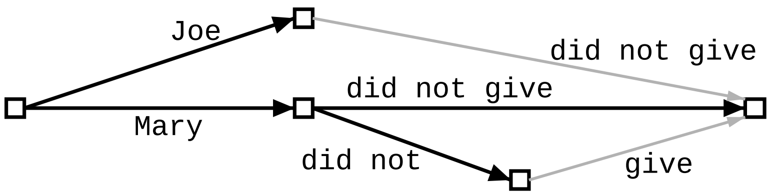
Word Lattice Generation



- **Search graph** can be easily converted into a **word lattice**
  - can be further mined for **n-best lists**

$\longrightarrow$ enables **reranking** approaches

$\longrightarrow$ enables **discriminative training**



Sample N-Best List

- Simple **N-best list**:

```
Translation ||| Reordering LM TM WordPenalty ||| Score
this is a small house ||| 0 -27.0908 -1.83258 -5 ||| -28.9234
this is a little house ||| 0 -28.1791 -1.83258 -5 ||| -30.0117
it is a small house ||| 0 -27.108 -3.21888 -5 ||| -30.3268
it is a little house ||| 0 -28.1963 -3.21888 -5 ||| -31.4152
this is an small house ||| 0 -31.7294 -1.83258 -5 ||| -33.562
it is an small house ||| 0 -32.3094 -3.21888 -5 ||| -35.5283
this is an little house ||| 0 -33.7639 -1.83258 -5 ||| -35.5965
this is a house small ||| -3 -31.4851 -1.83258 -5 ||| -36.3176
this is a house little ||| -3 -31.5689 -1.83258 -5 ||| -36.4015
it is an little house ||| 0 -34.3439 -3.21888 -5 ||| -37.5628
it is a house small ||| -3 -31.5022 -3.21888 -5 ||| -37.7211
this is an house small ||| -3 -32.8999 -1.83258 -5 ||| -37.7325
it is a house little ||| -3 -31.586 -3.21888 -5 ||| -37.8049
this is an house little ||| -3 -32.9837 -1.83258 -5 ||| -37.8163
the house is a little ||| -7 -28.5107 -2.52573 -5 ||| -38.0364
the is a small house ||| 0 -35.6899 -2.52573 -5 ||| -38.2156
is it a little house ||| -4 -30.3603 -3.91202 -5 ||| -38.2723
the house is a small ||| -7 -28.7683 -2.52573 -5 ||| -38.294
it 's a small house ||| 0 -34.8557 -3.91202 -5 ||| -38.7677
this house is a little ||| -7 -28.0443 -3.91202 -5 ||| -38.9563
it 's a little house ||| 0 -35.1446 -3.91202 -5 ||| -39.0566
this house is a small ||| -7 -28.3018 -3.91202 -5 ||| -39.2139
```

Moses: Open Source Toolkit

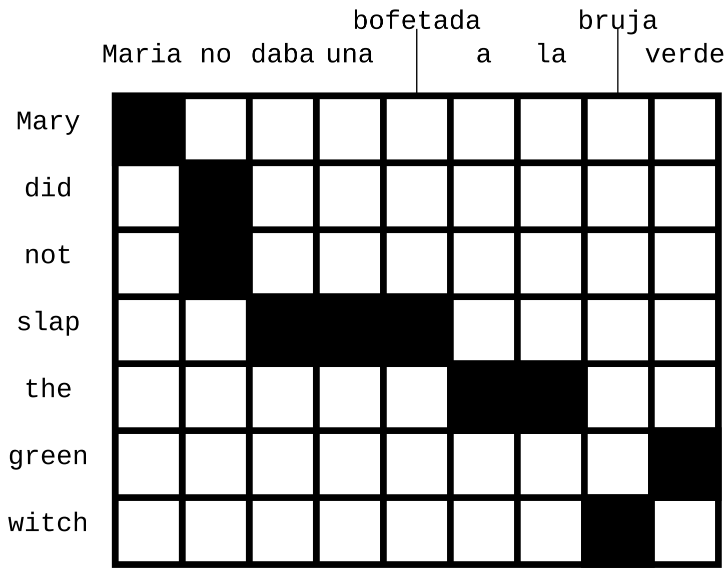
- **Open source** statistical machine translation system (developed from scratch 2006)
  - state-of-the-art *phrase-based* approach
  - novel methods: *factored translation models*, *confusion network decoding*
  - support for *very large models* through *memory-efficient* data structures

- Documentation, source code, binaries **available** at `http://www.statmt.org/moses/`
- Development also **supported by**
  - EC-funded *TC-STAR* project
  - *US* funding agencies DARPA, NSF
  - universities (Edinburgh, Maryland, MIT, ITC-irst, RWTH Aachen, ...)

Phrase-based models

Word alignment

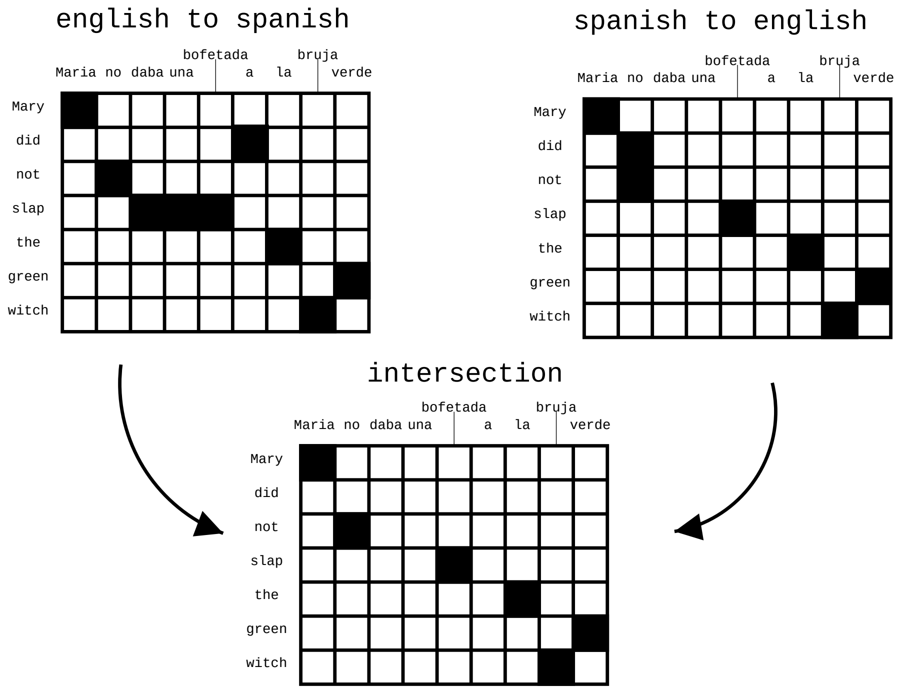
- Notion of **word alignment** valuable
- Shared task at NAACL 2003 and ACL 2005 workshops



Word alignment with IBM models

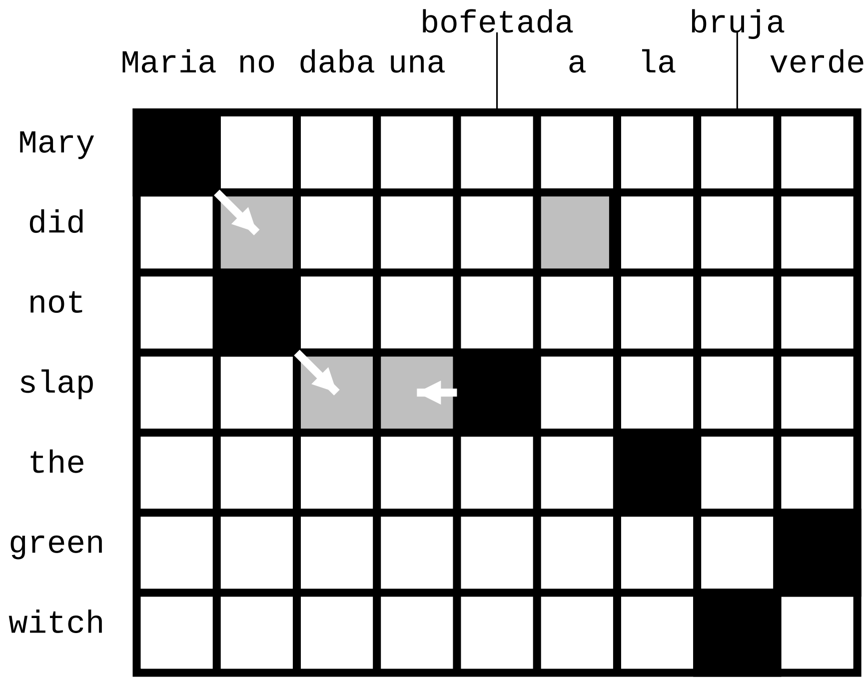
- IBM Models create a *many-to-one* mapping
  - words are aligned using an **alignment function**
  - a function may return the same value for different input (one-to-many mapping)
  - a function can not return multiple values for one input (*no many-to-one* mapping)
- But we need *many-to-many* mappings

Symmetrizing word alignments



- *Intersection* of GIZA++ bidirectional alignments

Symmetrizing word alignments



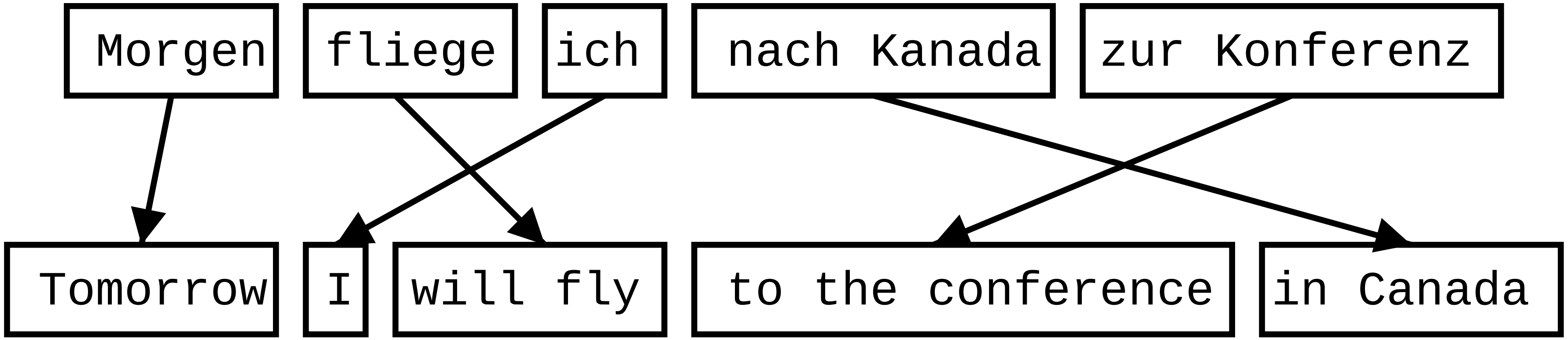
- *Grow* additional alignment points [Och and Ney, CompLing2003]

Growing heuristic

```
GROW-DIAG-FINAL(e2f,f2e):
  neighboring = ((-1,0),(0,-1),(1,0),(0,1),(-1,-1),(-1,1),(1,-1),(1,1))
  alignment = intersect(e2f,f2e);
  GROW-DIAG(); FINAL(e2f); FINAL(f2e);

GROW-DIAG():
  iterate until no new points added
    for english word e = 0 ... en
      for foreign word f = 0 ... fn
        if ( e aligned with f )
          for each neighboring point ( e-new, f-new ):
            if ( ( e-new not aligned and f-new not aligned ) and
                 ( e-new, f-new ) in union( e2f, f2e ) )
              add alignment point ( e-new, f-new )
FINAL(a):
  for english word e-new = 0 ... en
    for foreign word f-new = 0 ... fn
      if ( ( e-new not aligned or f-new not aligned ) and
           ( e-new, f-new ) in alignment a )
        add alignment point ( e-new, f-new )
```

Phrase-based translation



- Foreign input is segmented in phrases
  - any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered

Phrase-based translation model

- Major components of phrase-based model
  - **phrase translation model** $\phi(\mathbf{f}|\mathbf{e})$
  - **reordering model** $\omega^{\text{length}(\mathbf{e})}$
  - **language model** $p_{\text{LM}}(\mathbf{e})$
- Bayes rule

$$\begin{aligned} \text{argmax}_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) &= \text{argmax}_{\mathbf{e}} p(\mathbf{f}|\mathbf{e}) p(\mathbf{e}) \\ &= \text{argmax}_{\mathbf{e}} \phi(\mathbf{f}|\mathbf{e}) p_{\text{LM}}(\mathbf{e}) \omega^{\text{length}(\mathbf{e})} \end{aligned}$$

- Sentence $\mathbf{f}$ is decomposed into I phrases $\bar{f}_1^I = \bar{f}_1, ..., \bar{f}_I$
- Decomposition of $\phi(\mathbf{f}|\mathbf{e})$

$$\phi(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(a_i - b_{i-1})$$

Advantages of phrase-based translation

- *Many-to-many* translation can handle non-compositional phrases
- Use of *local context* in translation
- The more data, the *longer phrases* can be learned

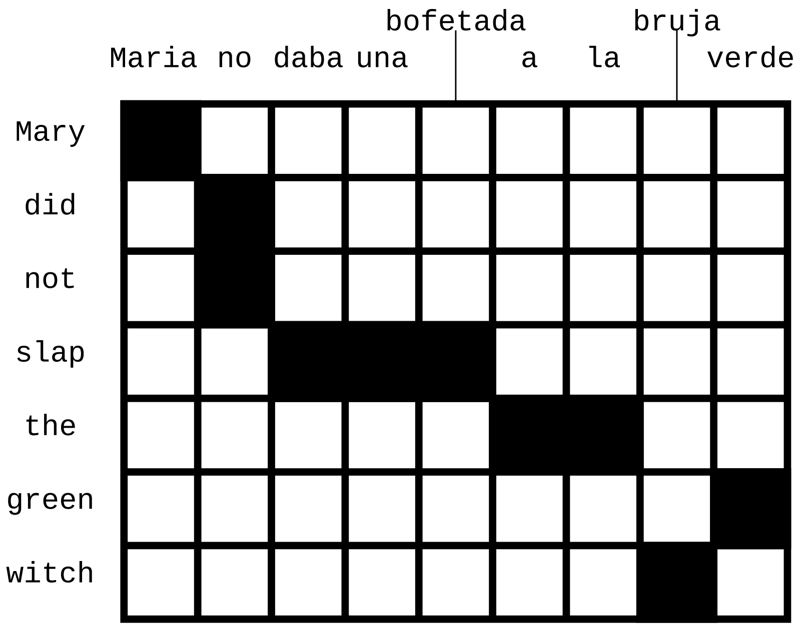
Phrase translation table

- Phrase translations for *den Vorschlag*

English	$\phi(e\|f)$	English	$\phi(e\|f)$
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

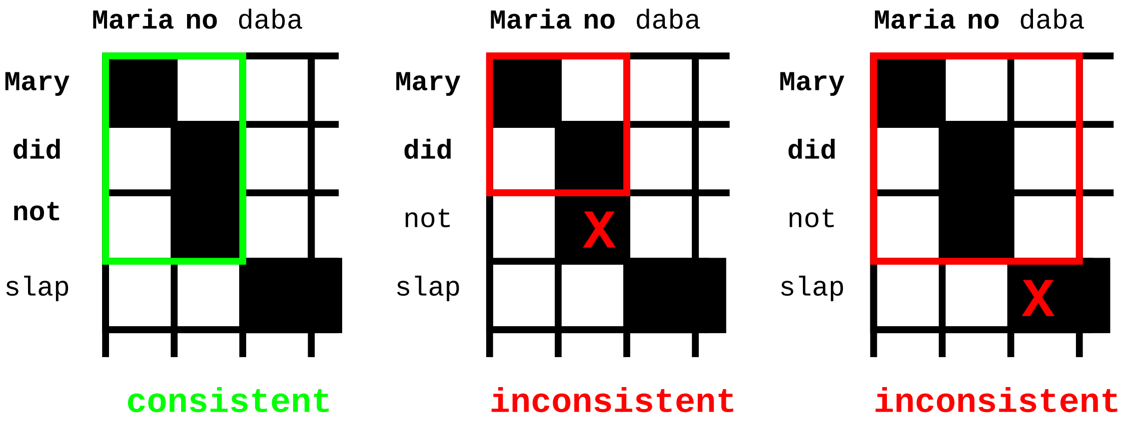
How to learn the phrase translation table?

- Start with the *word alignment*:



- Collect all phrase pairs that are **consistent** with the word alignment

Consistent with word alignment

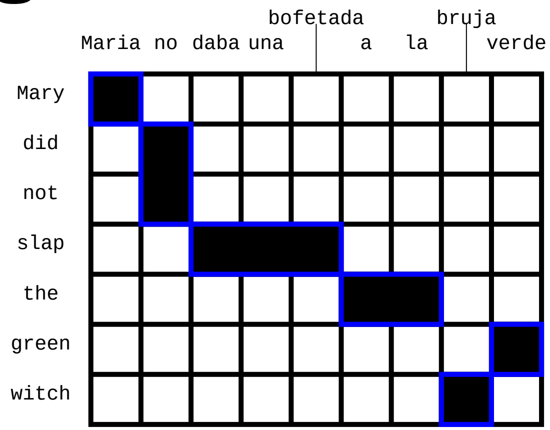


- **Consistent with the word alignment** :=

  phrase alignment has to *contain all alignment points* for all covered words

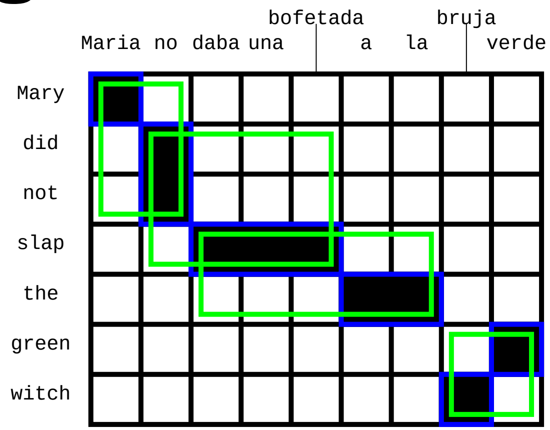
$$
\begin{aligned}
(\bar{e}, \bar{f}) \in BP \Leftrightarrow \quad & \forall e_i \in \bar{e} : (e_i, f_j) \in A \rightarrow f_j \in \bar{f} \\
\text{AND} \quad & \forall f_j \in \bar{f} : (e_i, f_j) \in A \rightarrow e_i \in \bar{e}
\end{aligned}
$$

Word alignment induced phrases



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)

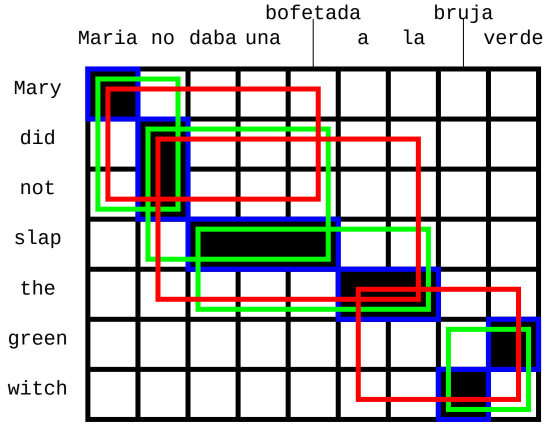
Word alignment induced phrases



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),

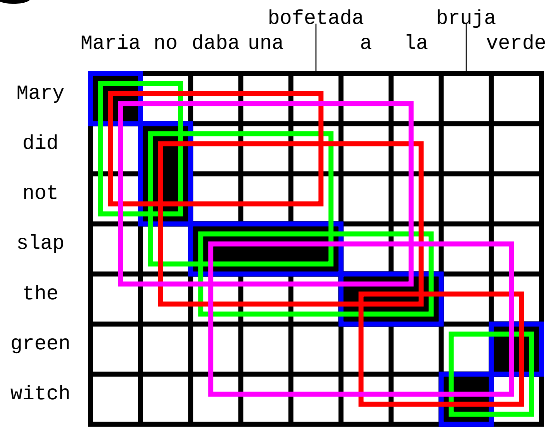
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the), (bruja verde, green witch)

Word alignment induced phrases



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)

Word alignment induced phrases



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),

(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),

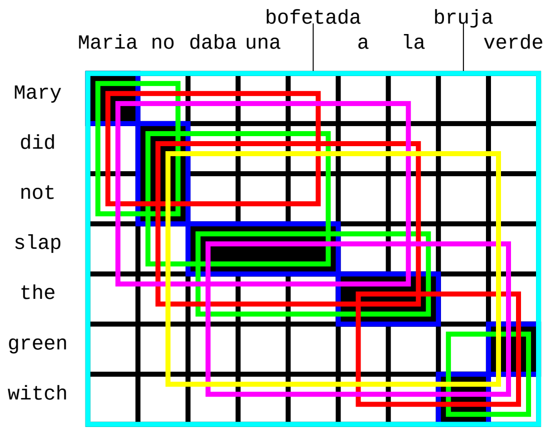
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),

(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),

(Maria no daba una bofetada a la, Mary did not slap the),

(daba una bofetada a la bruja verde, slap the green witch)

Word alignment induced phrases (5)



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde,
slap the green witch), (no daba una bofetada a la bruja verde, did not slap the green witch),
(Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)

Probability distribution of phrase pairs

- We need a **probability distribution** $\phi(\bar{f}|\bar{e})$ over the collected phrase pairs

⇒ Possible *choices*

- *relative frequency* of collected phrases: $\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{f},\bar{e})}{\sum_{\bar{f}} \text{count}(\bar{f},\bar{e})}$
- or, conversely $\phi(\bar{e}|\bar{f})$
- use *lexical translation probabilities*

Reordering

- *Monotone* translation
  - do not allow any reordering

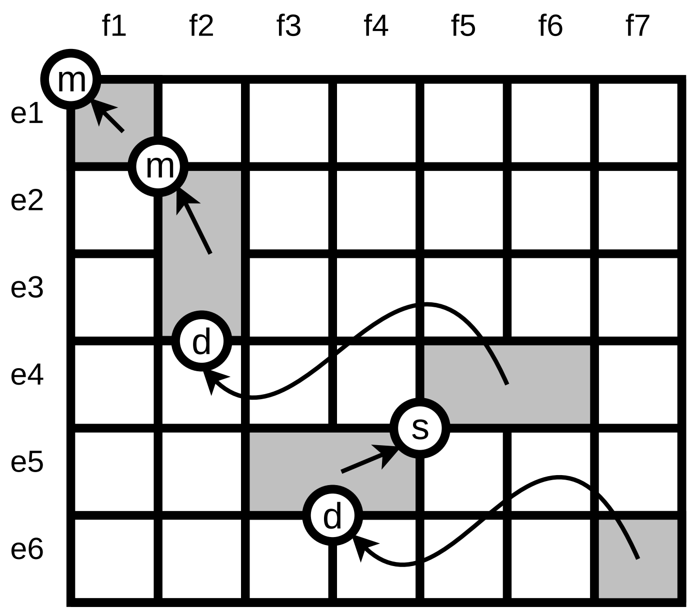
  $\rightarrow$ worse translations

- *Limiting* reordering (to movement over max. number of words) helps

- *Distance-based* reordering cost
  - moving a foreign phrase over n words: cost ω^n

- *Lexicalized* reordering model

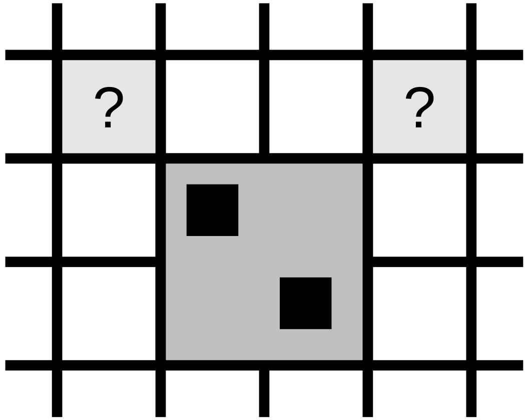
Lexicalized reordering models



[from Koehn et al., 2005, IWSLT]

- Three **orientation** types: **monotone**, **swap**, **discontinuous**
- Probability $p(swap|e, f)$ depends on foreign (and English) *phrase* involved

Learning lexicalized reordering models



[from Koehn et al., 2005, IWSLT]

- Orientation type is *learned during phrase extractions*
- *Alignment point* to the *top left* (monotone) or *top right* (swap)?
- For more, see [Tillmann, 2003] or [Koehn et al., 2005]

Log-linear models

- IBM Models provided mathematical justification for factoring *components* together

$$p_{LM} \times p_{TM} \times p_D$$

- These may be *weighted*

$$p_{LM}^{\lambda_{LM}} \times p_{TM}^{\lambda_{TM}} \times p_D^{\lambda_D}$$

- *Many components* p_i with weights λ_i

$$\Rightarrow \prod_i p_i^{\lambda_i} = exp(\sum_i \lambda_i log(p_i))$$
$$\Rightarrow log \prod_i p_i^{\lambda_i} = \sum_i \lambda_i log(p_i)$$

Knowledge sources

- Many different **knowledge sources** useful
  - language model
  - reordering (distortion) model
  - phrase translation model
  - word translation model
  - word count
  - phrase count
  - drop word feature
  - phrase pair frequency
  - additional language models
  - additional features

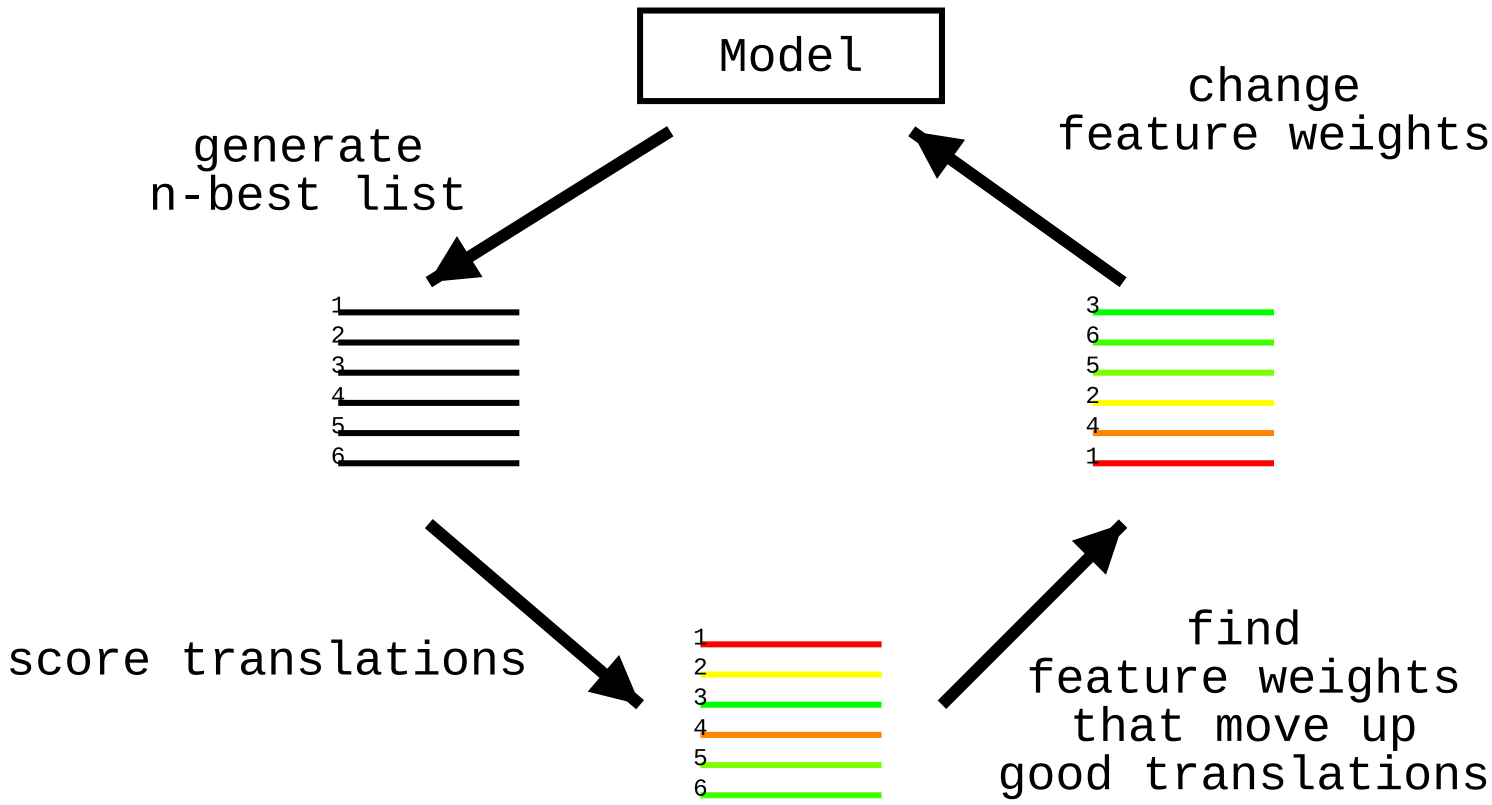
Set feature weights

- Contribution of components p_i determined by weight λ_i

- Methods
  - *manual setting* of weights: try a few, take best
  - *automate* this process

- Learn weights
  - set aside a **development corpus**
  - set the weights, so that **optimal translation performance** on this development corpus is achieved
  - requires *automatic scoring* method (e.g., BLEU)

Learn feature weights



Discriminative vs. generative models

- Generative models
  - translation process is broken down to *steps*
  - each step is modeled by a *probability distribution*
  - each probability distribution is estimated from the data by *maximum likelihood*
- Discriminative models
  - model consist of a number of *features* (e.g. the language model score)
  - each feature has a *weight*, measuring its value for judging a translation as correct
  - feature weights are *optimized on development data*, so that the system output matches correct translations as close as possible

Discriminative training

- Training set (*development set*)
  - different from original training set
  - small (maybe 1000 sentences)
  - must be different from test set

- Current model *translates* this development set
  - *n-best list* of translations (n=100, 10000)
  - translations in n-best list can be *scored*

- Feature weights are *adjusted*

- N-Best list generation and feature weight adjustment repeated for a number of iterations

Learning task

- Task: *find weights*, so that feature vector of the correct translations *ranked first*

rank	translation	LM	TM	WP	SER
1	Mary not give slap witch green .	-17.2	-5.2	-7	1
2	Mary not slap the witch green .	-16.3	-5.7	-7	1
3	Mary not give slap of the green witch .	-18.1	-4.9	-9	1
4	Mary not give of green witch .	-16.5	-5.1	-8	1
5	Mary did not slap the witch green .	-20.1	-4.7	-8	1
6	Mary did not slap green witch .	-15.5	-3.2	-7	1
7	Mary not slap of the witch green .	-19.2	-5.3	-8	1
8	Mary did not give slap of witch green .	-23.2	-5.0	-9	1
9	Mary did not give slap of the green witch .	-21.8	-4.4	-10	1
10	Mary did slap the witch green .	-15.5	-6.9	-7	1
11	**Mary did not slap the green witch .**	**-17.4**	**-5.3**	**-8**	**0**
12	Mary did slap witch green .	-16.9	-6.9	-6	1
13	Mary did slap the green witch .	-14.3	-7.1	-7	1
14	Mary did not slap the of green witch .	-24.2	-5.3	-9	1
15	Mary did not give slap the witch green .	-25.2	-5.5	-9	1

rank translation **feature vector**

Methods to adjust feature weights

- **Maximum entropy** [Och and Ney, ACL2002]
  - match *expectation* of feature values of model and data
- **Minimum error rate** training [Och, ACL2003]
  - try to *rank best translations first* in n-best list
  - can be adapted for various error metrics, even BLEU
- **Ordinal regression** [Shen et al., NAACL2004]
  - *separate* k worst from the k best translations

Syntax-Based
Statistical Machine Translation

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

Kevin Knight

USC/Information Sciences Institute
USC/Computer Science Department

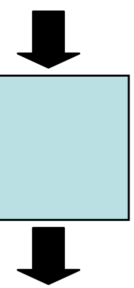
MT Summit, Copenhagen, September, 2007

Topics

- **Quick review of statistical MT essentials**
  - bilingual text
  - phrase substitution models
  - language models
  - decoding
- **Syntax-based statistical MT**
  - syntax-based translation models
  - learning syntactic transformation rules from data
  - decoding
  - tree automata

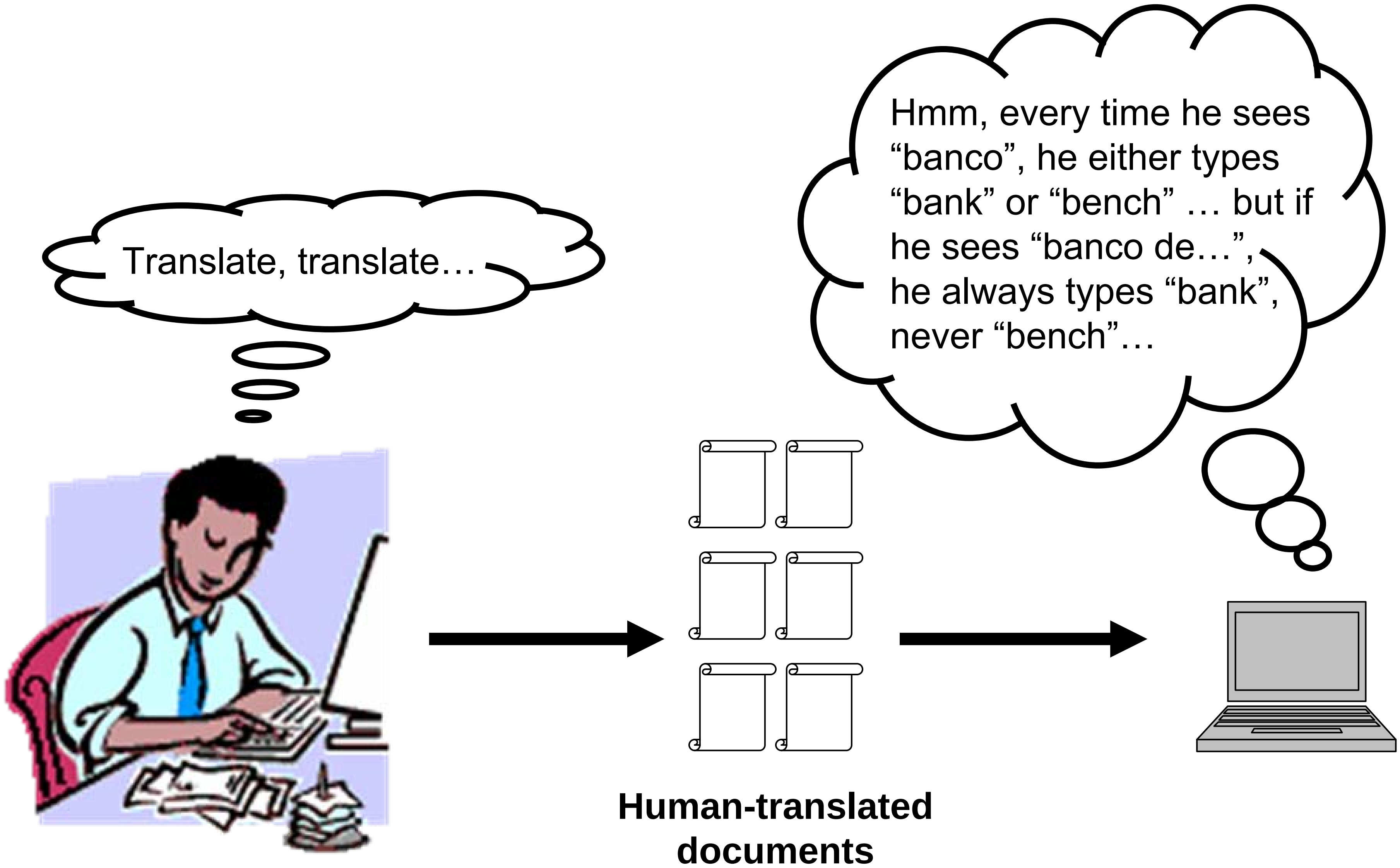
Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。



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

Statistical Machine Translation



Spanish/English corpus

Translate: Clients do not sell pharmaceuticals in Europe.

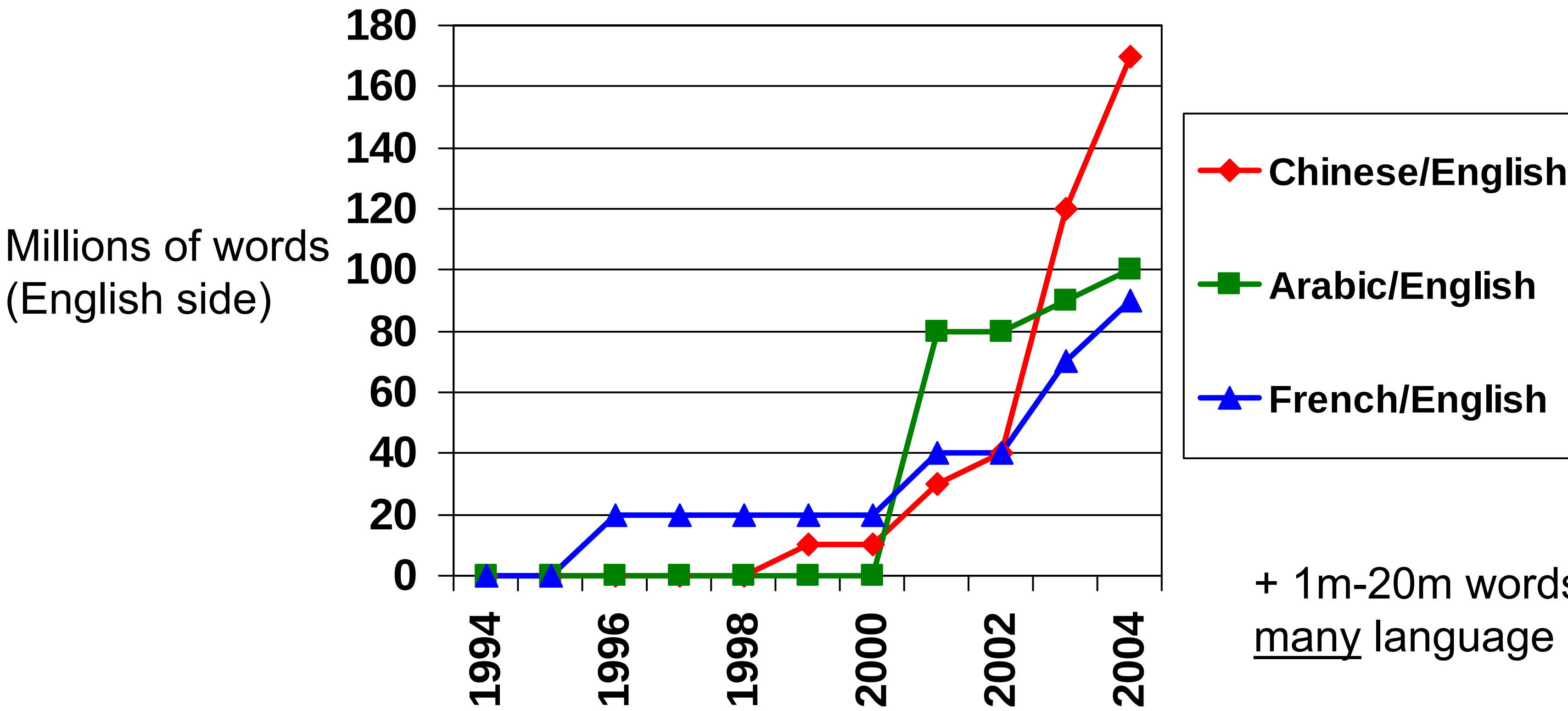
1a. Garcia and associates . 1b. Garcia y asociados .	7a. the clients and the associates are enemies . 7b. los clients y los asociados son enemigos .
2a. Carlos Garcia has three associates . 2b. Carlos Garcia tiene tres asociados .	8a. the company has three groups . 8b. la empresa tiene tres grupos .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

Centauri/Arcturan [Knight 97]

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

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

Ready-to-Use Online Bilingual Data



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

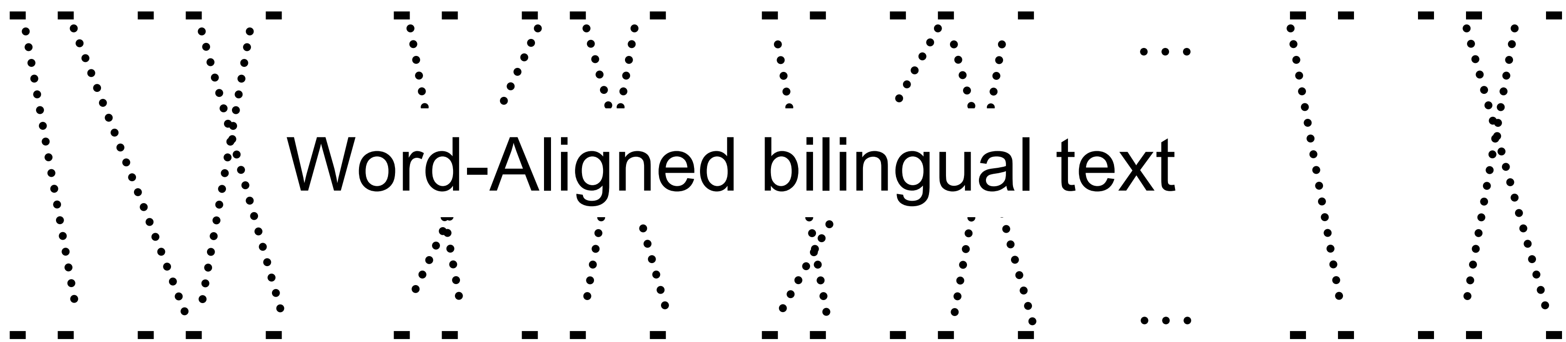
(Data stripped of formatting, in sentence-pair format, available from the Linguistic Data Consortium at UPenn).

Bilingual Text (200m words)

English strings

Word alignments

Chinese strings



Phrase Pair Extraction [Och & Ney, 2004]

Vast Database of Phrase Pairs

Phrase-Based Translation

<table>
<tr><th>这</th><th>7人</th><th>中包括</th><th>来自</th><th>法国</th><th>和</th><th>俄罗斯</th><th>的</th><th>宇航</th><th>员</th><th>.</th></tr>
<tr><td>**the**</td><td>7 people</td><td>including</td><td colspan="2">by some</td><td>**and**</td><td>the russian</td><td>**the**</td><td colspan="2">the astronauts</td><td>,</td></tr>
<tr><td>it</td><td colspan="2">7 people included</td><td colspan="2">by france</td><td>and the</td><td colspan="2">the russian</td><td>international astronautical</td><td colspan="2">of rapporteur .</td></tr>
<tr><td>this</td><td>7 out</td><td>including the</td><td>**from**</td><td>the french</td><td colspan="2">and the russian</td><td colspan="2">the fifth</td><td colspan="2">.</td></tr>
<tr><td>these</td><td>7 among</td><td colspan="2">including from</td><td colspan="2">the french and</td><td>of the russian</td><td>of</td><td>space</td><td>members</td><td>.</td></tr>
<tr><td>that</td><td>7 persons</td><td colspan="2">including from the</td><td>of france</td><td>and to</td><td>russian</td><td>of the</td><td>aerospace</td><td colspan="2">members .</td></tr>
<tr><td></td><td colspan="2">7 include</td><td>from the</td><td colspan="2">of france and</td><td colspan="2">russian</td><td colspan="2">**astronauts**</td><td>. the</td></tr>
<tr><td></td><td colspan="2">7 numbers include</td><td colspan="2">**from france**</td><td colspan="2">and russian</td><td colspan="3">of astronauts who</td><td>. ”</td></tr>
<tr><td></td><td colspan="2">7 populations include</td><td colspan="2">those from france</td><td colspan="3">and russian</td><td colspan="3">astronauts .</td></tr>
<tr><td></td><td colspan="2">7 deportees included</td><td>come from</td><td>**france**</td><td colspan="2">**and russia**</td><td>in</td><td>astronautical</td><td>personnel</td><td>;</td></tr>
<tr><td></td><td>7 philtrum</td><td colspan="2">including those from</td><td colspan="2">**france and**</td><td>**russia**</td><td colspan="2">a space</td><td>**member**</td><td></td></tr>
<tr><td></td><td></td><td colspan="2">including representatives from</td><td colspan="2">france and the</td><td colspan="2">**russia**</td><td colspan="2">astronaut</td><td></td></tr>
<tr><td></td><td></td><td>include</td><td>came from</td><td colspan="3">**france and russia**</td><td colspan="3">by cosmonauts</td><td></td></tr>
<tr><td></td><td></td><td colspan="2">include representatives from</td><td>french</td><td colspan="3">**and russia**</td><td colspan="2">cosmonauts</td><td></td></tr>
<tr><td></td><td></td><td>include</td><td colspan="2">came from france</td><td colspan="3">and russia 's</td><td colspan="3">cosmonauts .</td></tr>
<tr><td></td><td></td><td>**includes**</td><td>coming from</td><td colspan="2">french and</td><td colspan="2">russia 's</td><td colspan="2">cosmonaut</td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td colspan="3">french and russian</td><td>'s</td><td>astronavigation</td><td colspan="2">member .</td></tr>
<tr><td></td><td></td><td></td><td></td><td>french</td><td colspan="2">**and russia**</td><td colspan="3">**astronauts**</td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td colspan="3">and russia 's</td><td></td><td>special rapporteur</td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td>, and</td><td>**russia**</td><td></td><td></td><td>rapporteur</td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td colspan="2">, and russia</td><td></td><td></td><td colspan="2">rapporteur .</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td colspan="3">, and russia</td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td>or</td><td colspan="2">russia 's</td><td></td><td></td><td></td></tr>
</table>

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

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

Phrase-Based Translation

这	7人	中包括	来自	法国	和	俄罗斯	的	宇航	员	.
the	7 people	including	by some		**and**	the russian	**the**	the astronauts		,
it	7 people included		by france		and the	the russian		international astronautical	of rapporteur .	
this	7 out	including the	**from**	the french	and the russian		the fifth		.	
these	7 among	including from		the french and		of the russian	of	space	members	.
that	7 persons	including from the		of france	and to	russian	of the	aerospace	members .	
	7 include		from the	of france and		russian		**astronauts**		. the
	7 numbers include		**from france**		and russian		of astronauts who			. "
	7 populations include		those from france		and russian			astronauts .		
	7 deportees included		come from	**france**	**and russia**		in	astronautical	personnel	;
	7 philtrum	including those from		**france and**		**russia**	a space		**member**	
		including representatives from		france and the		**russia**		astronaut		
		include	came from	**france and russia**			by cosmonauts			
		include representatives from		french	**and russia**			cosmonauts		
		include	came from france		and russia 's			cosmonauts .		
		includes	coming from	french and		russia 's		cosmonaut		
				french and russian			's	astronavigation	member .	
				french	**and russia**		**astronauts**			
					and russia 's				special rapporteur	
					, and	**russia**			rapporteur	
					, and russia				rapporteur .	
					, and russia					
					or	russia 's				

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

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

Phrase-Based Translation

这	7人	中包括	来自	法国	和	俄罗斯	的	宇航	员	.
the	7 people	including	by some		**and**	the russian	**the**	the astronauts		,
it	7 people included		by france		and the	the russian		international astronautical	of rapporteur .	
this	7 out	including the	**from**	the french	and the russian		the fifth		.	
these	7 among	including from		the french and		of the russian	of	space	members	.
that	7 persons	including from the		of france	and to	russian	of the	aerospace	members	
	7 include		from the	of france and		russian		**astronauts**		. the
	7 numbers include		**from france**		and russian		of astronauts who			. "
	7 populations include		those from france		and russian			astronauts .		
	7 deportees included		come from	**france**	**and russia**		in	astronautical	personnel	;
	7 philtrum	including those from		**france and**		**russia**	a space		**member**	
		including representatives from		france and the		**russia**		astronaut		
		include	came from	**france and russia**			by cosmonauts			
		include representatives from		french	**and russia**			cosmonauts		
		include	came from france		and russia 's			cosmonauts .		
		includes	coming from	french and		russia 's		cosmonaut		
				french and russian			's	astronavigation	member .	
				french	**and russia**		**astronauts**			
					and russia 's				special rapporteur	
					, and	**russia**			rapporteur	
					, and russia				rapporteur .	
					, and russia					
					or	russia 's				

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

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

Phrase-Based Translation

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

Scoring: Try to use phrase pairs that have been frequently observed.

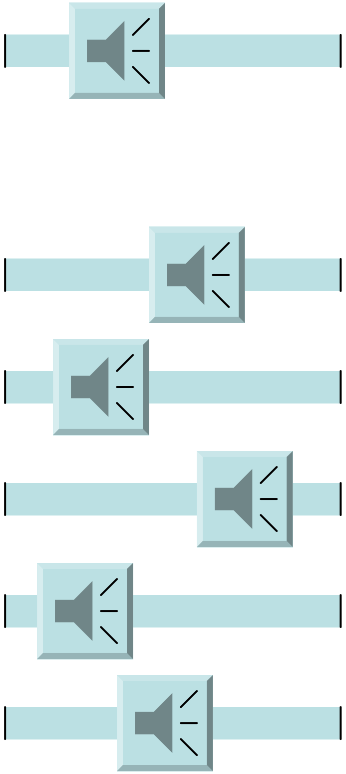
Try to output a sentence with frequent English word sequences.

Components

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

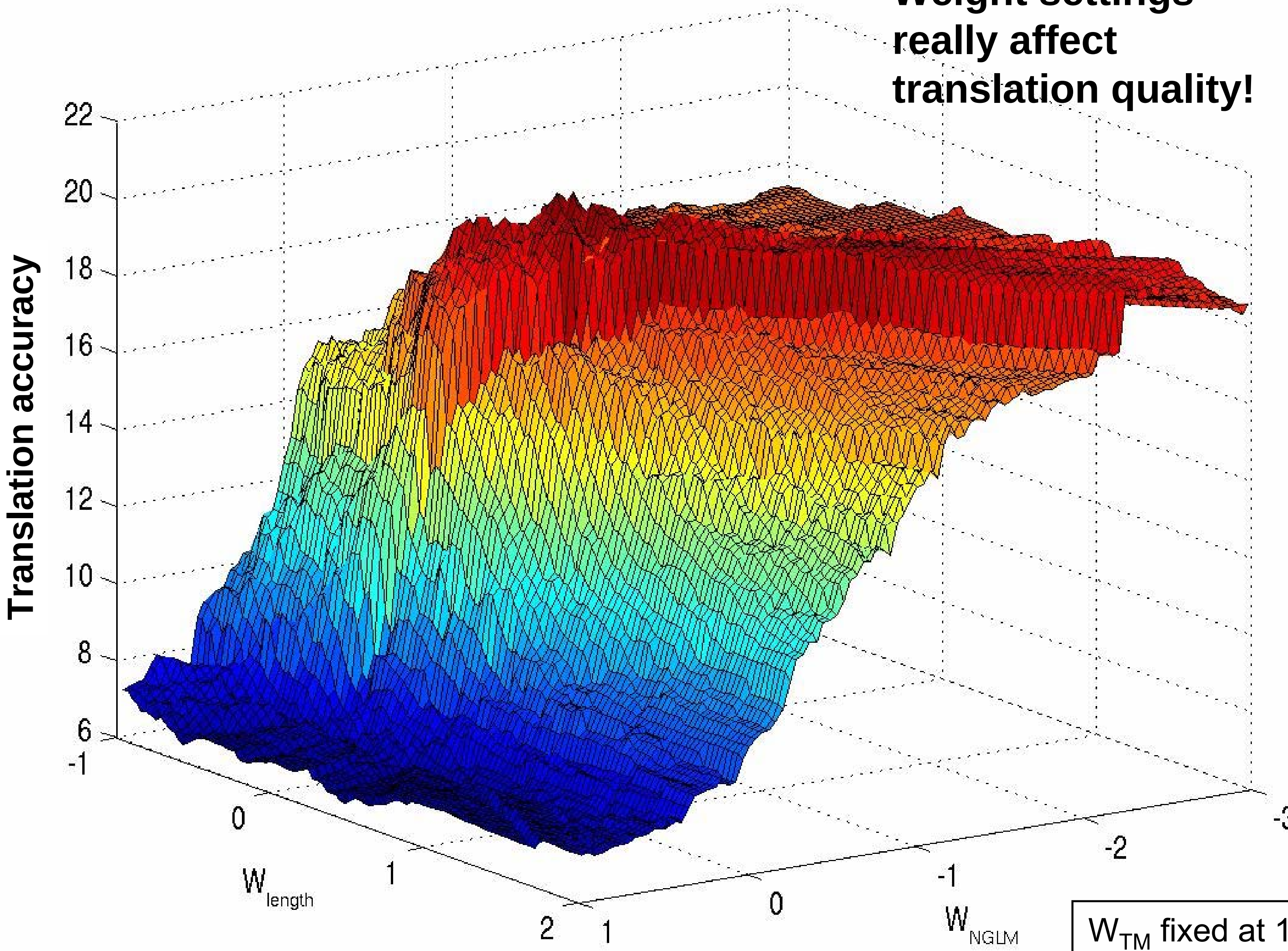
Features and Tuning

- English n-gram language model
- Phrase pairs
  - Corpus probability of phrase pair
  - Bad-phrase spotter
  - Word-drop spotter
  - “Move Me” preference
- English output length



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

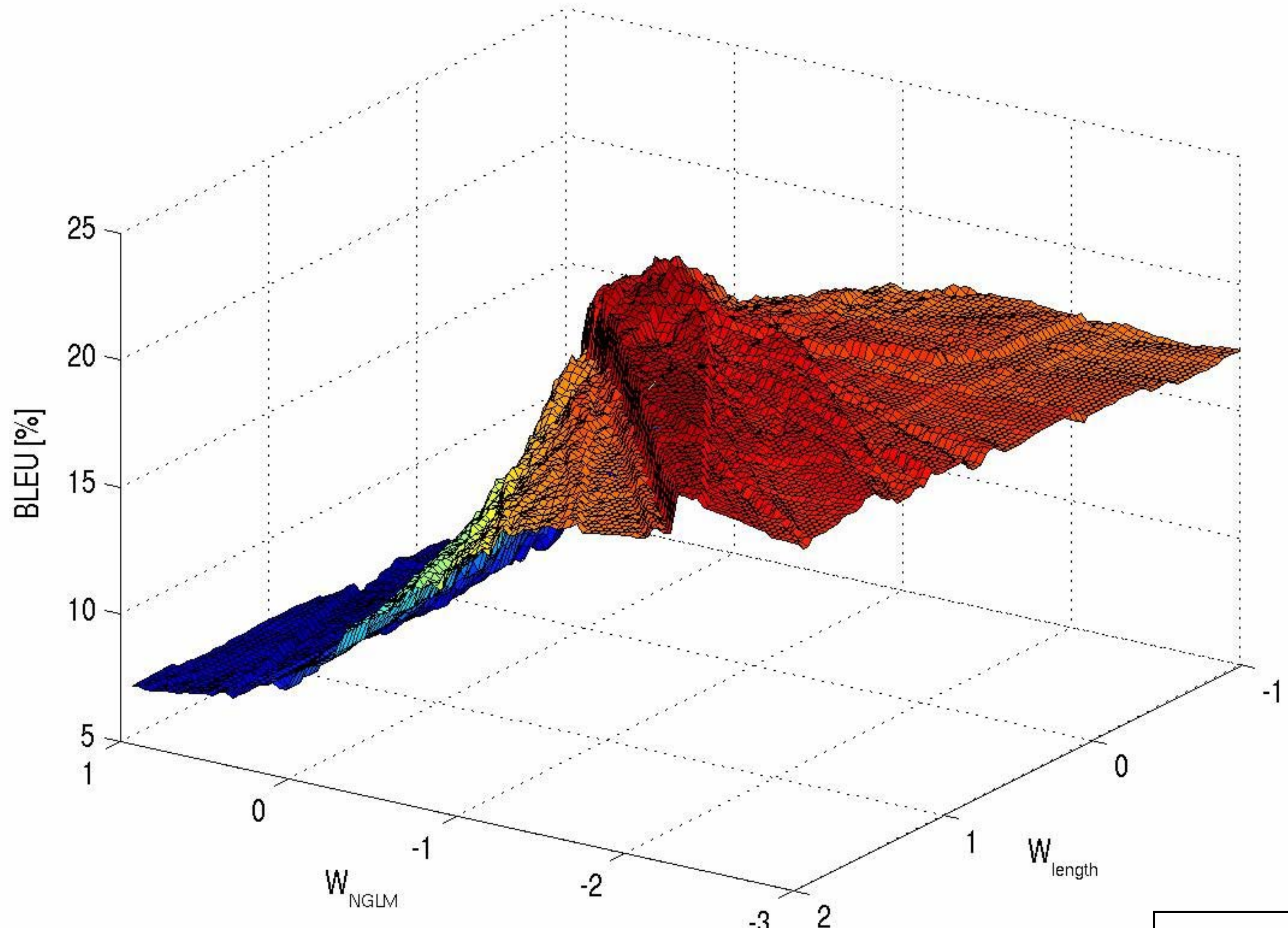
Weight settings really affect translation quality!



plot by Emil Ettelaie

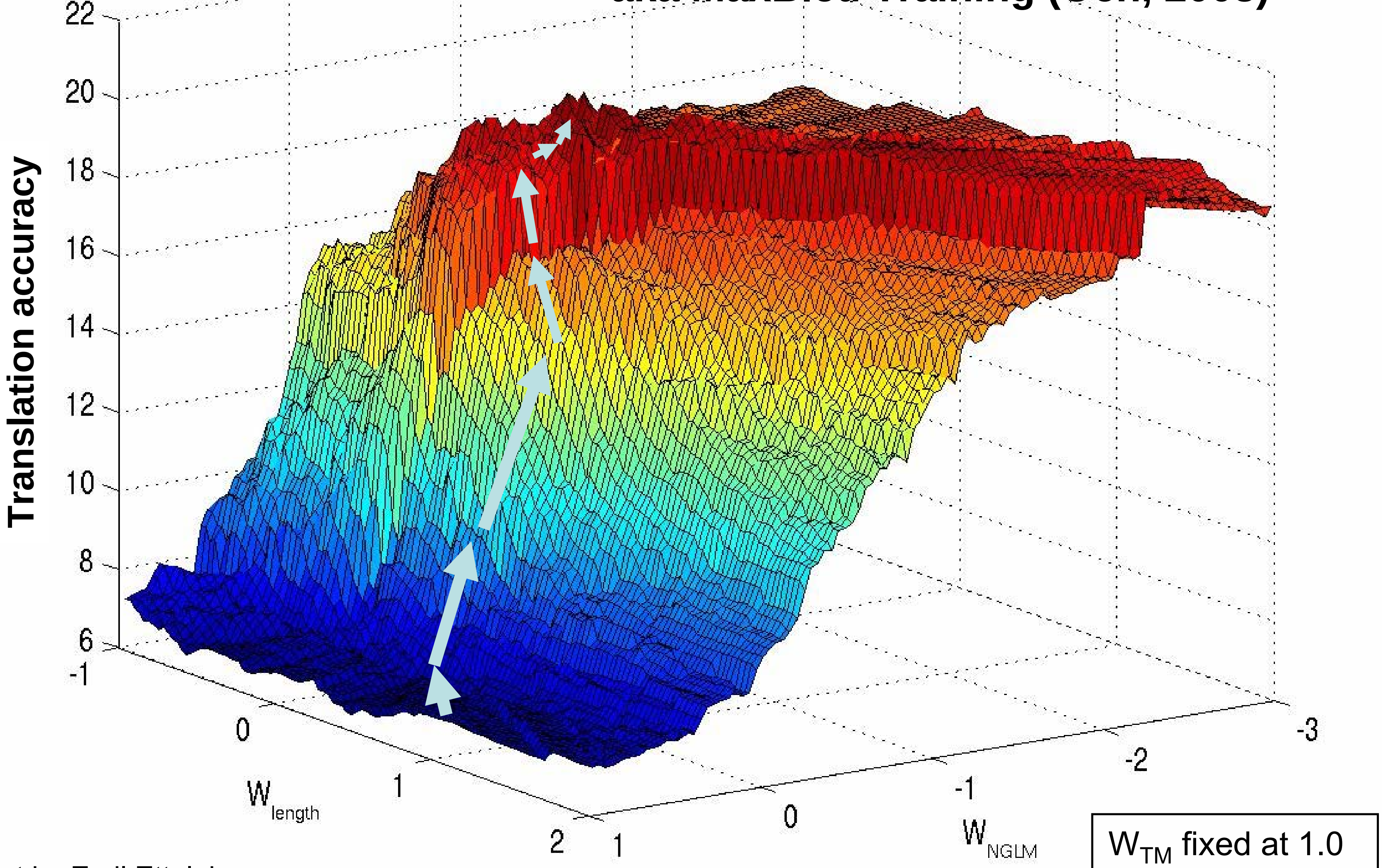
W_{TM} fixed at 1.0

(A View from the Back)



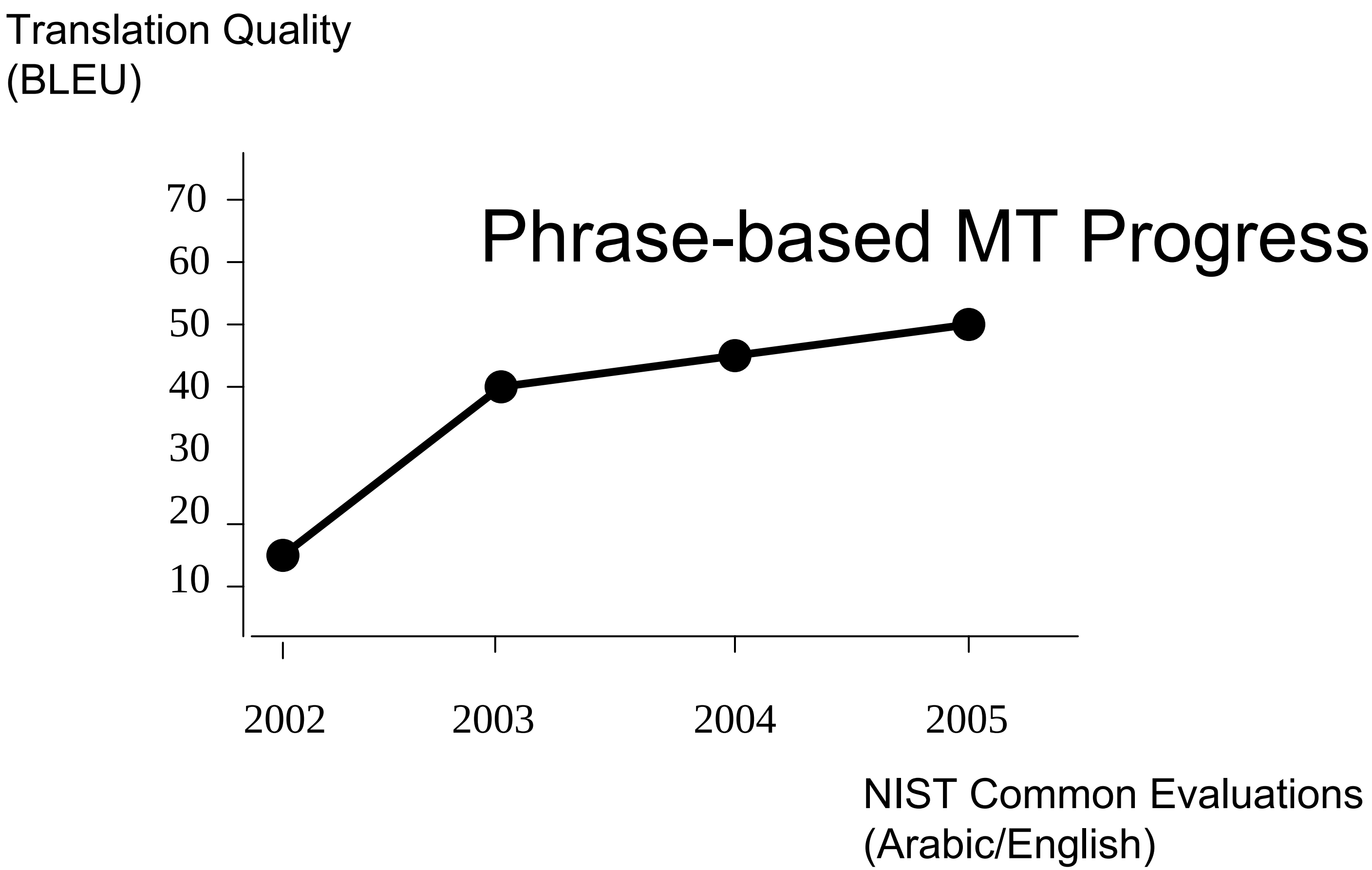
W_{TM} fixed at 1.0

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



plot by Emil Ettelaie

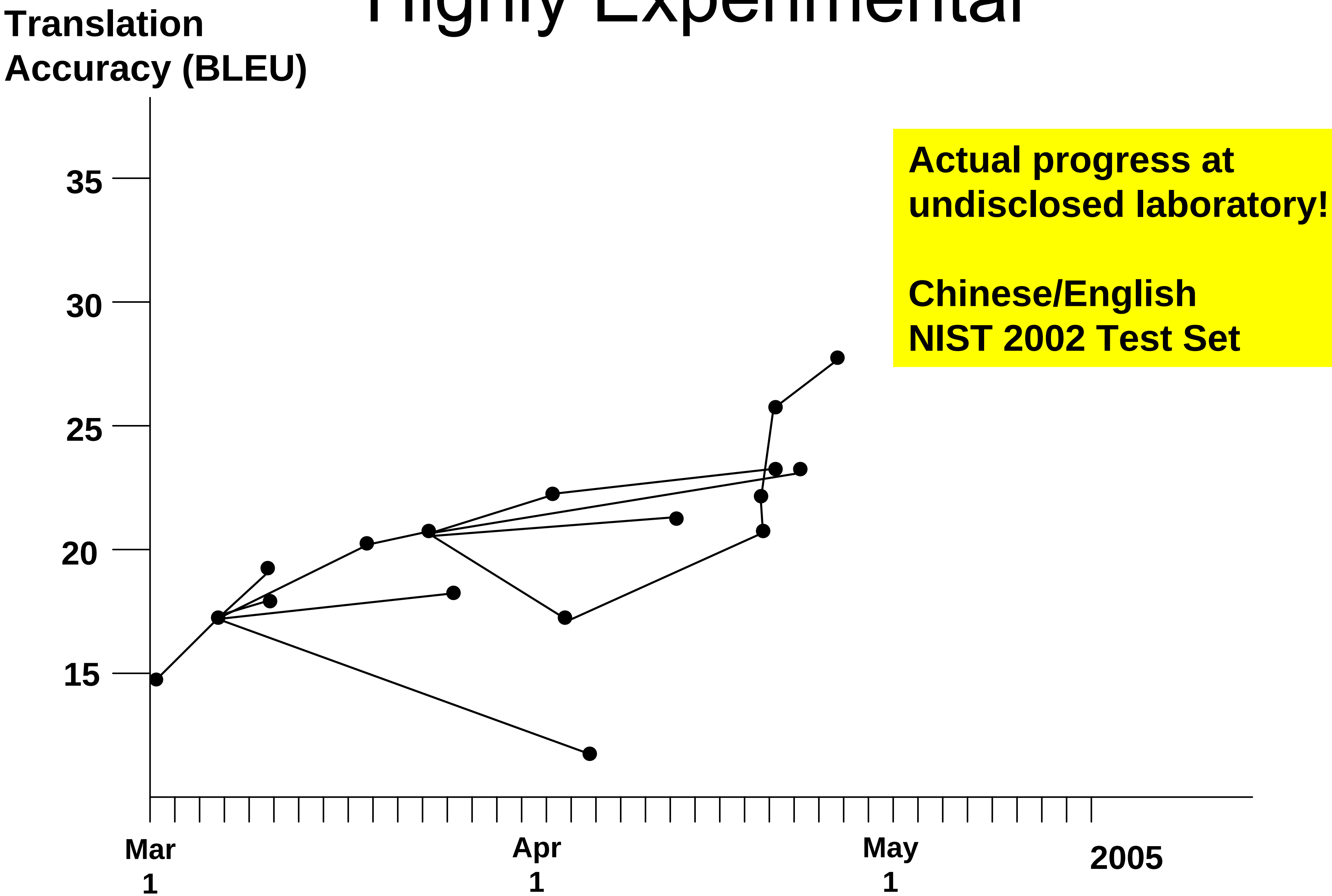
These Ideas Work!



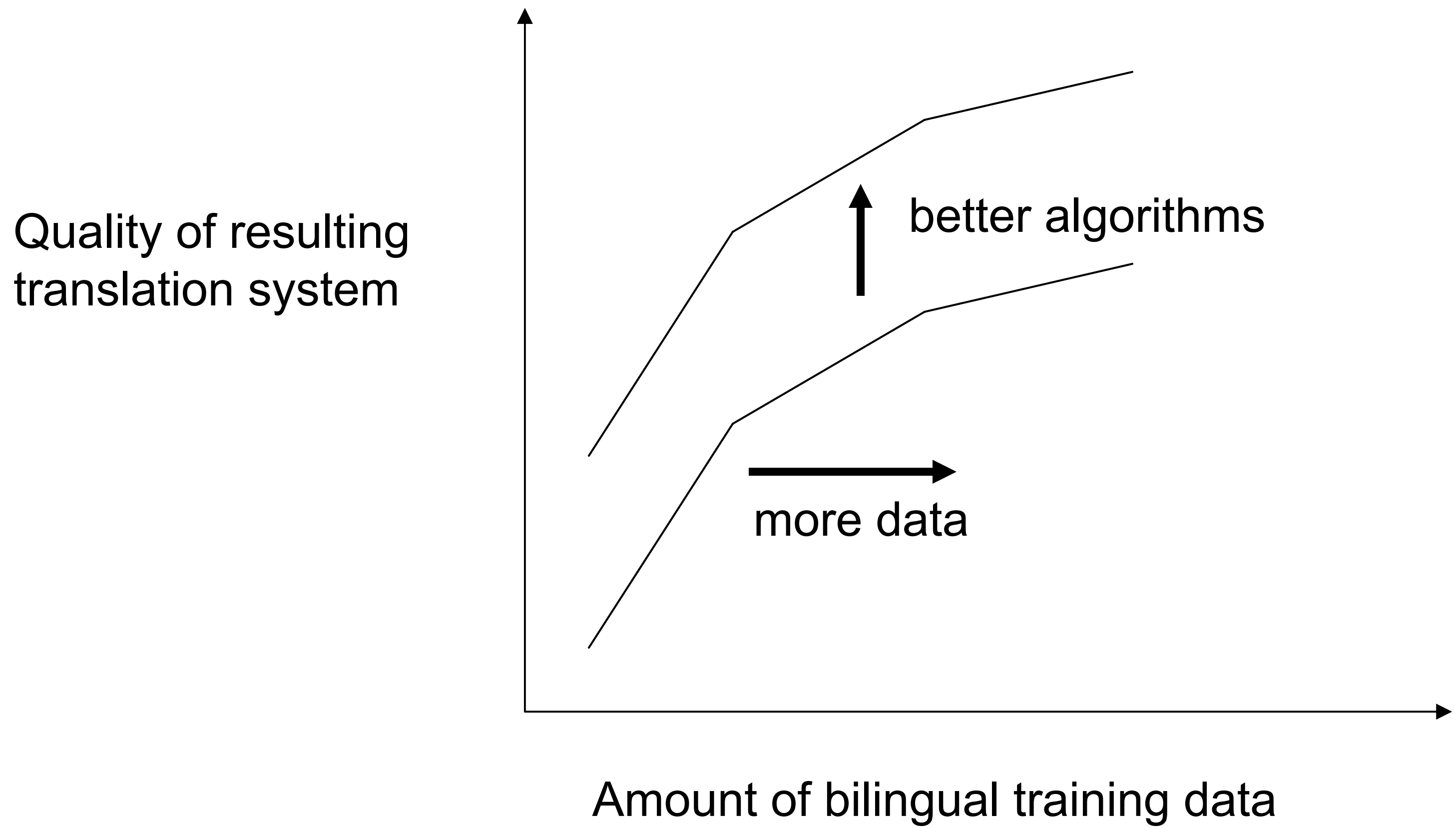
Some Lessons

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

Statistical MT Research is Highly Experimental



Two Ways to Improve Statistical MT Systems



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

- Of course
- But the output is often bad

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

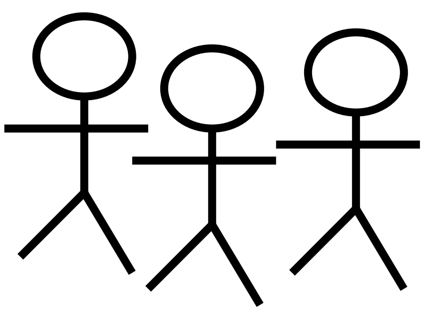
- Our phrase-based story might need some work

Syntax

Maybe we need some grammar?

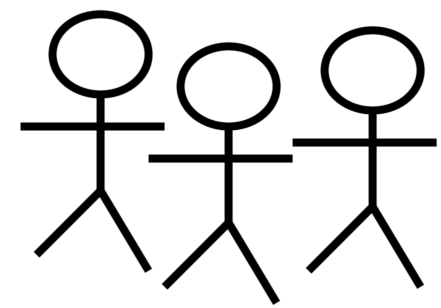
MT Research Landscape

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

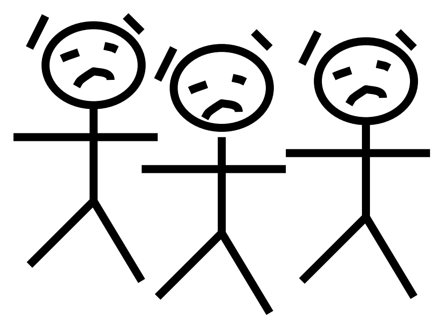


ACL Language Engineers

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

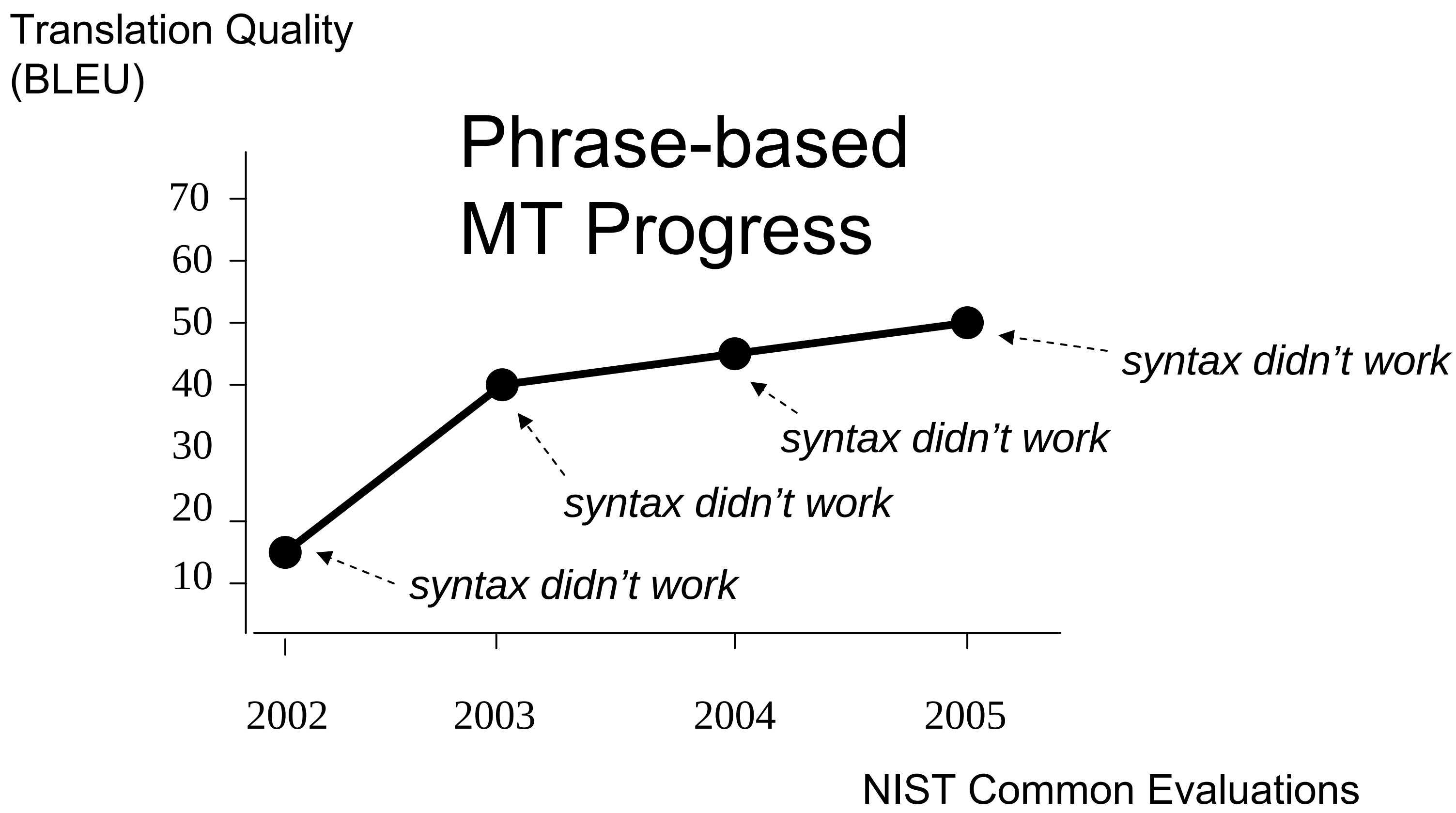


AAAI Fellows

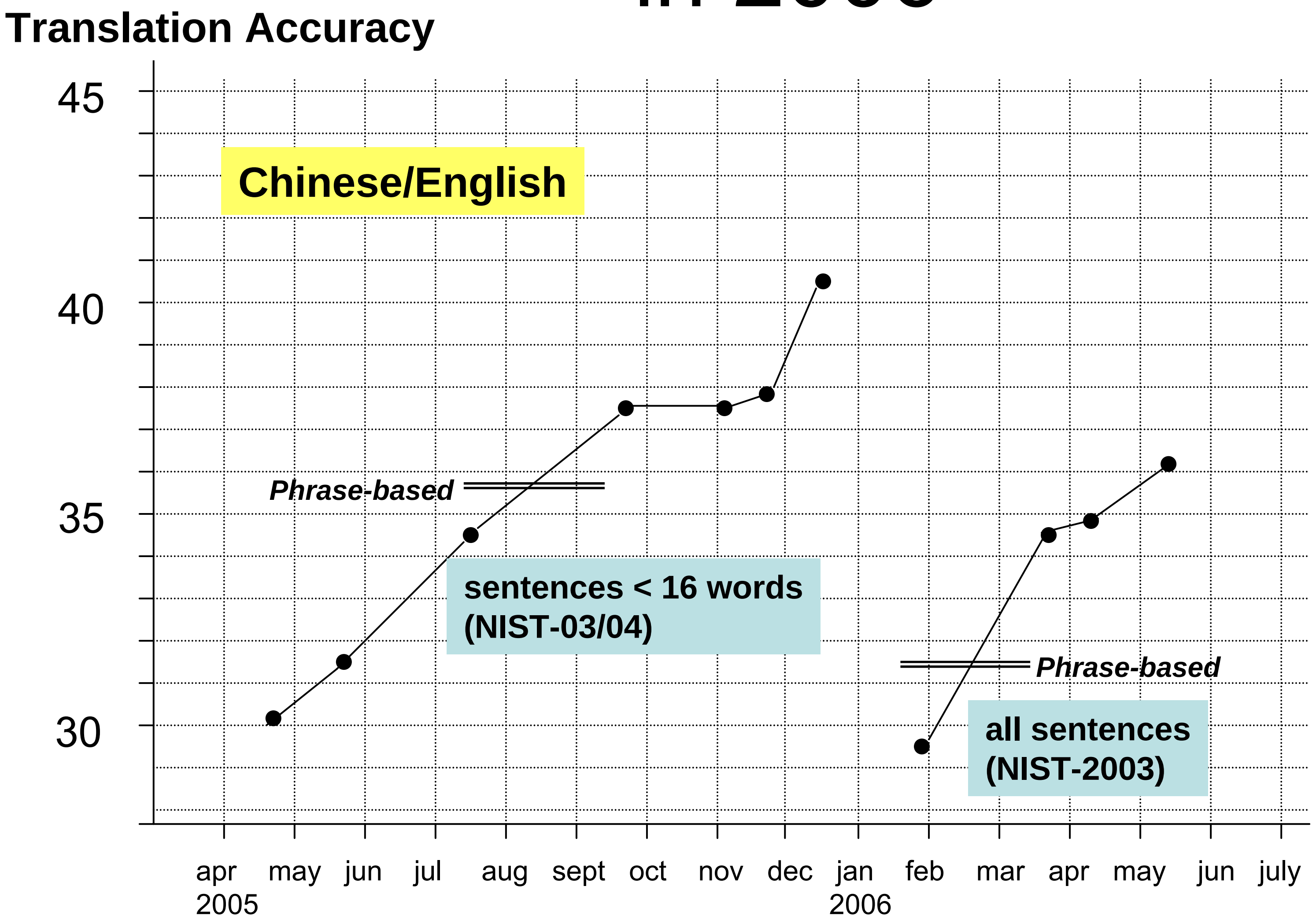


Working on syntax-based approach
to translation (nouns, verbs,
prepositional phrases…)

MT Progress


Translation Quality
(BLEU)
Phrase-based
MT Progress
70
60
50
40
30
20
10
syntax didn't work
syntax didn't work
syntax didn't work
syntax didn't work
2002
2003
2004
2005
NIST Common Evaluations

Syntax Started to Be Helpful in 2006



How to Add Syntax?

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

Phrase-Based Output

枪手 被 警方 击毙 .

Gunman of police killed .

Decoder Hypothesis #1

Phrase-Based Output

枪手 被 警方 击毙 .

Gunman of police attack .

Decoder Hypothesis #7

Phrase-Based Output

枪手 被 警方 击毙 .

Gunman by police killed .

Decoder Hypothesis #12

Phrase-Based Output

枪手 被 警方 击毙 .

Killed gunman by police .

Decoder Hypothesis #134

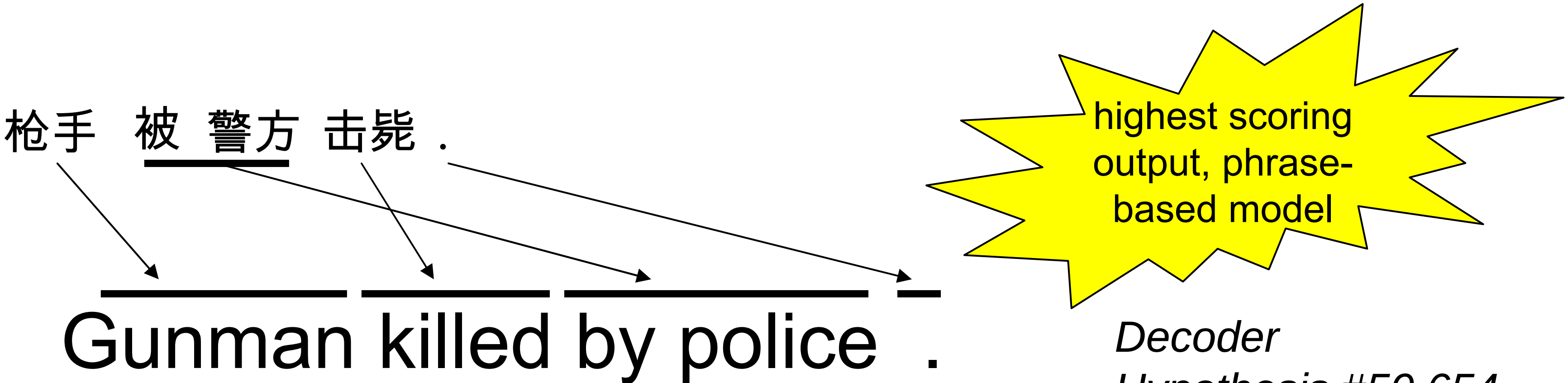
Phrase-Based Output

枪手 被 警方 击毙 .

Gunman killed the police .

Decoder Hypothesis #9,329

Phrase-Based Output



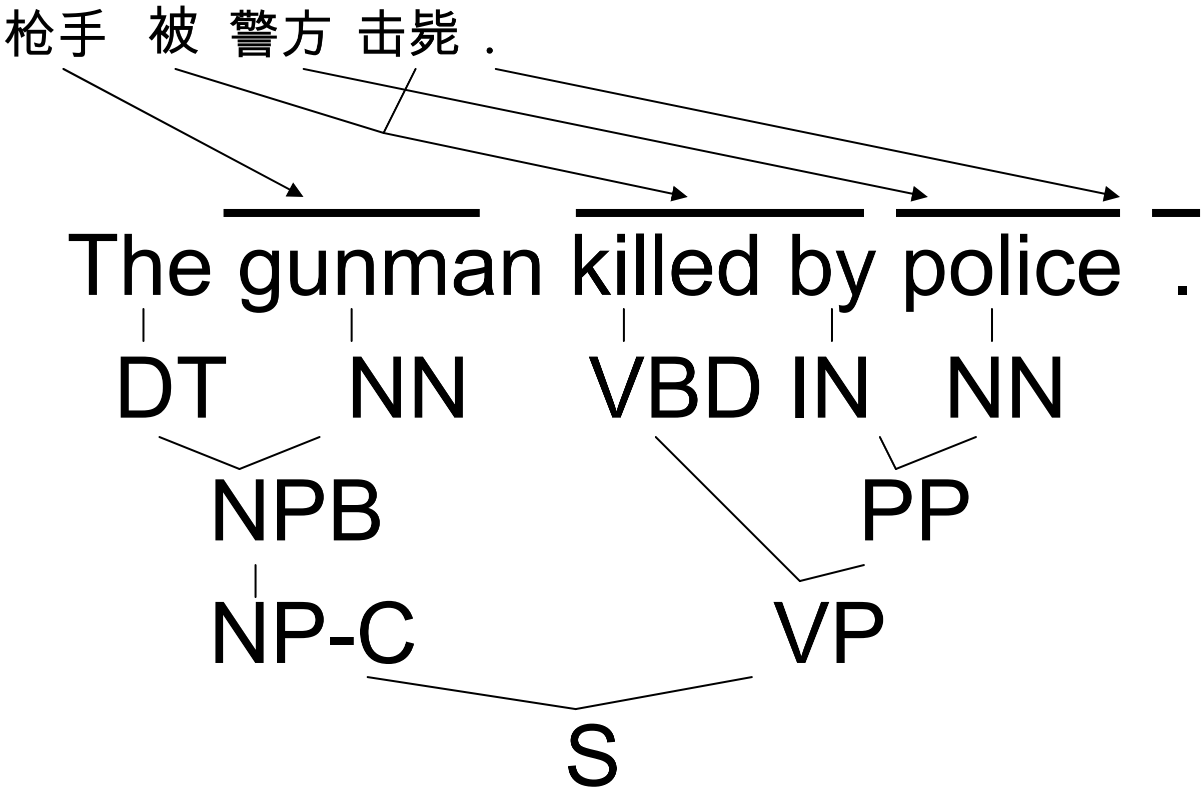
Decoder Hypothesis #50,654

Problematic:

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

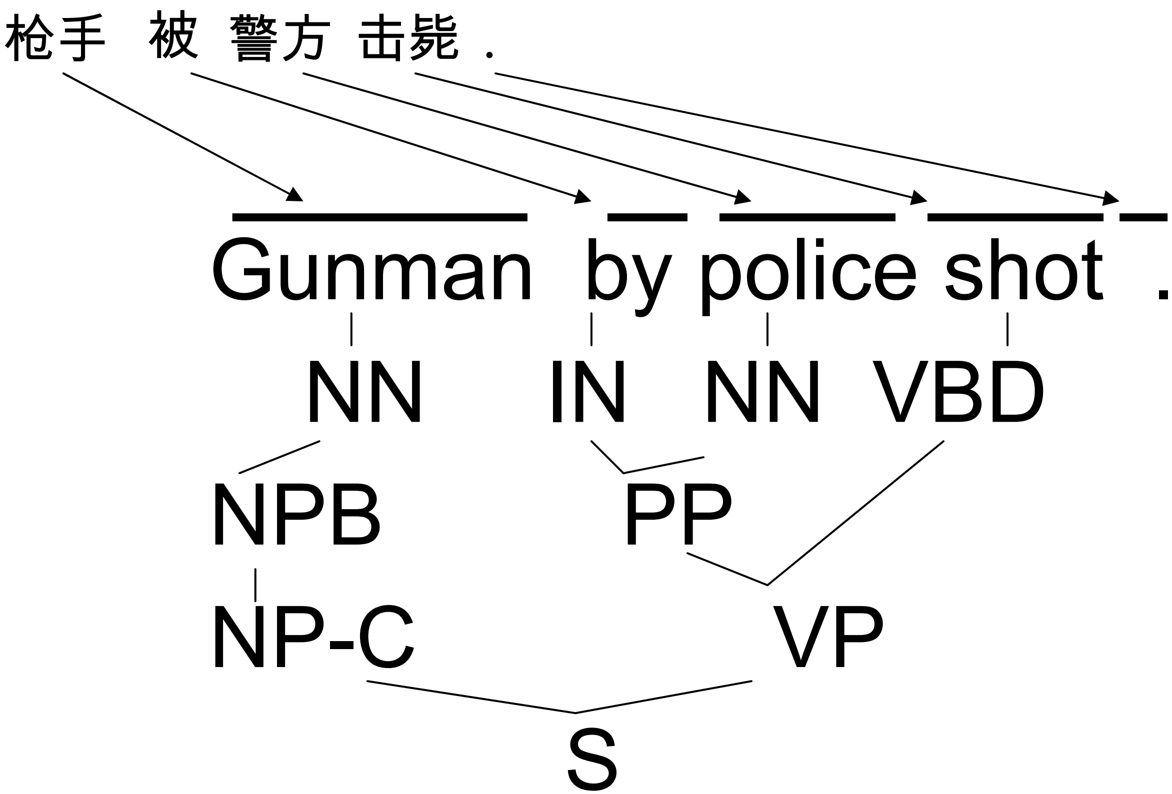
Can’t enforce/encourage any of this!

Syntax-Based Output



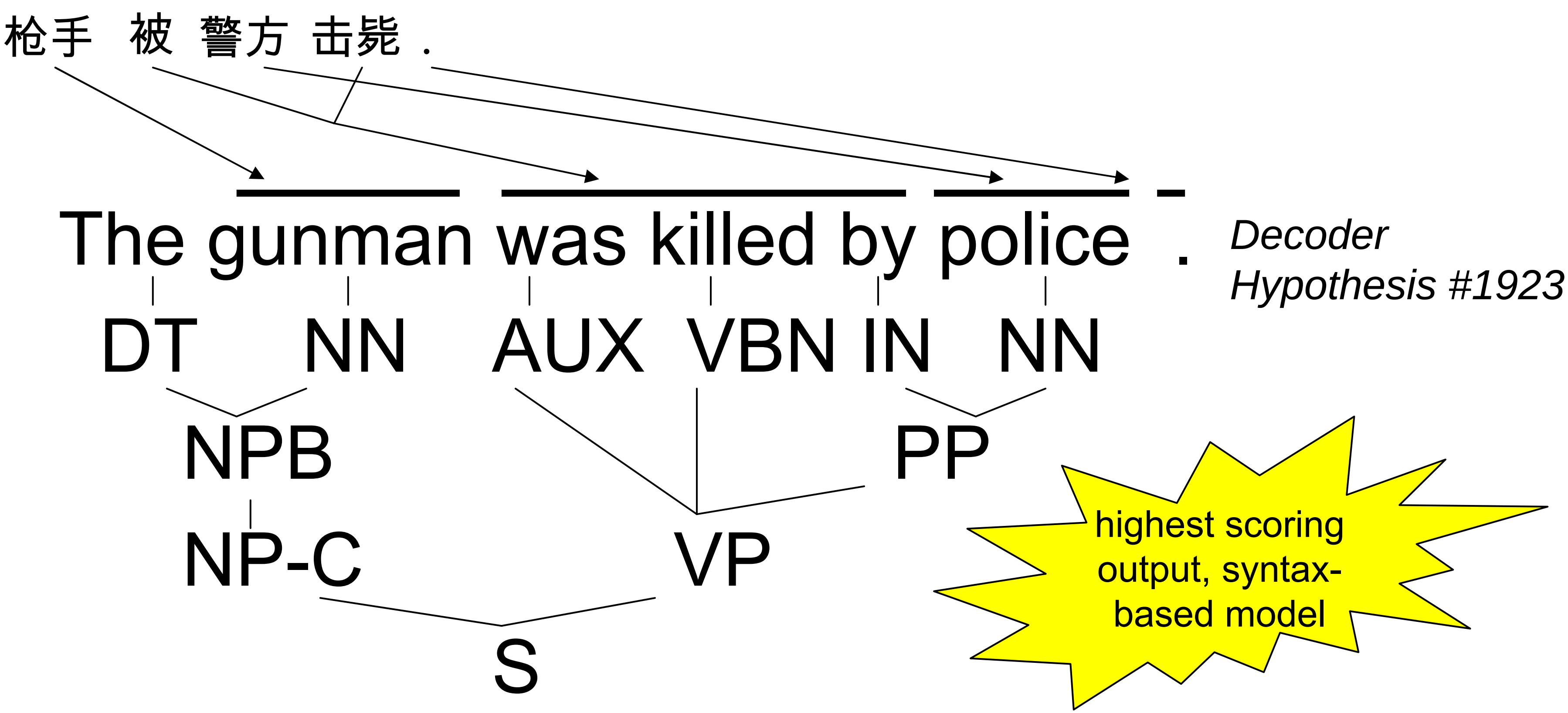
Decoder Hypothesis #1

Syntax-Based Output



Decoder Hypothesis #16

Syntax-Based Output

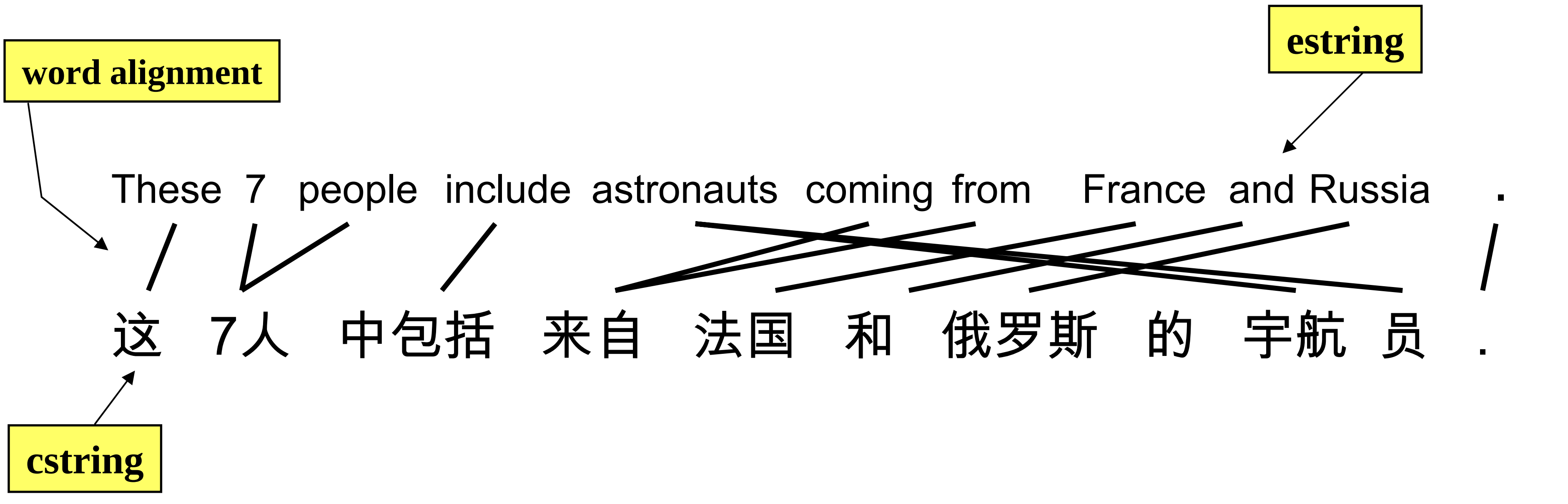


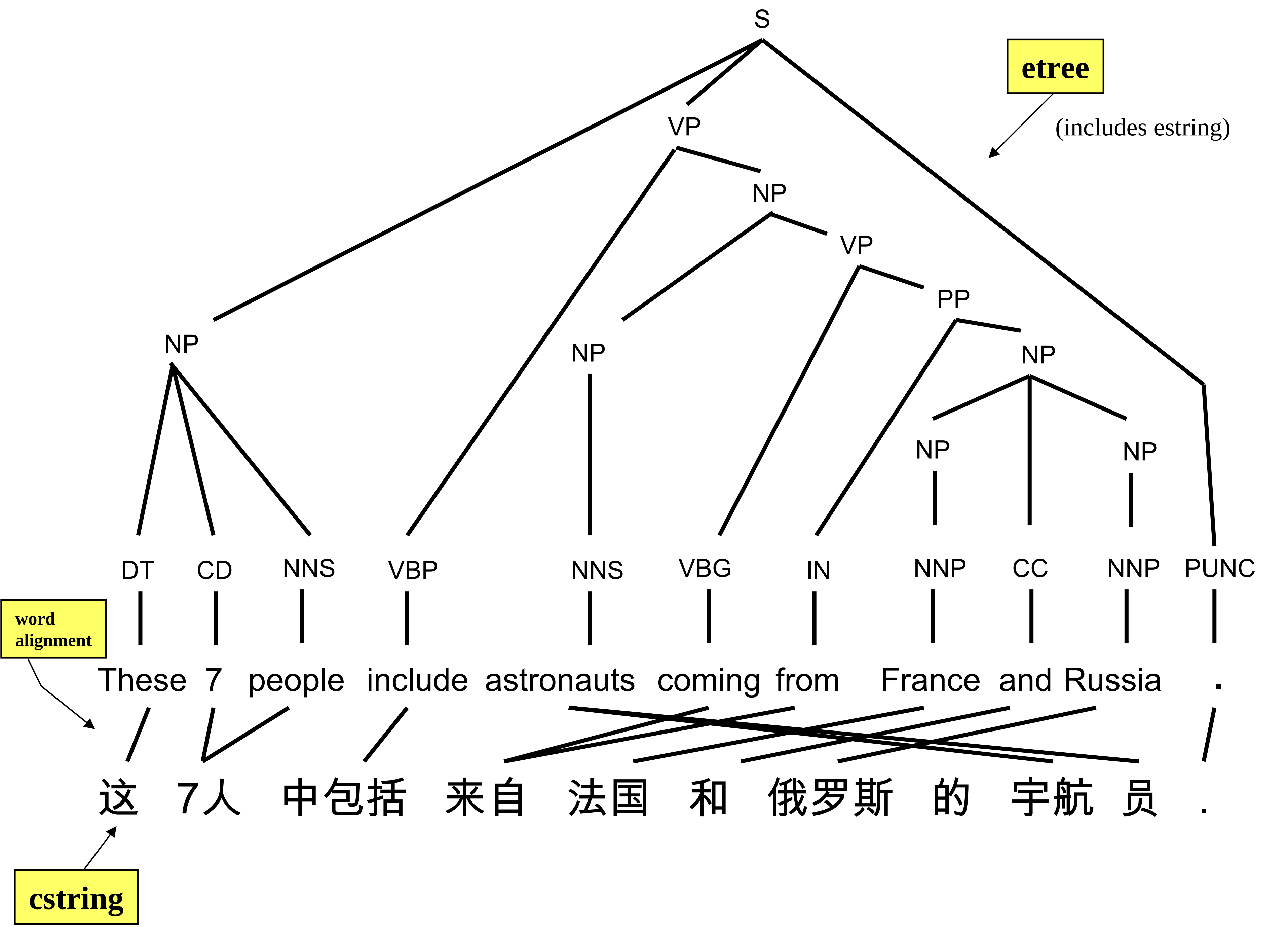
Syntax-Based Output

- Better modeling of target language structure
  - Always a verb
  - Verb is always in the right place
- Better handling of function words
  - They often don’t translate
  - But they control how the translation goes
- Better generalization in translation patterns

Syntax-Based Statistical MT

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


word alignment
estring
These 7 people include astronauts coming from France and Russia .
这 7人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .
cstring


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

Mathematical Framework

- String-based system

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

- Tree-based system

$$\operatorname{argmax}_{etree,a} \; \underbrace{P(etree, a, c)^{\alpha}}_{\text{translation model}} \cdot \underbrace{P(etree)^{\beta}}_{\text{language model}} \cdot \underbrace{|etree|^{\gamma}}_{\text{length bonus}} \cdot \ldots$$

String-to-Tree

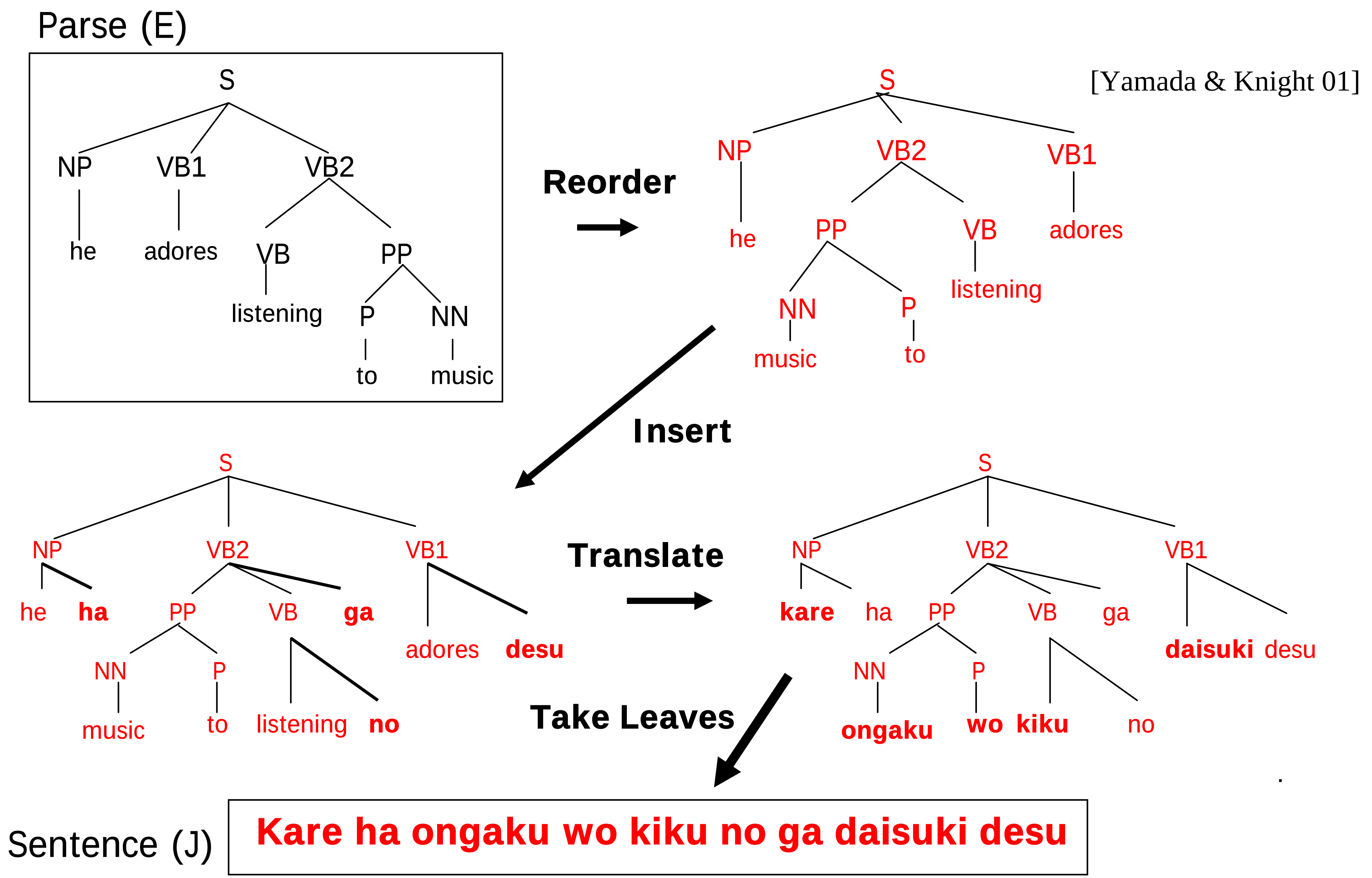
- Mathematically, we want a weighted relation with pairs drawn from:
  - (the infinite) set of Chinese strings
  - (the infinite) set of English trees

- Good pairs should have a high weight
- Bad pairs should have a low weight

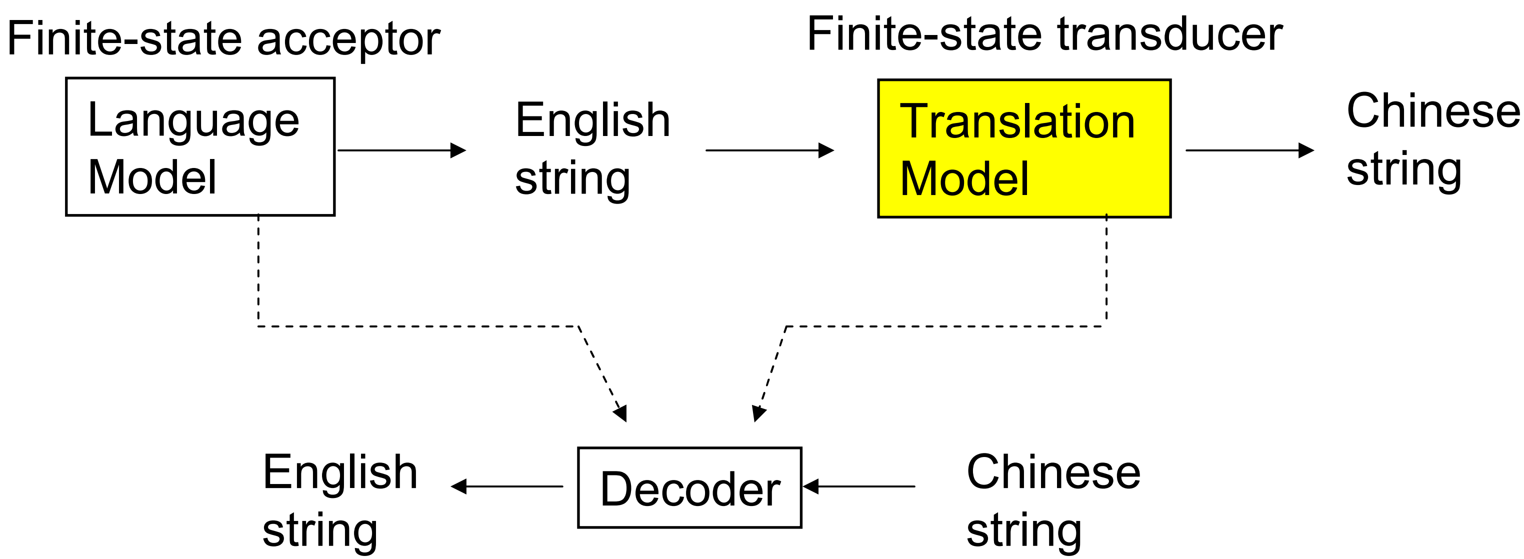
- Probabilistic generative modeling approach
  - How does a Chinese string become an English tree (or vice-versa)?

An Early Syntactic Model of Translation

[Yamada & Knight 01]

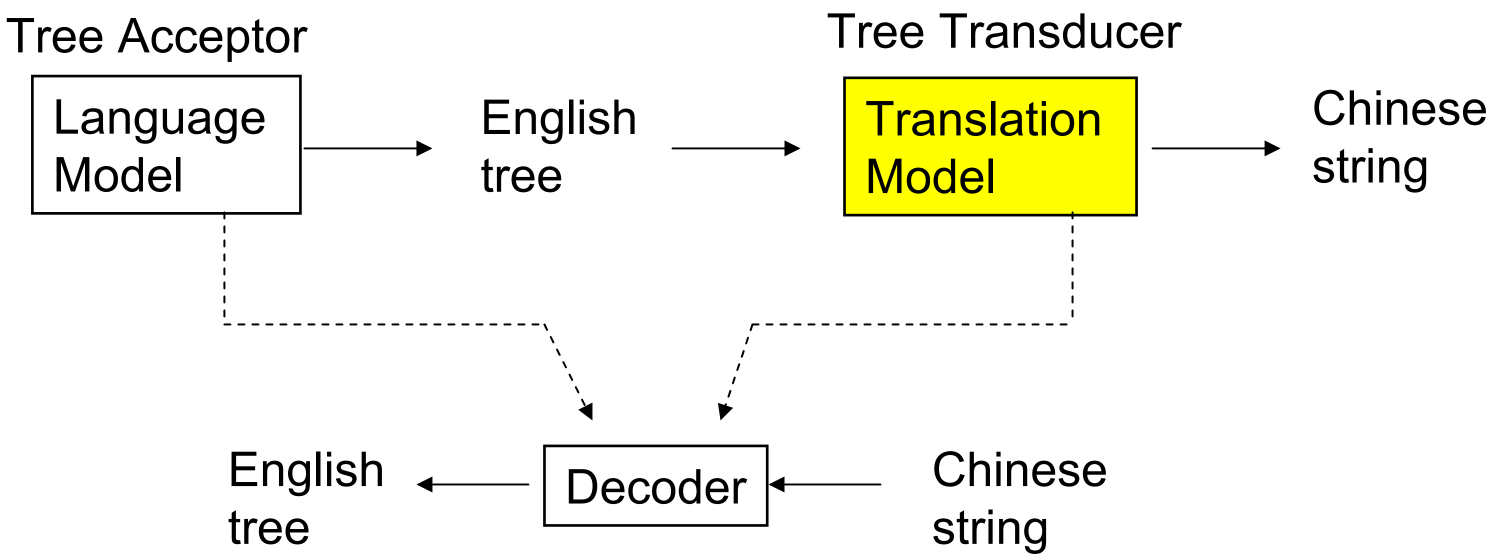


Phrase-Based



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

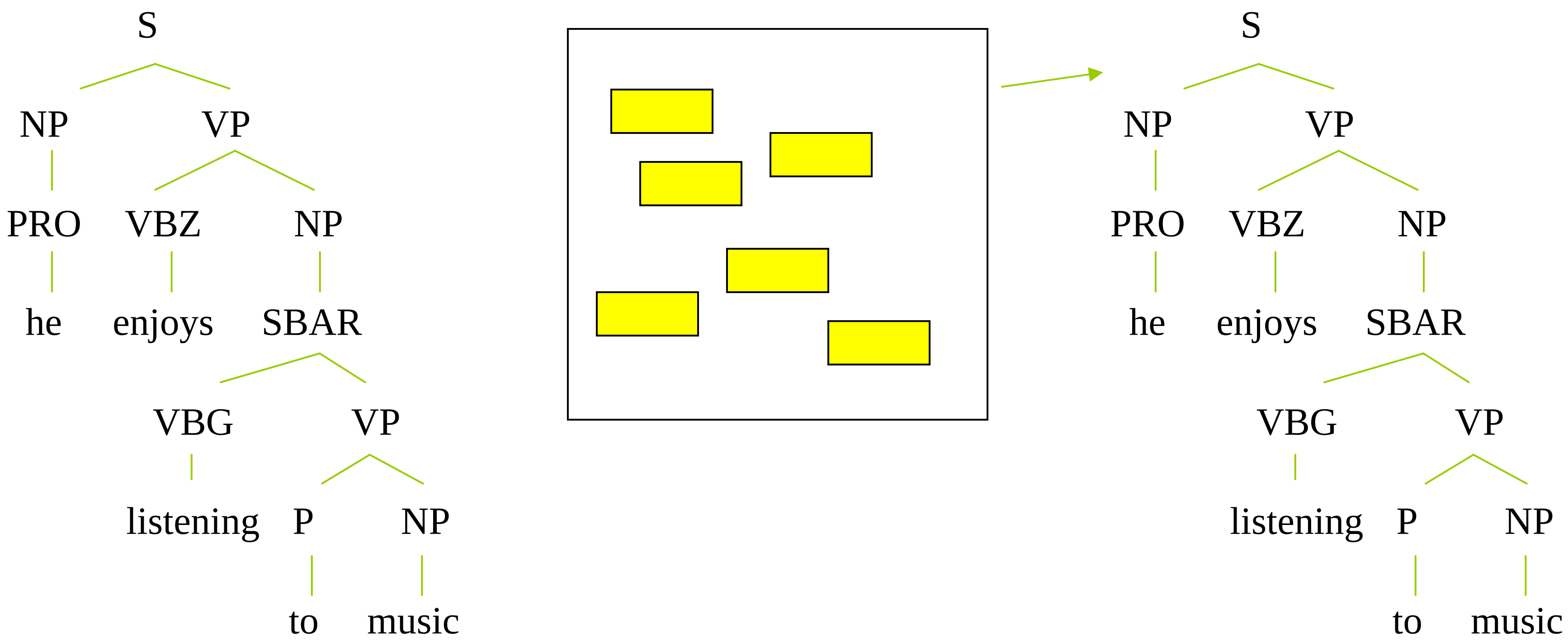
Syntax-Based



- Grab a chunk of English input tree
- Decide how to translate it
- Recurse of remaining subtrees
  - Can be modeled by tree transducer
  - [Rounds, 1970] → [Graehl & Knight, 2004, HLT]

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

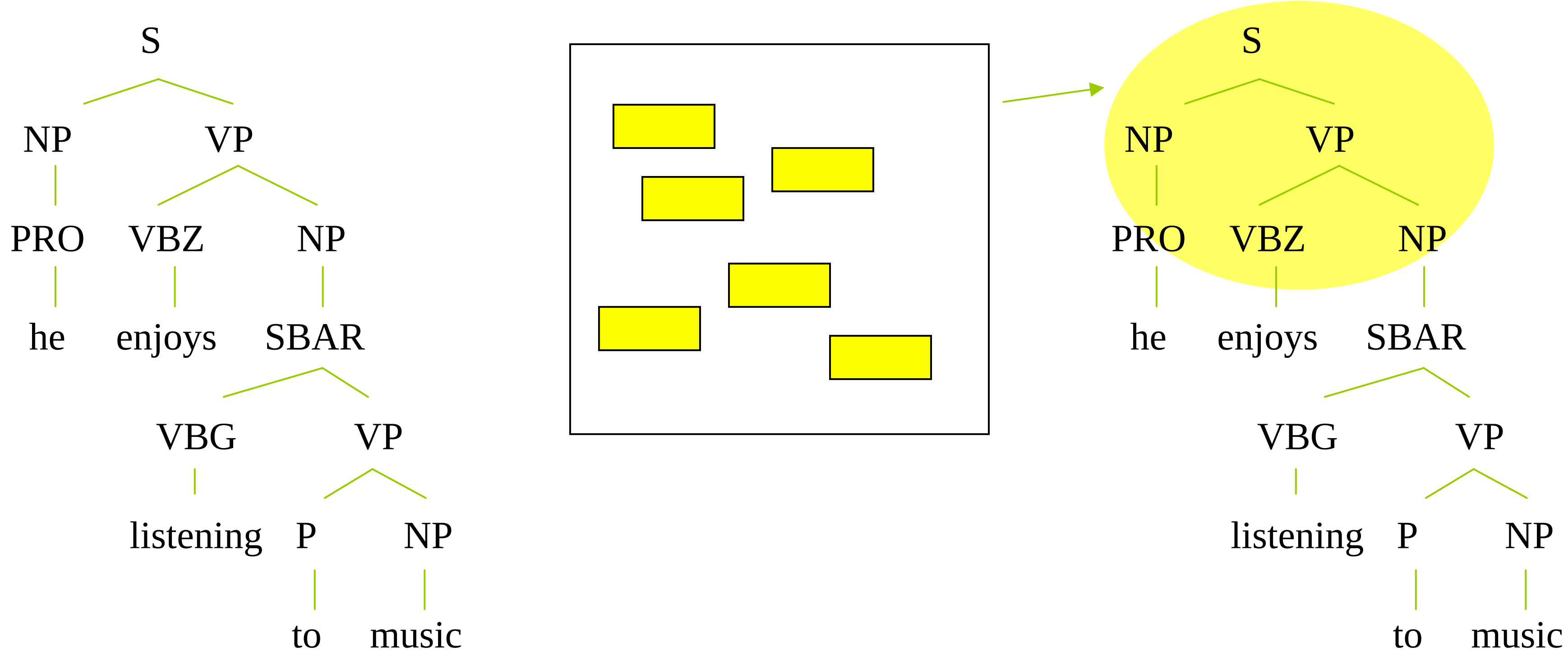
Original input: Transformation:



Top-Down Tree Transducer

(W. Rounds 1970; J. Thatcher 1970)

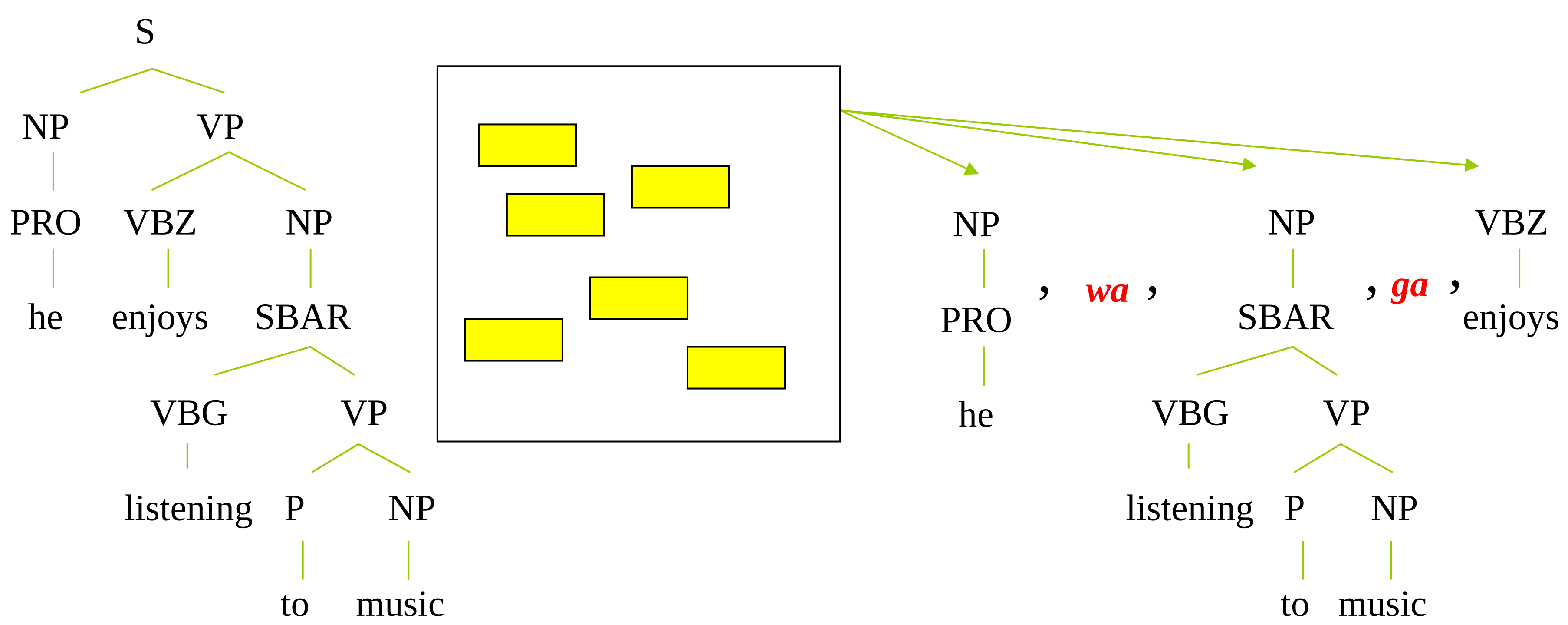
Original input: Transformation:



Top-Down Tree Transducer

(W. Rounds 1970; J. Thatcher 1970)

Original input: Transformation:

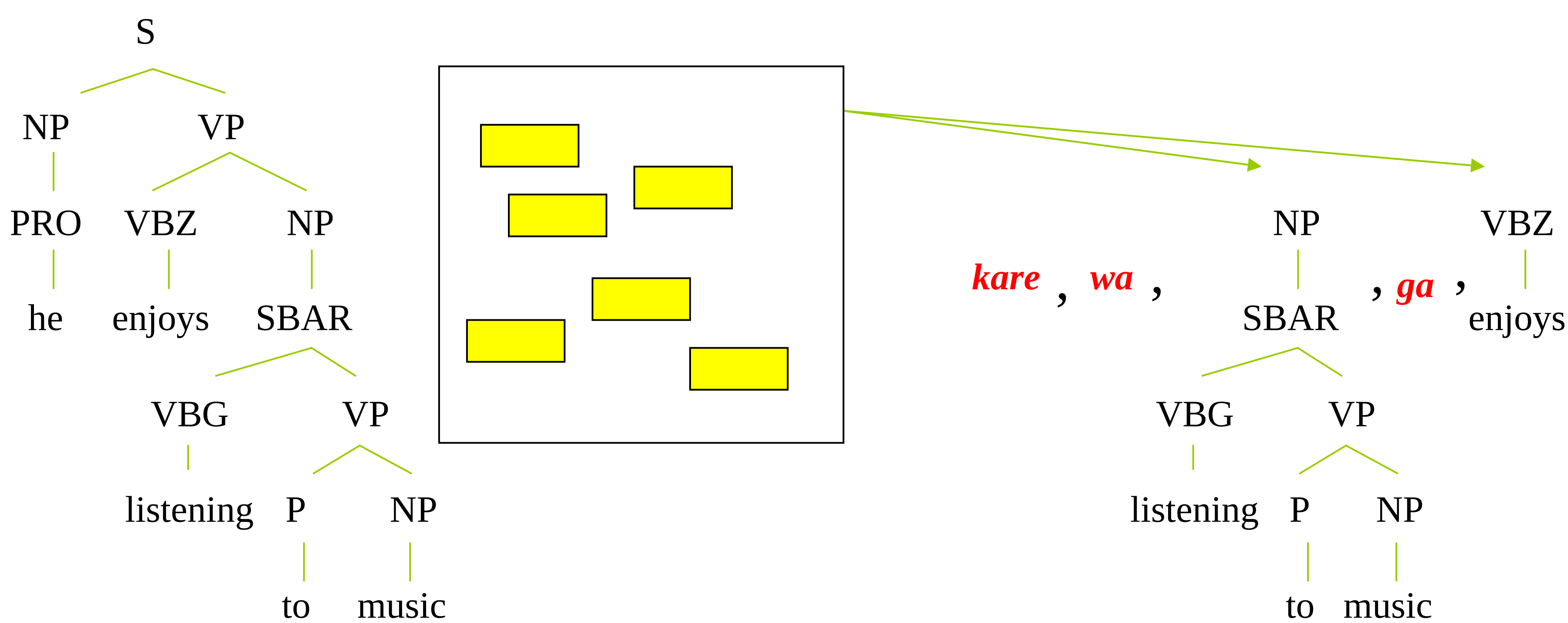


Top-Down Tree Transducer

(W. Rounds 1970; J. Thatcher 1970)

Original input:

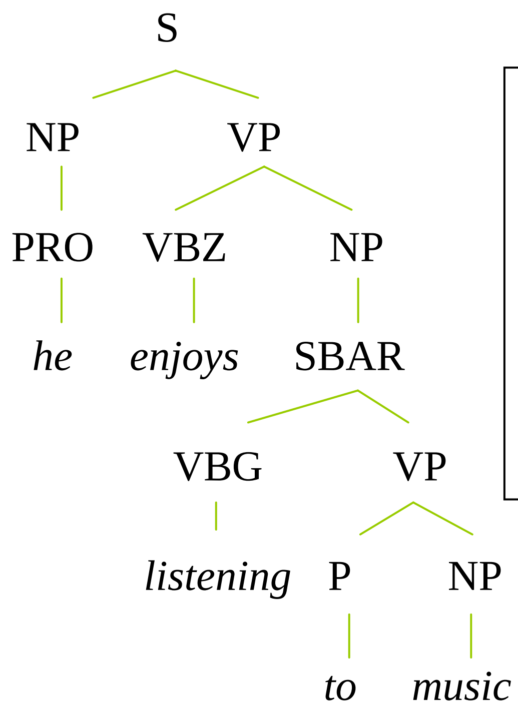
Transformation:

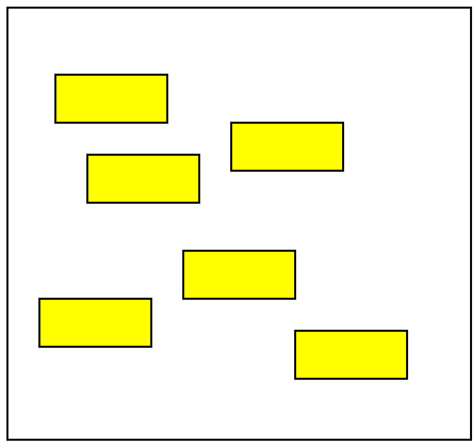


Top-Down Tree Transducer

(W. Rounds 1970; J. Thatcher 1970)

Original input:





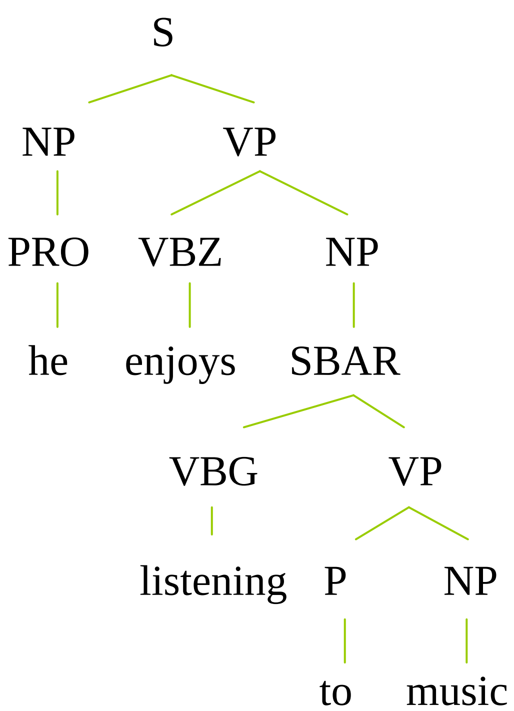
Final output:

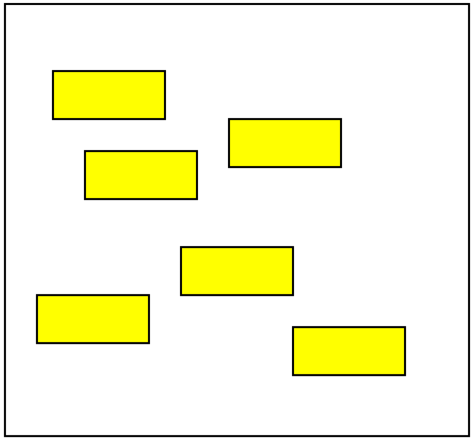
kare*, *wa* , *ongaku*, *o* , *kiku*, *no*, *ga*, *daisuki*, *desu

Top-Down Tree Transducer

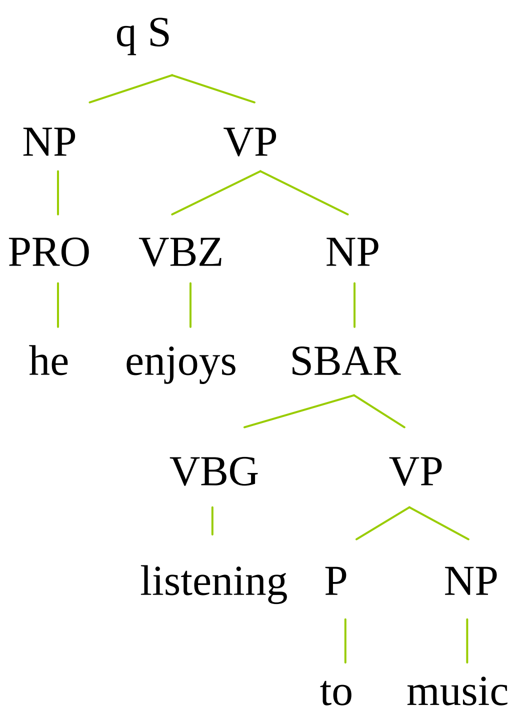
(W. Rounds 1970; J. Thatcher 1970)

Original input:





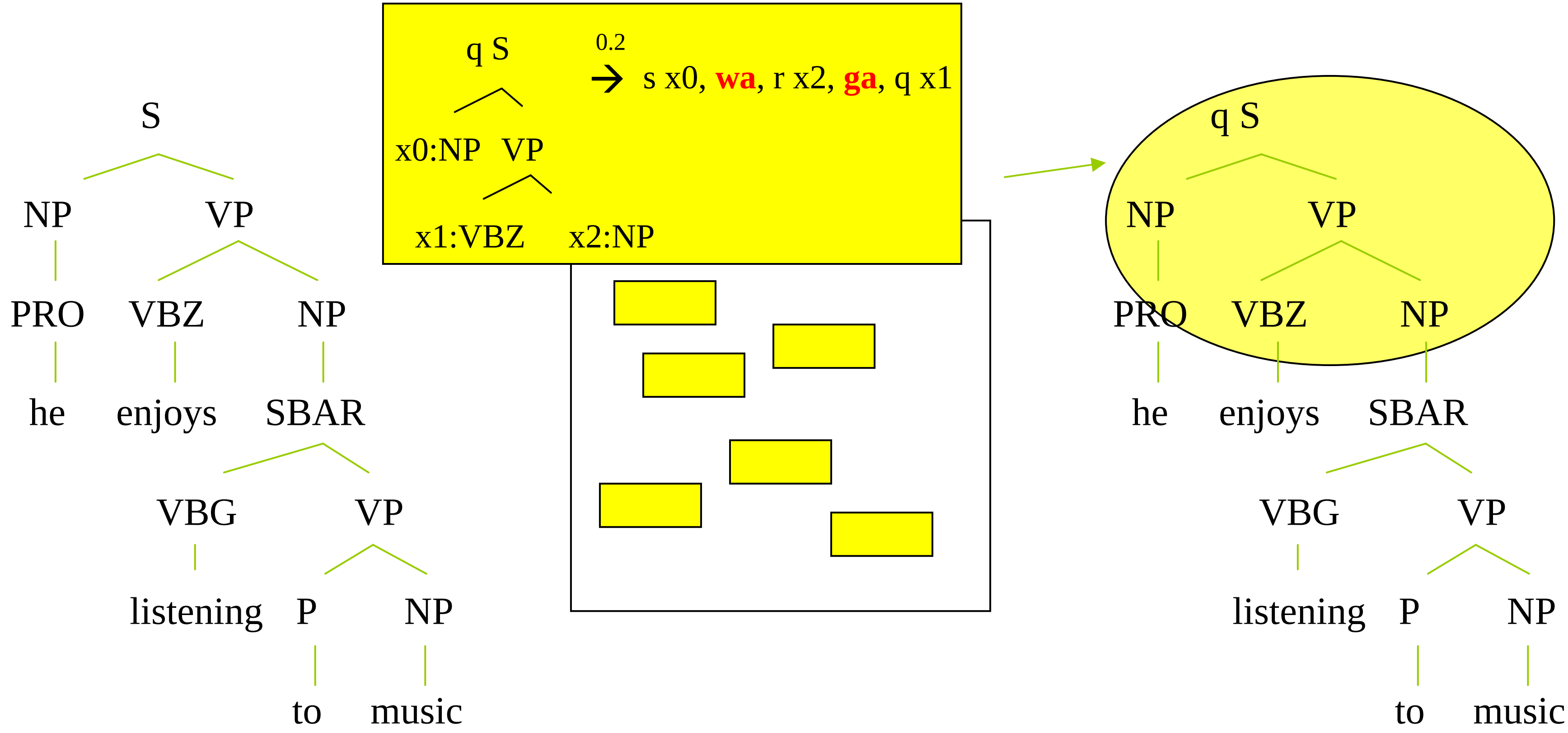
Transformation:



Top-Down Tree Transducer

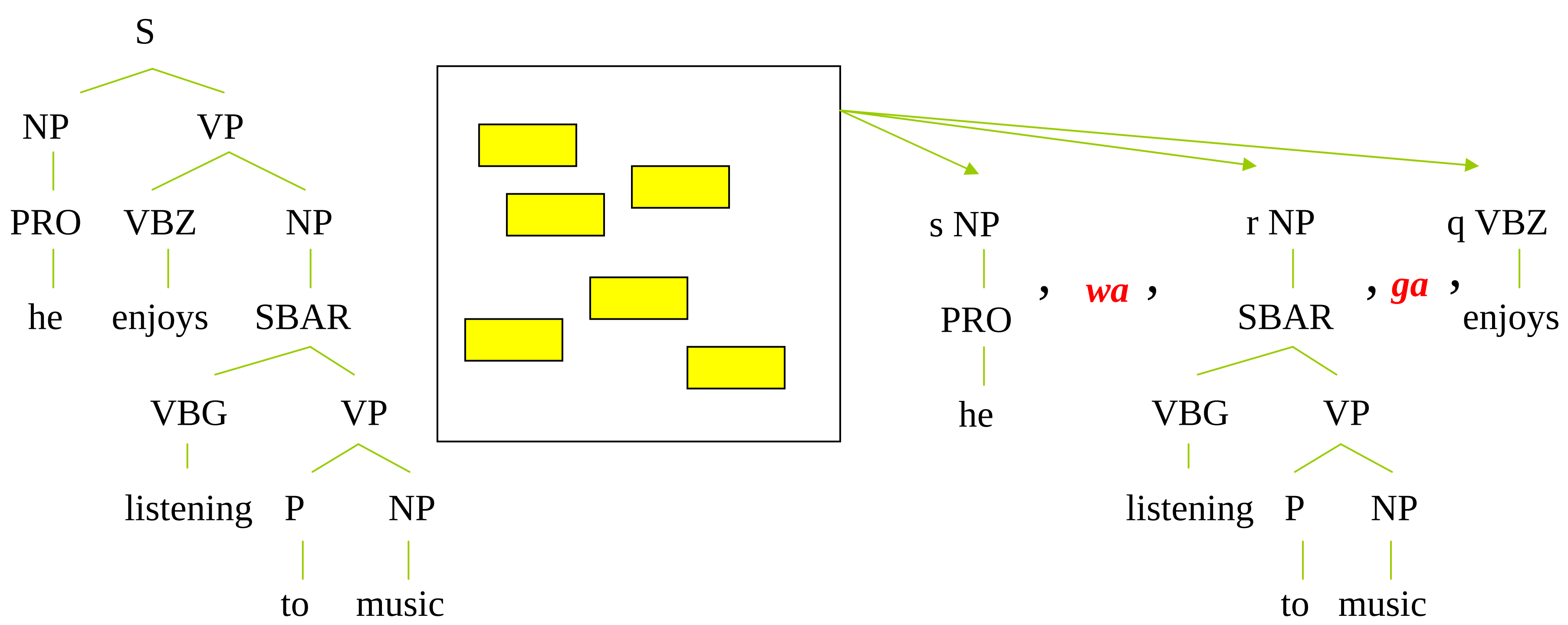
(W. Rounds 1970; J. Thatcher 1970)

Original input: Transformation:



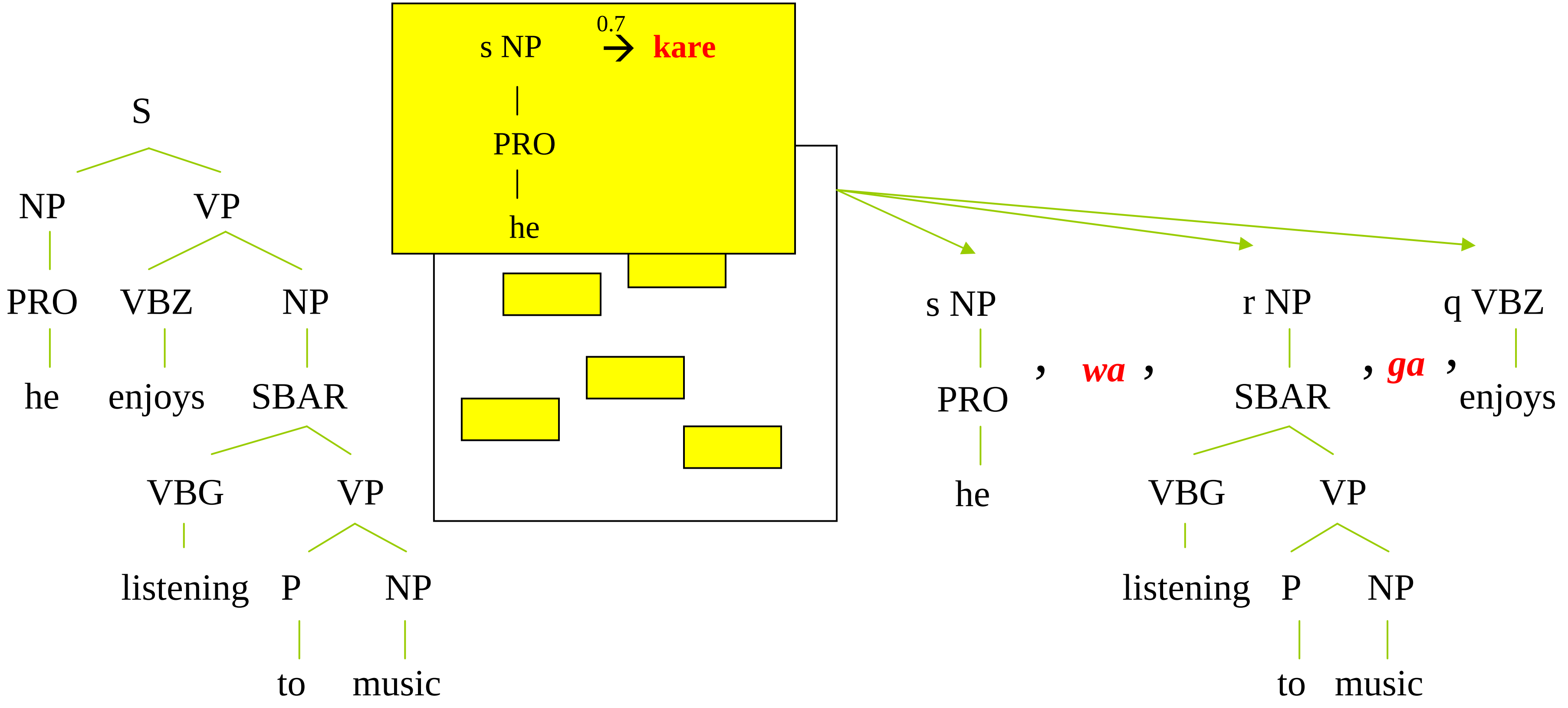
Top-Down Tree Transducer
(W. Rounds 1970; J. Thatcher 1970)

Original input: Transformation:



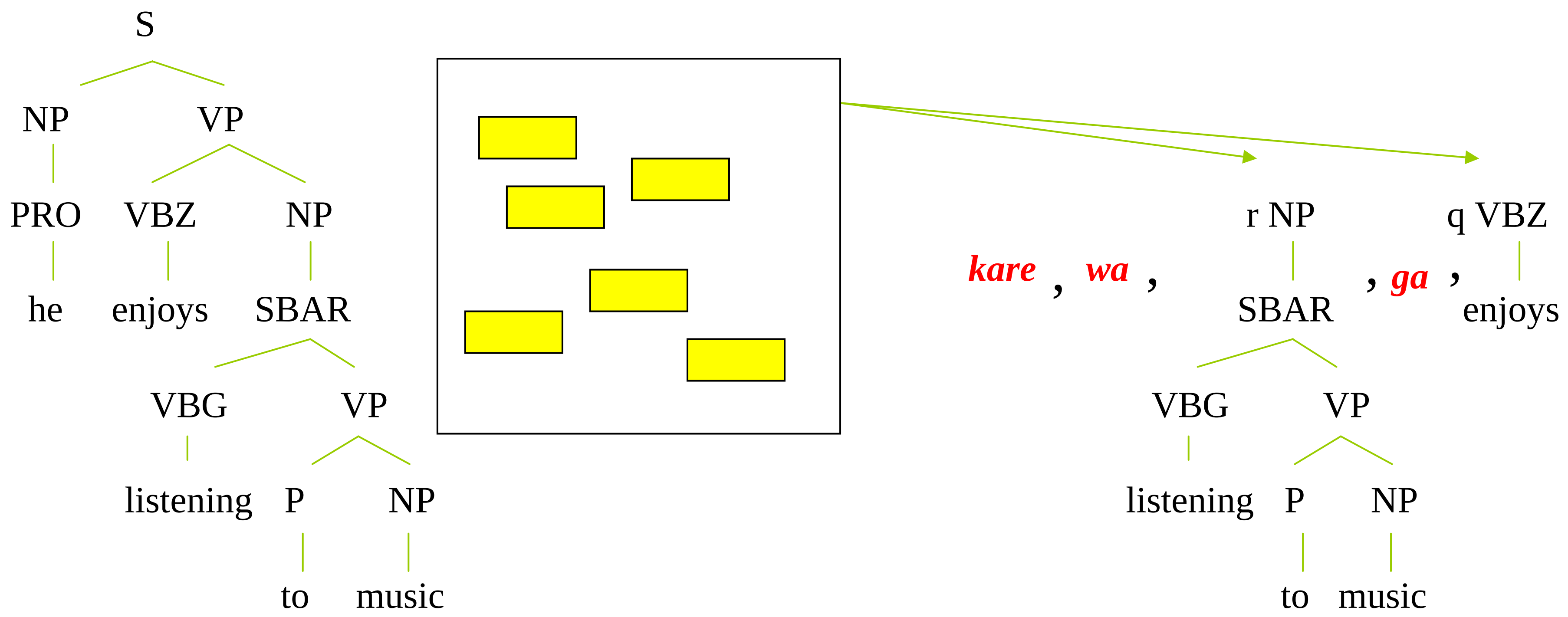
Top-Down Tree Transducer
(W. Rounds 1970; J. Thatcher 1970)

Original input: Transformation:



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

Original input: Transformation:

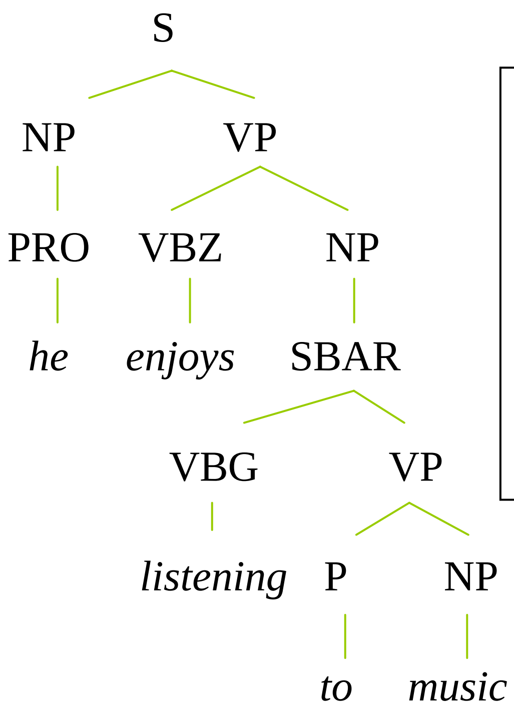


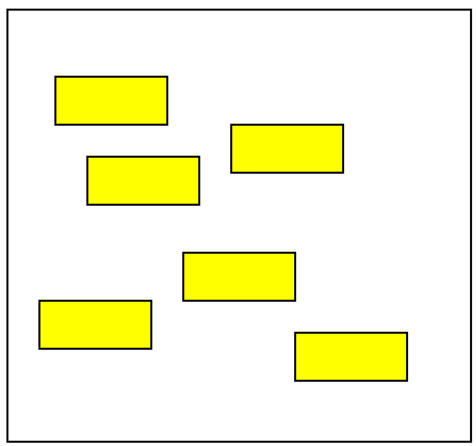
Top-Down Tree Transducer

(W. Rounds 1970; J. Thatcher 1970)

Original input:

Final output:



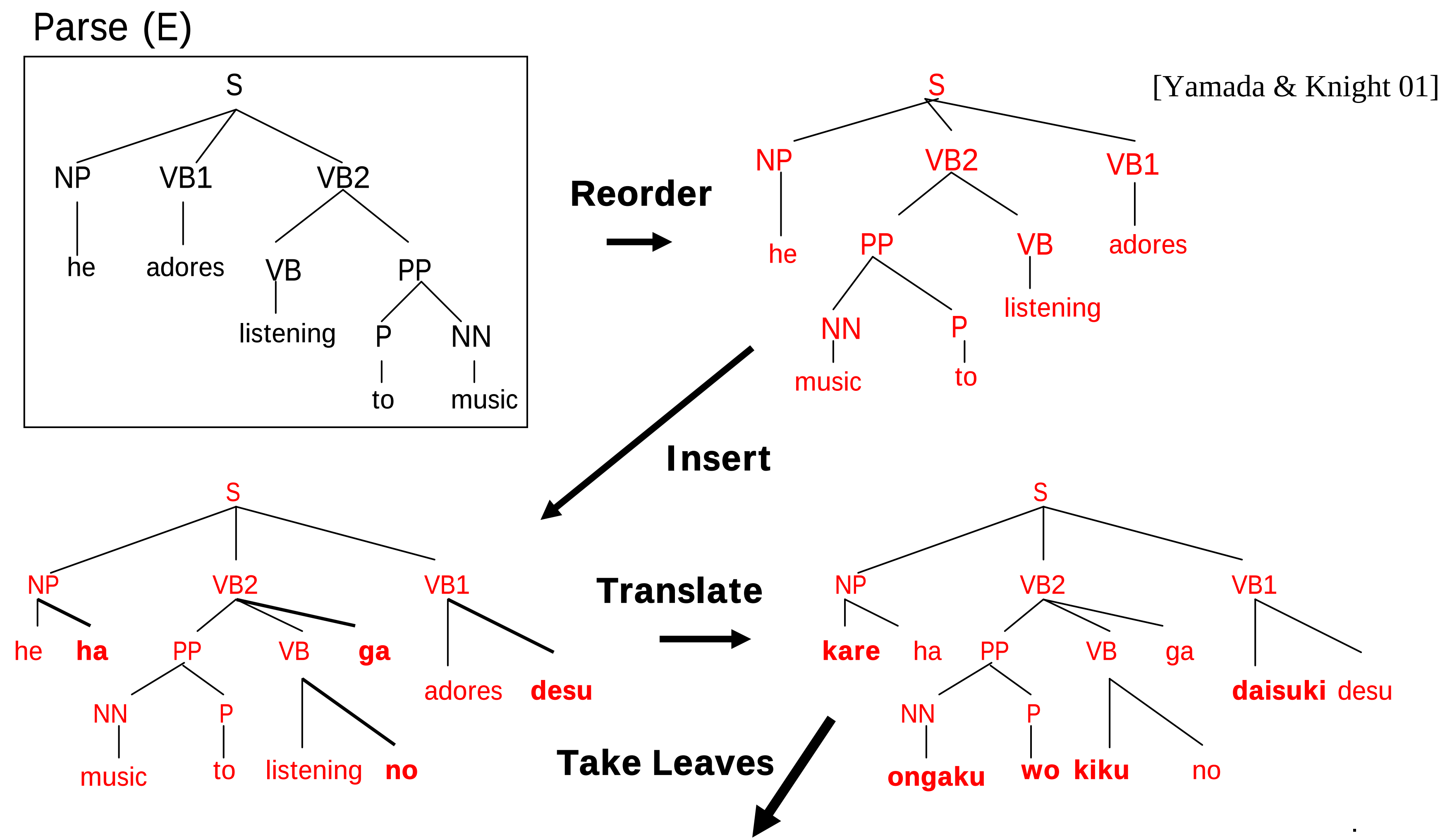


kare*, *wa* , *ongaku*, *o* , *kiku*, *no*, *ga*, *daisuki*, *desu

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

An Early Syntactic Model of Translation

[Yamada & Knight 01]

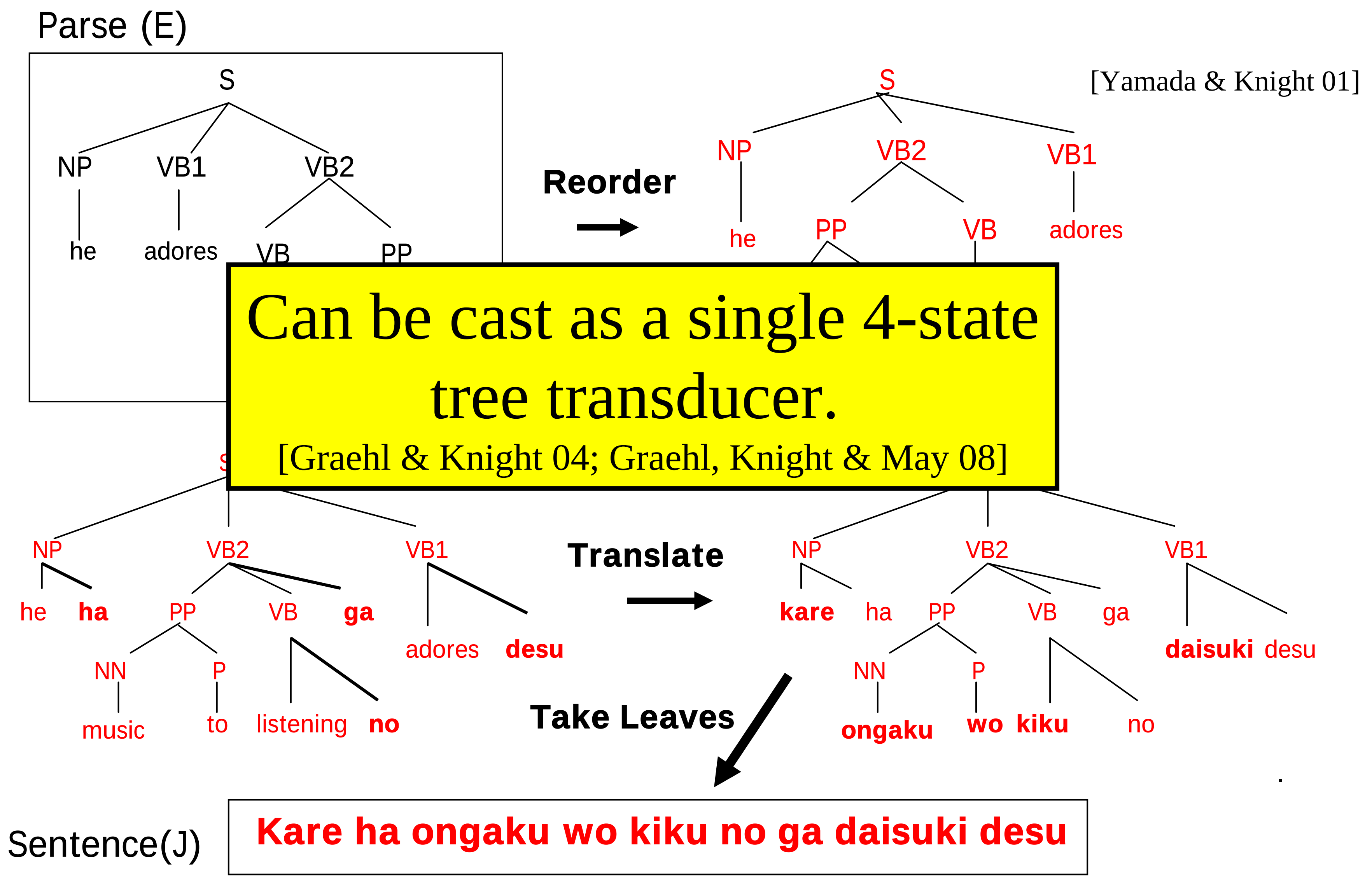


Sentence(J)

Kare ha ongaku wo kiku no ga daisuki desu

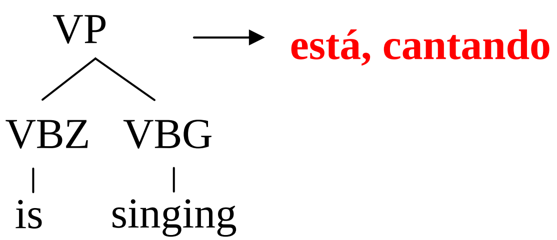
An Early Syntactic Model of Translation

[Yamada & Knight 01]

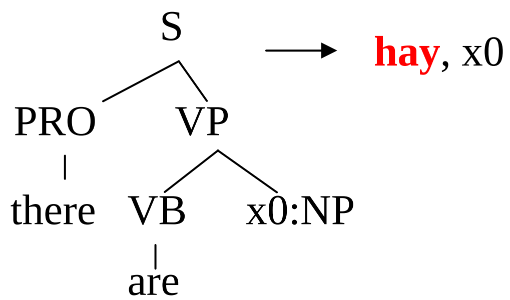


Tree Transducers are Expressive

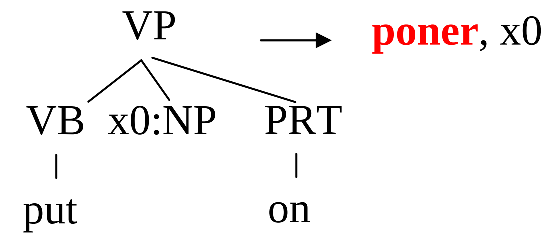
Phrasal Translation



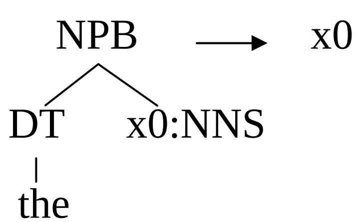
Non-constituent Phrases



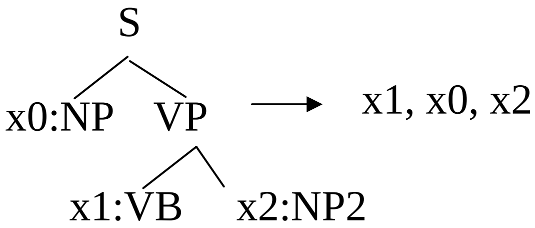
Non-contiguous Phrases



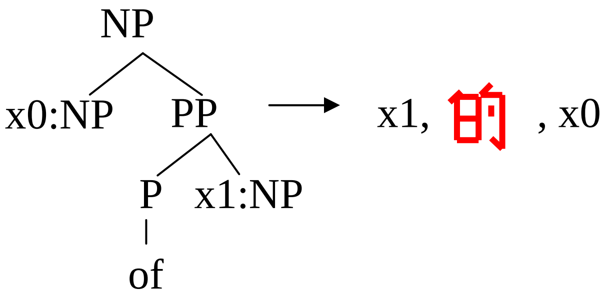
Context-Sensitive Word Insertion



Multilevel Re-Ordering



Lexicalized Re-Ordering

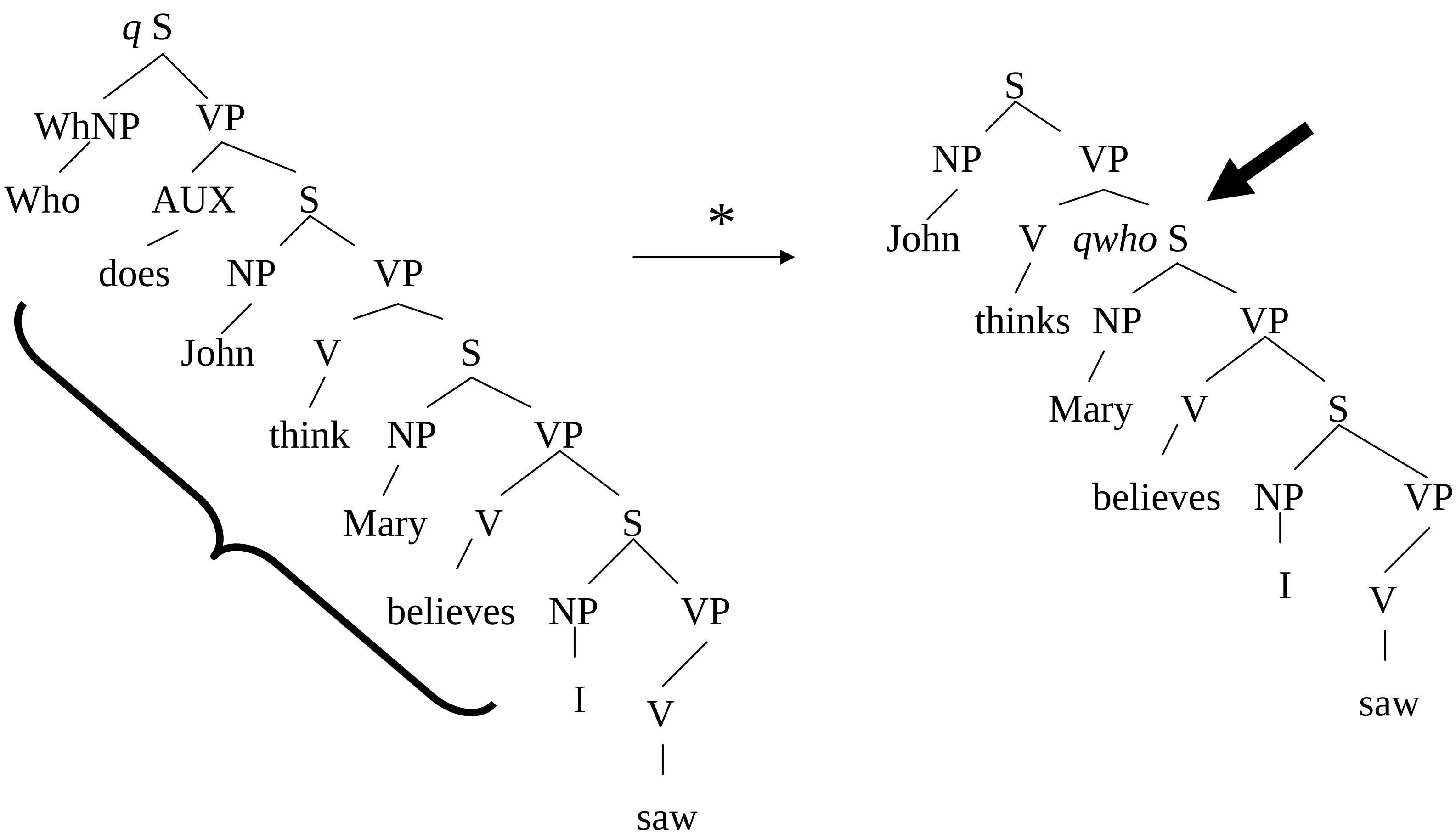


also QA, compression, paraphrasing, etc

most probabilistic tree-based models proposed 2000-2005 can be so cast

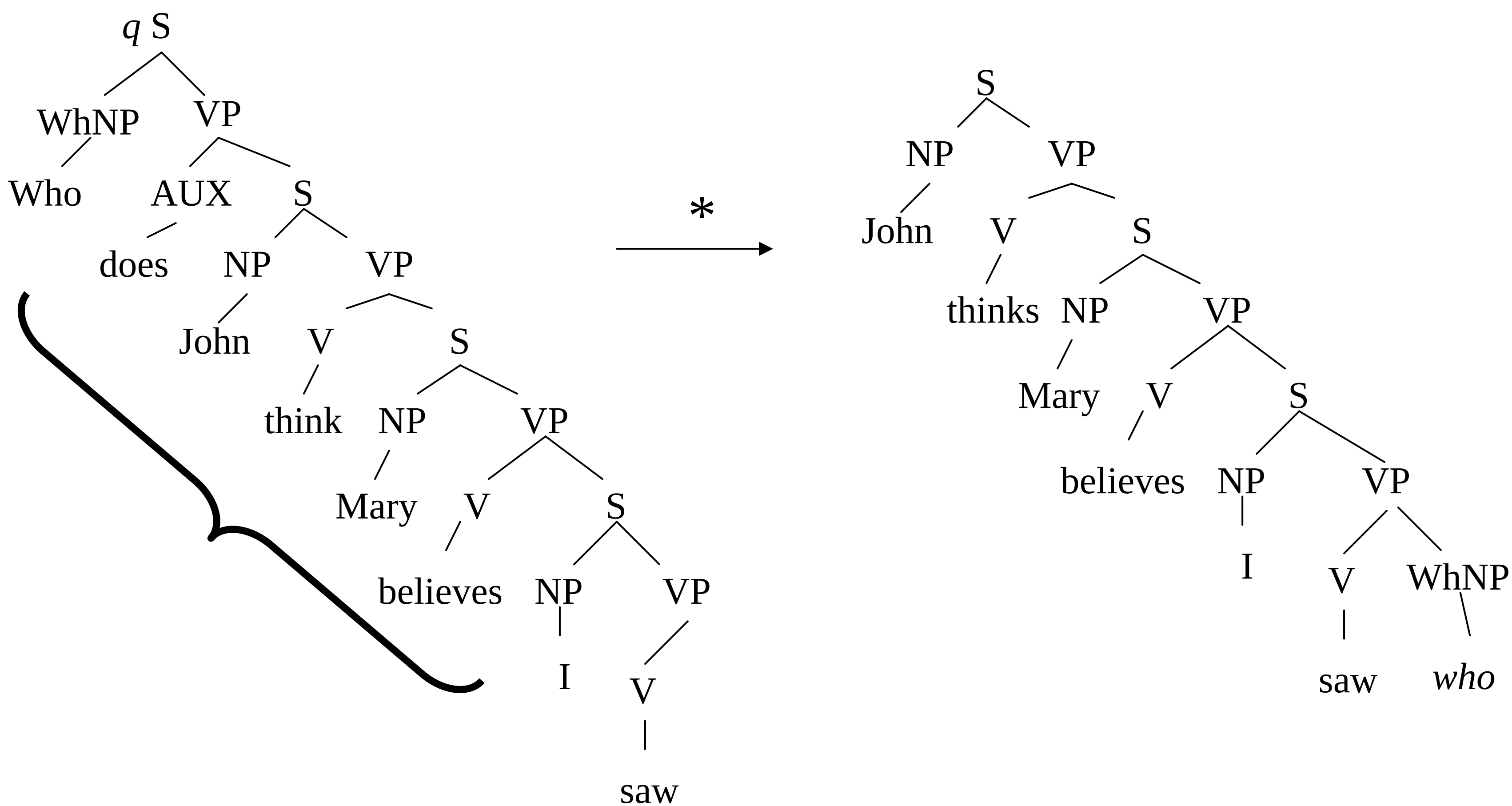
Limitations of the Top-Down Transducer Model

Who does John think Mary believes I saw? → John thinks Mary believes I saw *who*?



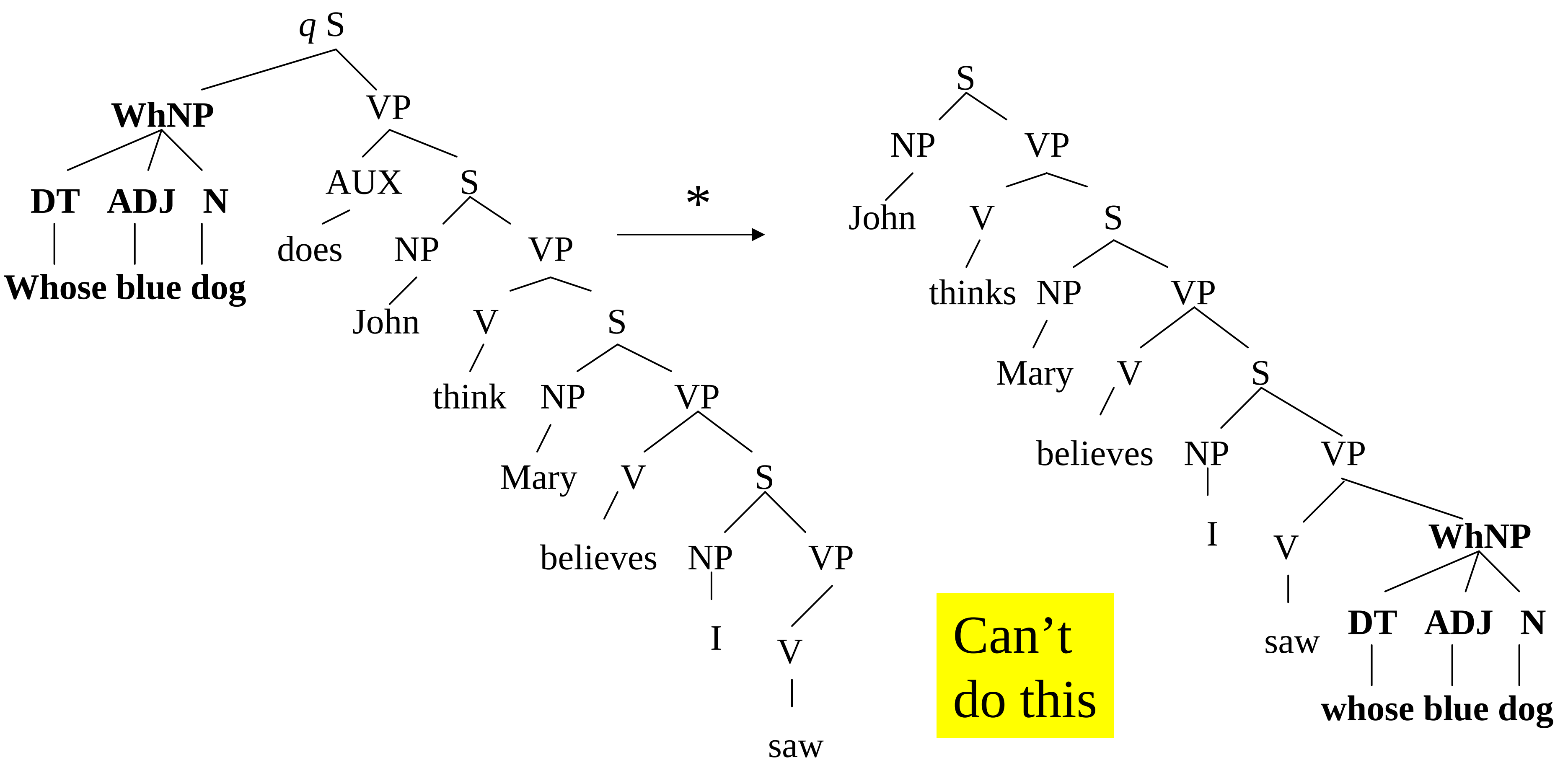
Limitations of the Top-Down Transducer Model

Who does John think Mary believes I saw? → John thinks Mary believes I saw *who*?



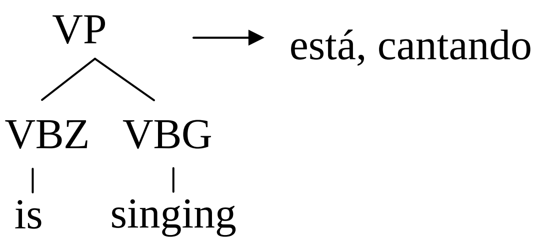
Limitations of the Top-Down Transducer Model

Whose blue dog does John think Mary believes I saw**?** → John thinks Mary believes I saw ***whose blue dog***?

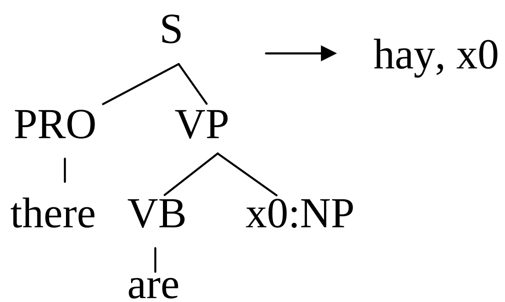


Computer-Friendly Format for Tree Transducer Rules

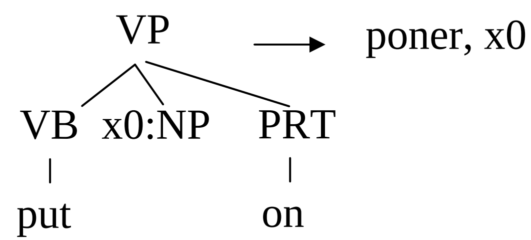
Phrasal Translation



Non-constituent Phrases



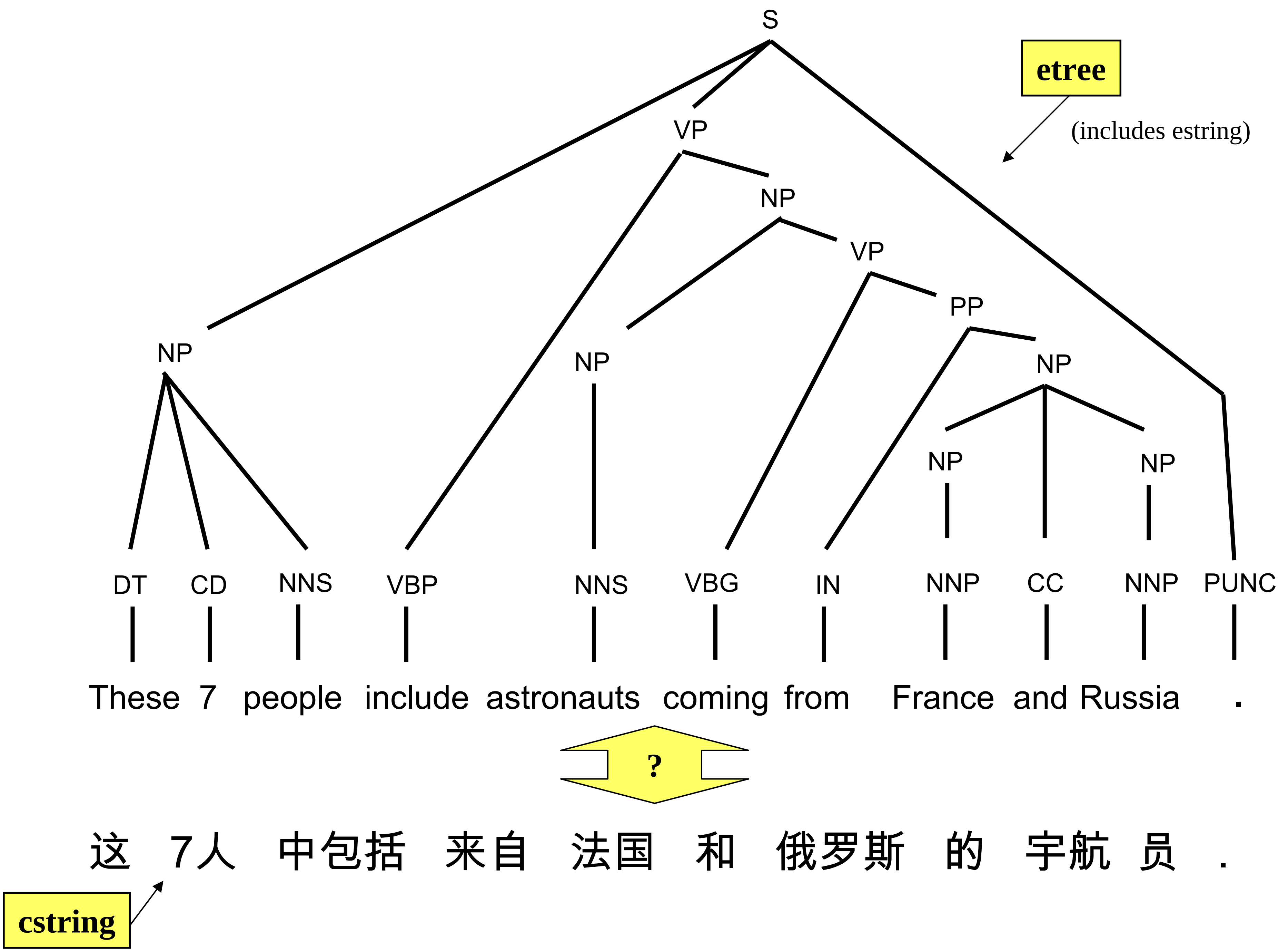
Non-contiguous Phrases



VP(VBZ(is), VBG(singing)) → está, cantando

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

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


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

Tree Transformations

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

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

Contiguous phrase pair substitution rules (alignment templates)

Higher-level rules

Tree Transformations

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

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

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

pair

tes)

Higher-level rules

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

这	7人	中包括	来自	法国	和	俄罗斯	的	宇航	员	.
the	7 people	including	by some		**and**	the russian	**the**	the astronauts		,
it	7 people included		by france		and the	the russian		international astronautical	of rapporteur .	
this	7 out	including the	**from**	the french	and the russian		the fifth		**.**	
these	7 among	including from		the french and		of the russian	of	space	members	**.**
that	7 persons	including from the		of france	and to	russian	of the	aerospace	members .	
	7 include		from the	of france and		russian		**astronauts**		. the
	7 numbers include		**from france**		and russian		of astronauts who			. "
	7 populations include		those from france		and russian			astronauts .		
	7 deportees included		come from	**france**	**and russia**		in	astronautical	personnel	;
	7 philtrum	including those from		**france and**		**russia**	a space		**member**	
		including representatives from		france and the		**russia**		astronaut		
		include	came from	**france and russia**			by cosmonauts			
		include representatives from		french	**and russia**			cosmonauts		
		include	came from france		and russia 's			cosmonauts .		
		includes	coming from	french and		russia 's		cosmonaut		
				french and russian			's	astronavigation	member .	
				french	**and russia**		**astronauts**			
					and russia 's				special rapporteur	
					, and	**russia**			rapporteur	
					, and russia				rapporteur .	
					, and russia					
					or	russia 's				

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

lattice

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

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

这	7人	中包括	来自	法国	和	俄罗斯	的	宇航	员	.
the	7 people	including	by some		**and**	the russian	**the**	the astronauts		,
it	7 people included		by france		and the	the russian		international astronautical	of rapporteur .	
this	7 out	including the	**from**	the french	and the russian		the fifth		**.**	
these	7 among	including from		the french and		of the russian	of	space	members	**.**
that	7 persons	including from the		of france	and to	russian	of the	aerospace	members .	
	7 include		from the	of france and		russian		**astronauts**		. the
	7 numbers include		**from france**		and russian		of astronauts who			. ”
	7 populations include		those from france		and russian			astronauts .		
	7 deportees included		come from	**france**	**and russia**		in	astronautical	personnel	;
	7 philtrum	including those from		**france and**		**russia**	a space		**member**	
		including representatives from		france and the		**russia**		astronaut		
		include	came from	**france and russia**			by cosmonauts			
		include representatives from		french	**and russia**			cosmonauts		
		include	came from france		and russia 's			cosmonauts .		
		includes	coming from	french and		russia 's		cosmonaut		
				french and russian			's	astronavigation	member .	
				french	**and russia**		**astronauts**			
					and russia 's				special rapporteur	
					, and	**russia**			rapporteur	
					, and russia				rapporteur .	
					, and russia					
					or	russia 's				

lattice

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

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

Tree Transformations

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

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

rase pair
es
plates)

Higher-level rules

Tree Transformations

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

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

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

pair

es)

Higher-level rules

Tree Trans

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

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

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

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

r

Higher-level rules

Tree Transformations

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

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

NNS(astronauts) → 宇航 , 员

NP(x0:NNS) → x0

pair

es)

Higher-level rules

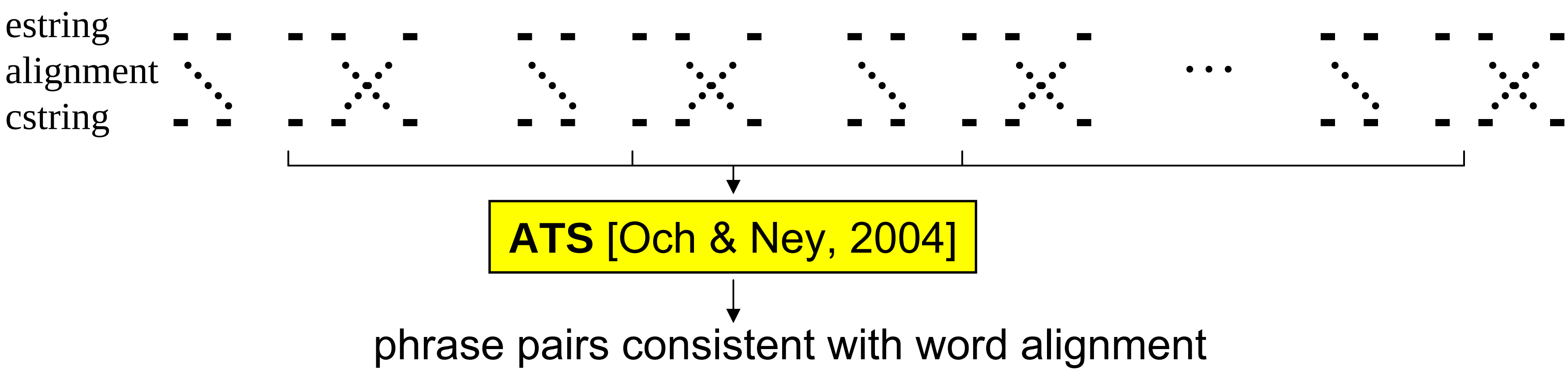
Tree Transformations

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

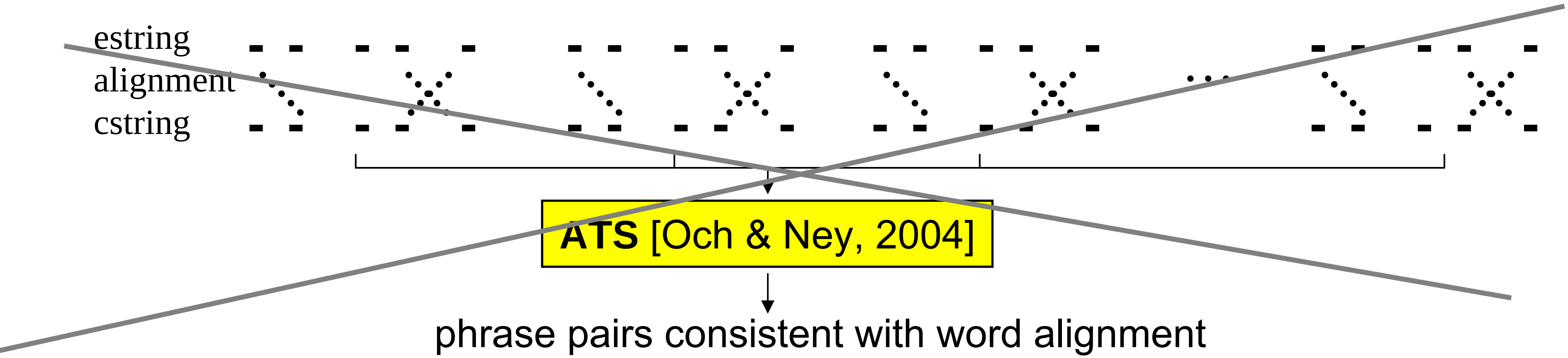
Okay, these rules look interesting.

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

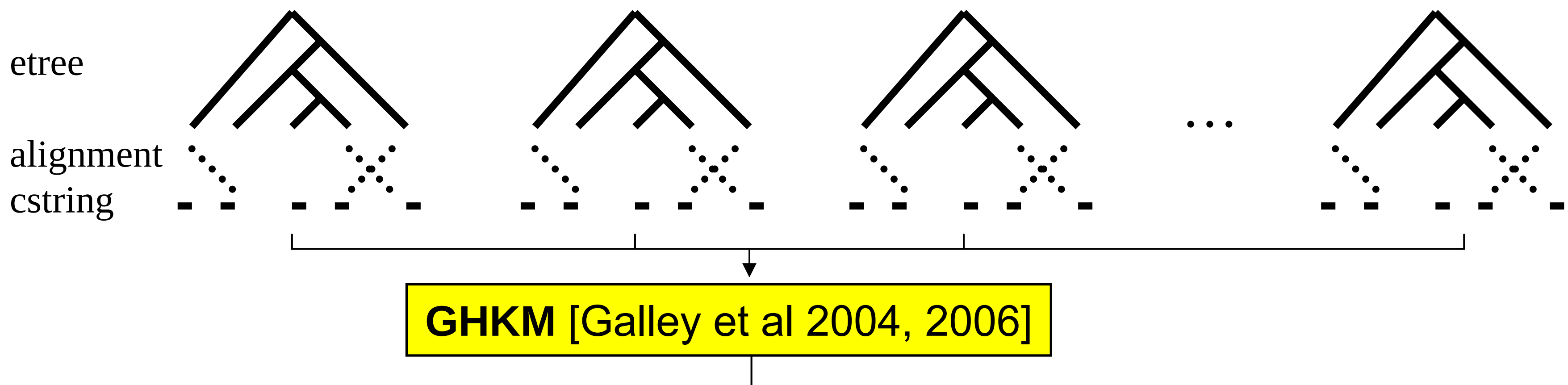
Phrase-Based and Syntax-Based Pattern Extraction



Phrase-Based and Syntax-Based Pattern Extraction



phrase pairs consistent with word alignment



syntax transformation rules consistent with word alignment

Tree Transducers Can be Extracted from Data

(Galley, Hopkins, Knight, Marcu, 2004)

RULES ACQUIRED:

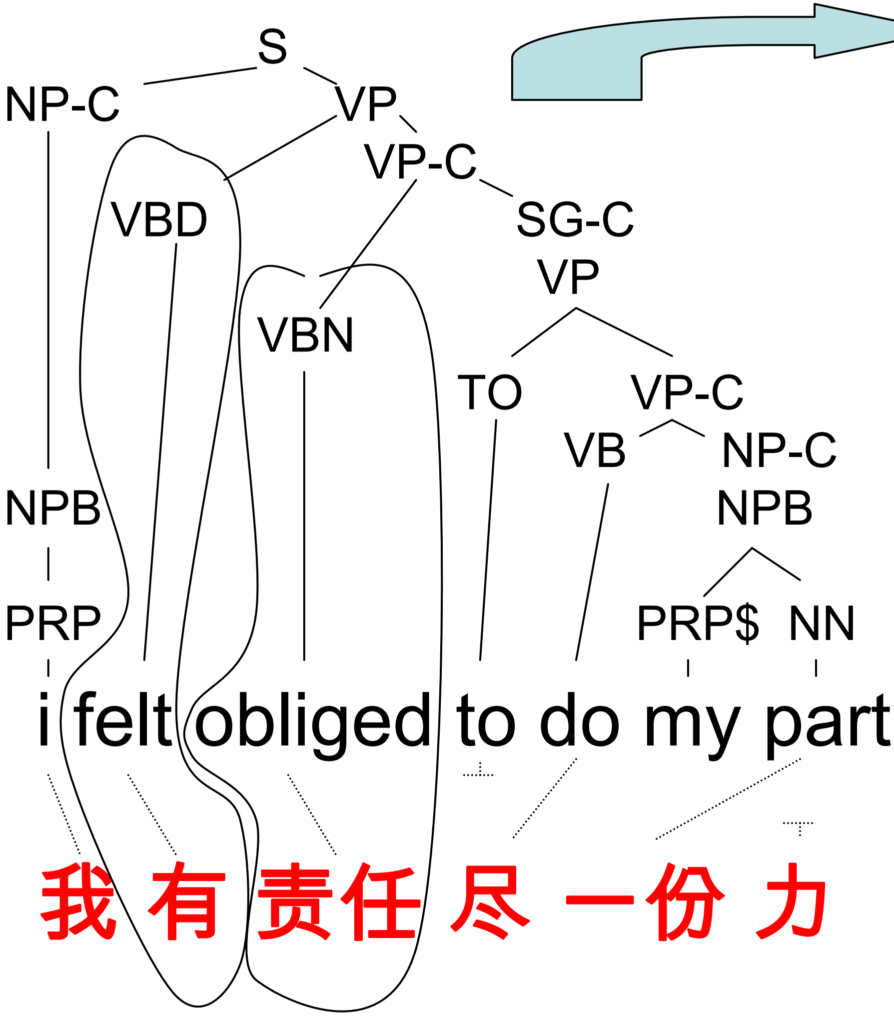
VBD(felt) → 有
VBN(obliged) → 责任
VB(do) → 尽
NN(part) → 一份
NN(part) → 一份 力
VP-C(x0:VBN x1:SG-C) → x0 x1
VP(TO(to) x0:VP-C) → x0
…
S(x0:NP-C x1:VP) → x0 x1

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

RULES ACQUIRED:

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

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

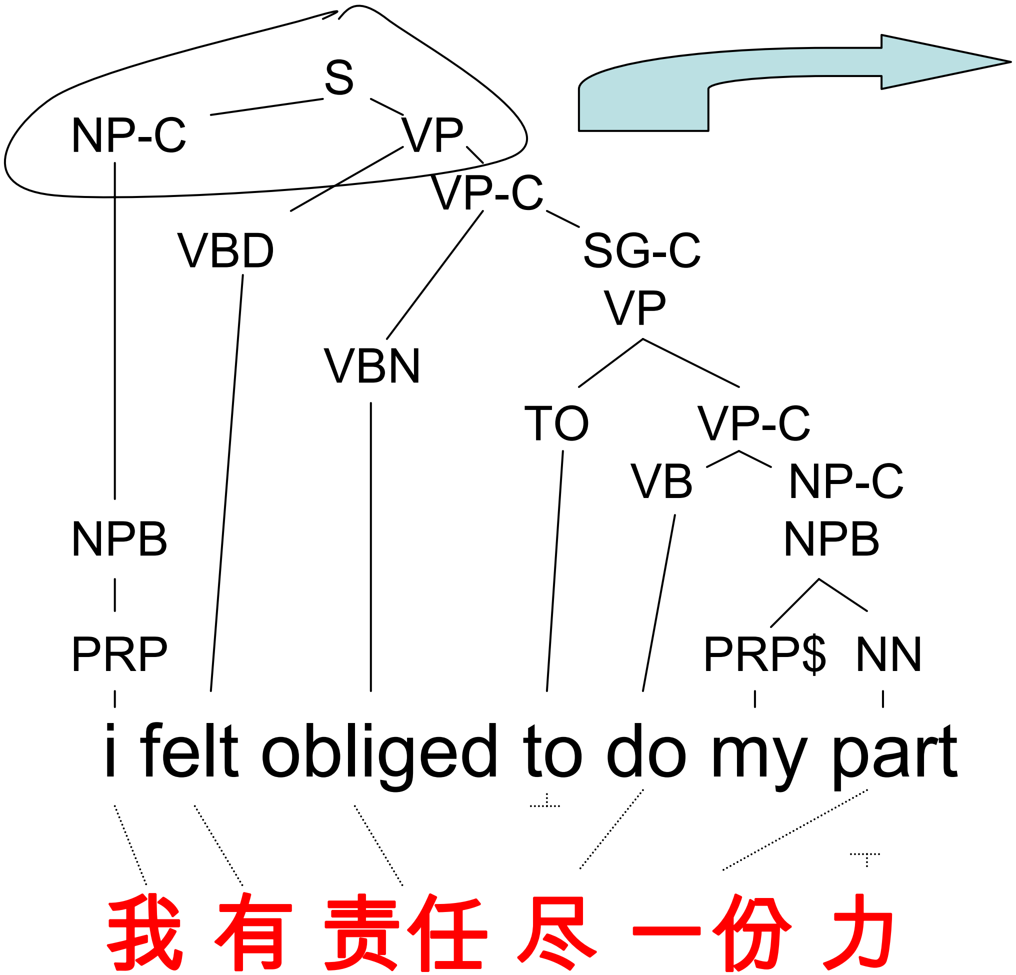


RULES ACQUIRED:

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

Tree Transducers Can be Extracted from Data

(Galley, Hopkins, Knight, Marcu, 2004)

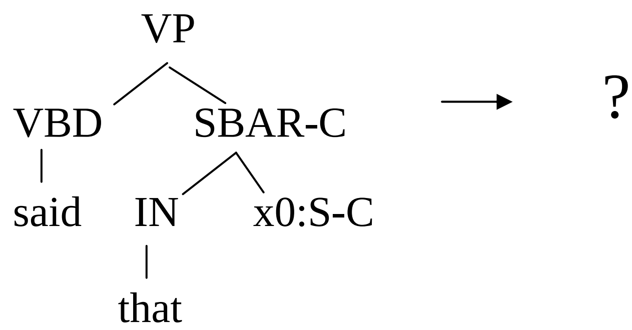


RULES ACQUIRED:

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

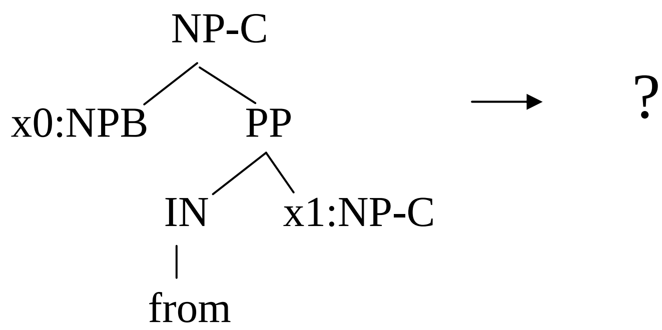
There is a unique tiling that identifies minimal translation units.

Sample "said that" rules



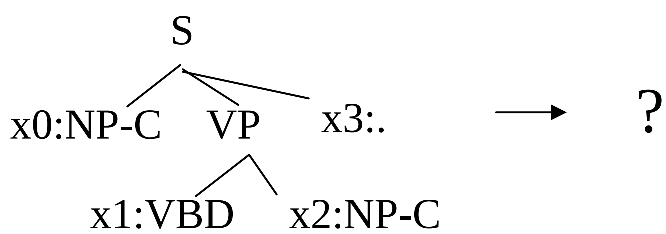
0.57	VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 说 , x0
0.09	VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 说 x0
0.02	VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 他 说 , x0
0.02	VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 指出 , x0
0.02	VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> x0
0.01	VP(VBD("said")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

Sample “NP-from-NP” rules



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

Sample SVO rules



CHINESE / ENGLISH

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

ARABIC / ENGLISH

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

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

Enumerate all ways of dealing with unaligned Chinese words.

Generate rule counts which can be normalized into probabilities.

Language Models

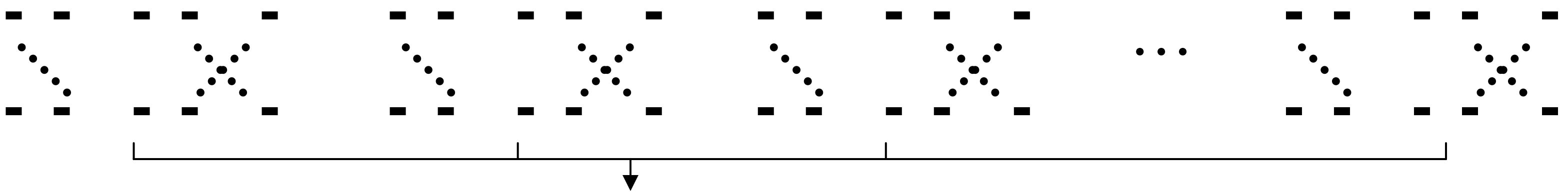
- Syntax-based Language Model
  - Assigns P(tree)
    - [Collins 97; Charniak 01]
  - NOTE: Unlike parser, must be trained on domain data
- Ngram Language Model
  - Standard trigram model
  - Only judges a tree by its leaves

BREAK

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

Phrase-Based and Syntax-Based Pattern Extraction

estring
alignment
cstring

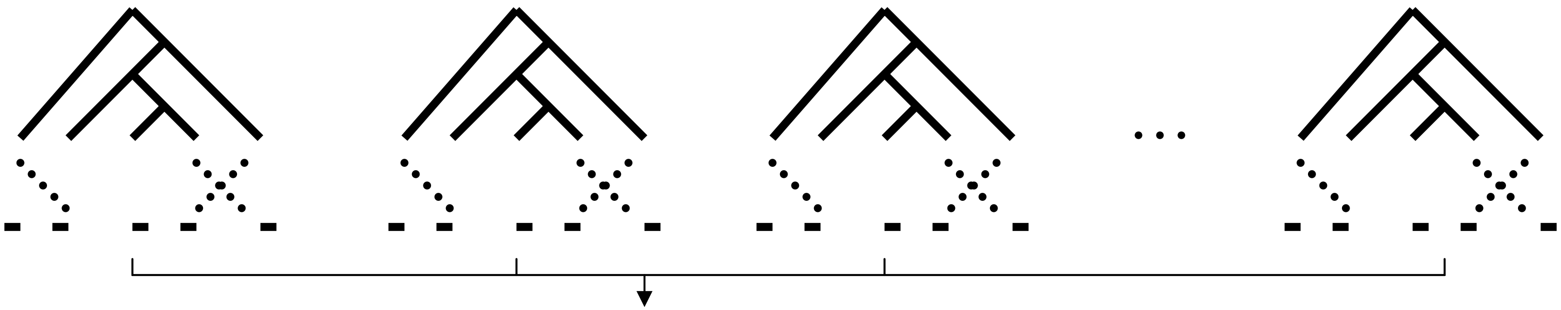


ATS [Och & Ney, 2004]

phrase pairs consistent with word alignment

etree

alignment
cstring



GHKM [Galley et al 2004, 2006]

syntax transformation rules consistent with word alignment

Tree Transducers Can be Extracted from Data

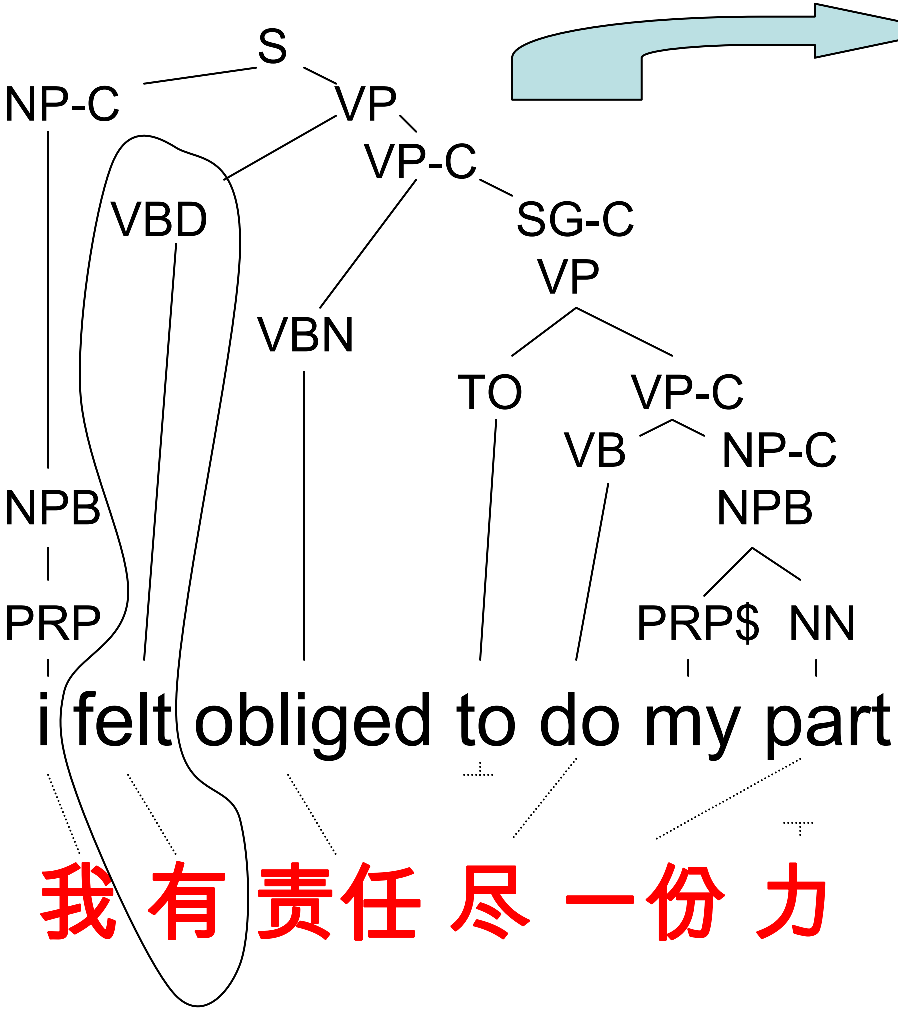
(Galley, Hopkins, Knight, Marcu, 2004)

RULES ACQUIRED:

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

Tree Transducers Can be Extracted from Data

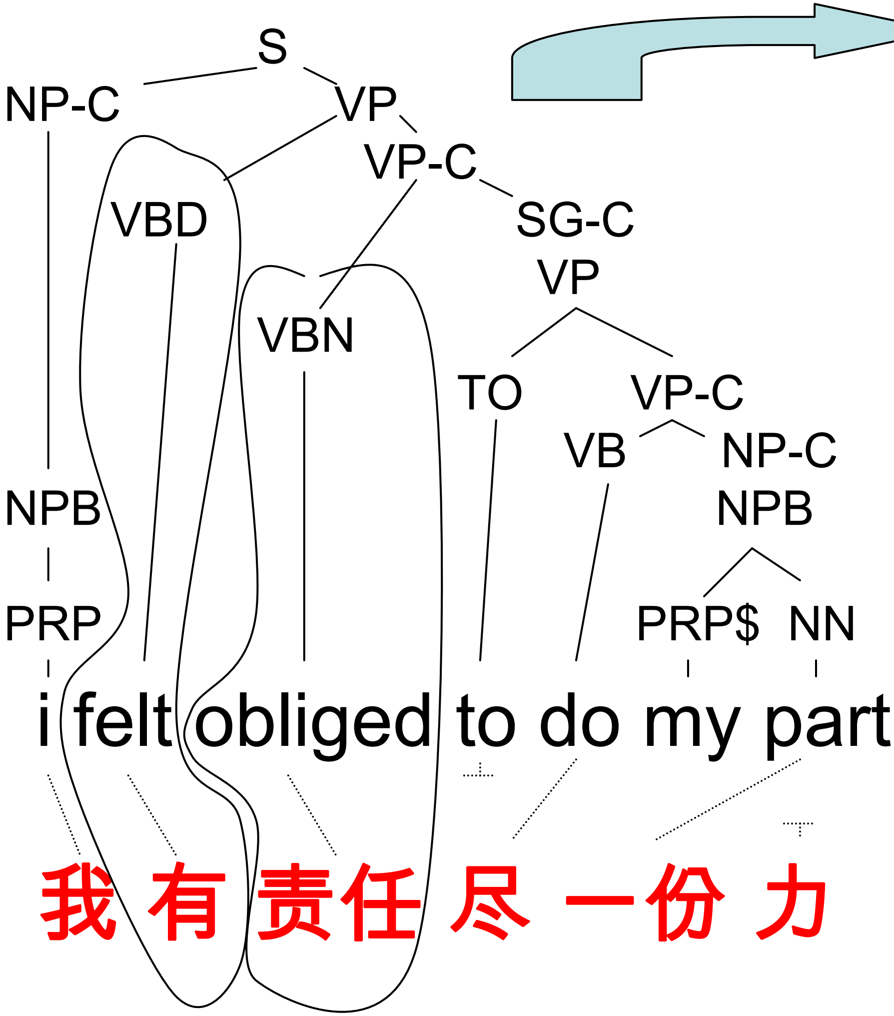
(Galley, Hopkins, Knight, Marcu, 2004)



RULES ACQUIRED:

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

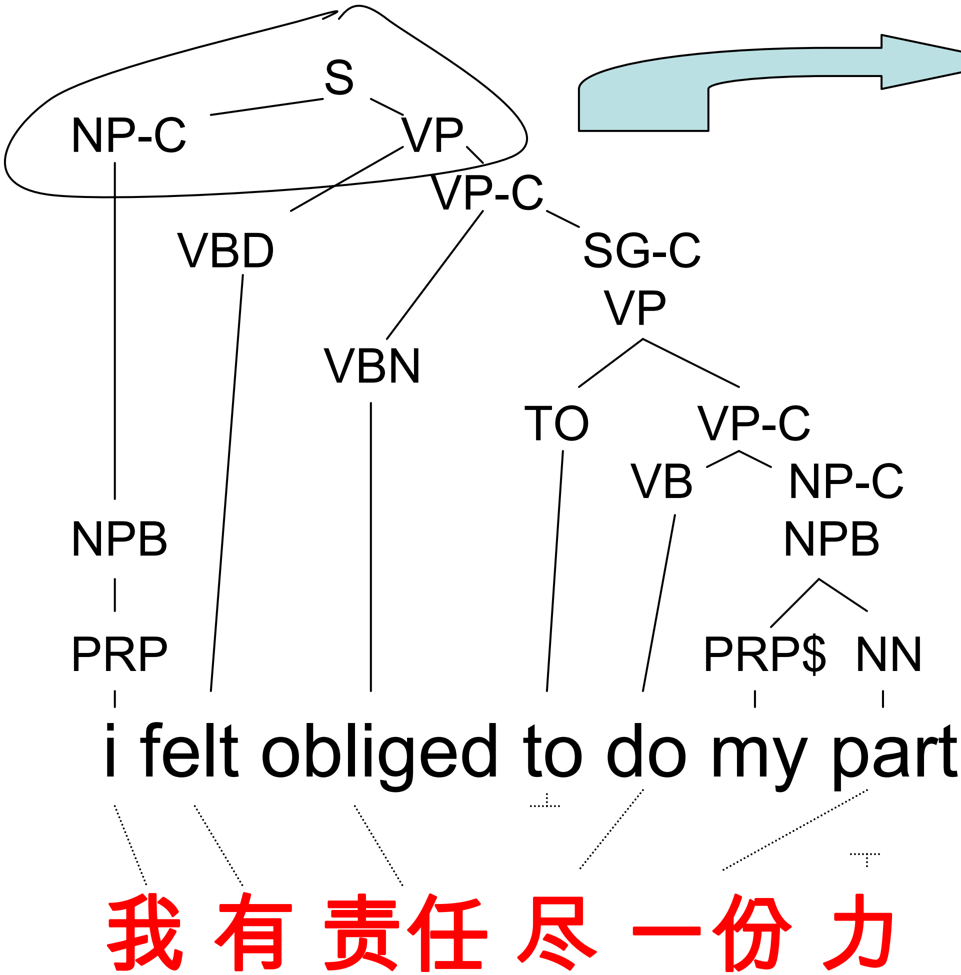
Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)



RULES ACQUIRED:

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

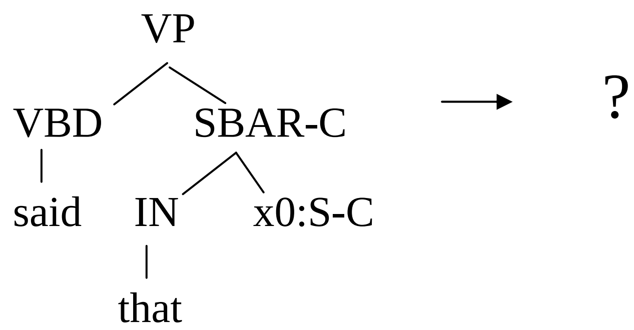
Tree Transducers Can be Extracted from Data
(Galley, Hopkins, Knight, Marcu, 2004)



RULES ACQUIRED:

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

Sample "said that" rules



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

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

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

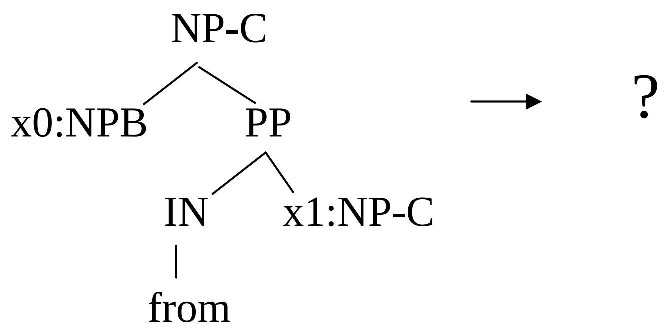
0.02 VP(VBD("said") SBAR-C(IN("that") x0:S-C)) -> 指出 , x0

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

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

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

Sample “NP-from-NP” rules



0.27 NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x1 x0

0.15 NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 来自 x1 x0

0.06 NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x1 的 x0

0.06 NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 从 x1 x0

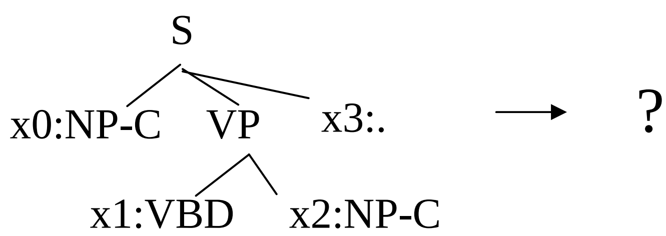
0.06 NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 来自 x1 的 x0

0.02 NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x0 从 x1

0.01 NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> 自 x1 x0

0.01 NP-C(x0:NPB PP(IN("from") x1:NP-C)) -> x1 x0 ，

Sample SVO rules



CHINESE / ENGLISH

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

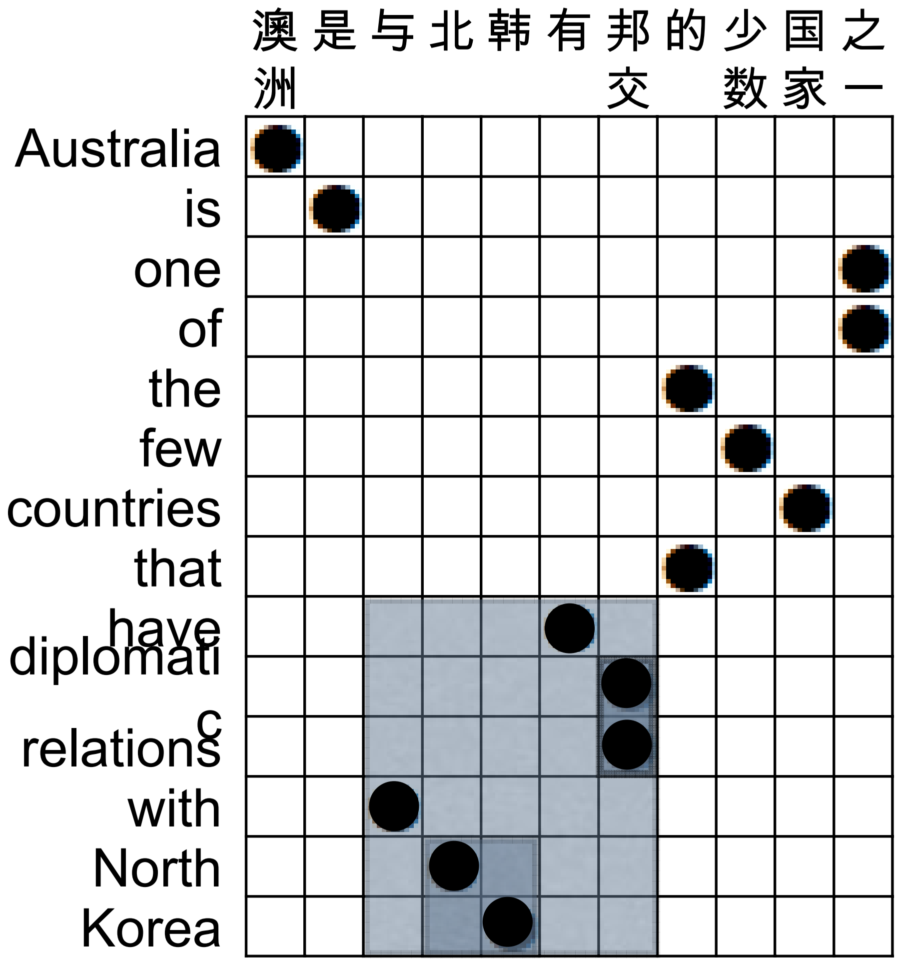
ARABIC / ENGLISH

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

Hiero (Chiang 05, 07)

- Phrase pairs with variables
  – e.g., “of X ←→ X de”
- Hierarchical decoding
  – the X itself could be created via other phrase pairs with variables
- Only one syntactic symbol in rules
  – X
- Translation patterns can be extracted without syntactically parses of the training data

Hiero Grammar Extraction



X(diplomatic relations)
→ 邦交

X(North Korea)
→ 北 韩

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

X(have x0:X with x1:X)
→ 与 x1 有 x0

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

Sample Hiero rules

(in tree transducer format)

X('s) → 的

X(the x0:X of x1:X) → x1 的 x0

X(the x0:X that x1:X) → x1 的 x0

X(in) → 在

X(under x0:X) → 在 x0 下

X(before x0:X) → 在 x0 前

X(x0:X this year) → 今年 x0

X(one of x0:X) → x0 之一

X(president x0:X) → x0 总统

Decoding
Reminder: phrase-based decoding

这	7人	中包括	来自	法国	和	俄罗斯	的	宇航	员	.
the	7 people	including	by some		**and**	the russian	**the**	the astronauts		,
it	7 people included		by france		and the	the russian		international astronautical	of rapporteur .	
this	7 out	including the	**from**	the french	and the russian		the fifth		.	
these	7 among	including from		the french and		of the russian	of	space	members	**.**
that	7 persons	including from the		of france	and to	russian	of the	aerospace	members	
	7 include		from the	of france and		russian		**astronauts**		. the
	7 numbers include		**from france**		and russian		of astronauts who			,
	7 populations include		those from france		and russian			astronauts .		
	7 deportees included		come from	**france**	**and russia**		in	astronautical	personnel	;
	7 philtrum	including those from		**france and**		**russia**	a space		**member**	
		including representatives from		france and the		**russia**		astronaut		
		include	came from	**france and russia**			by cosmonauts			
		include representatives from		french	**and russia**			cosmonauts		
		include	came from france		and russia 's			cosmonauts .		
		includes	coming from	french and		russia 's		cosmonaut		
				french and russian			's	astronavigation	member .	
				french	**and russia**		**astronauts**			
					and russia 's				special rapporteur	
					, and	**russia**			rapporteur	
					, and russia				rapporteur .	
					, and russia					
					or	russia 's				

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

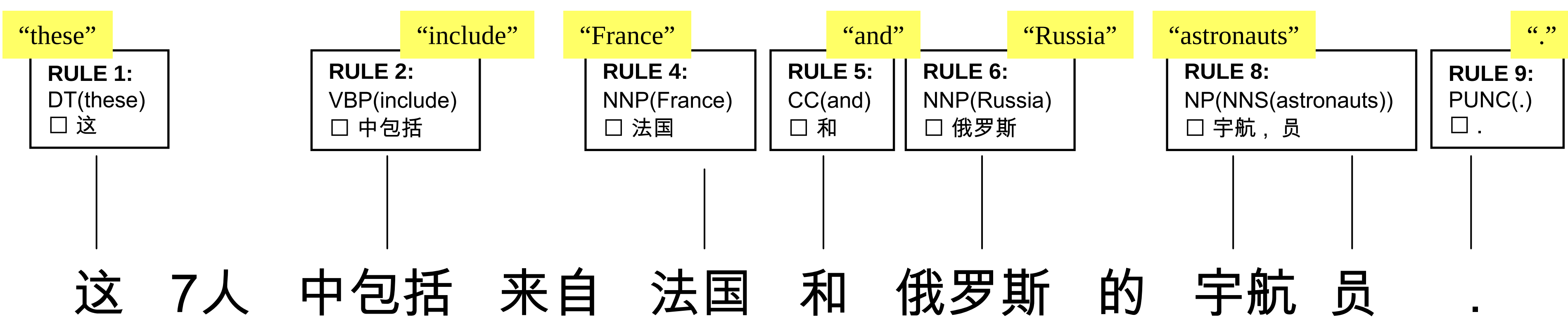
Scoring: Try to use phrase pairs that have been frequently observed.
Try to output a sentence with frequent English word sequences.

Syntax-Based Decoding

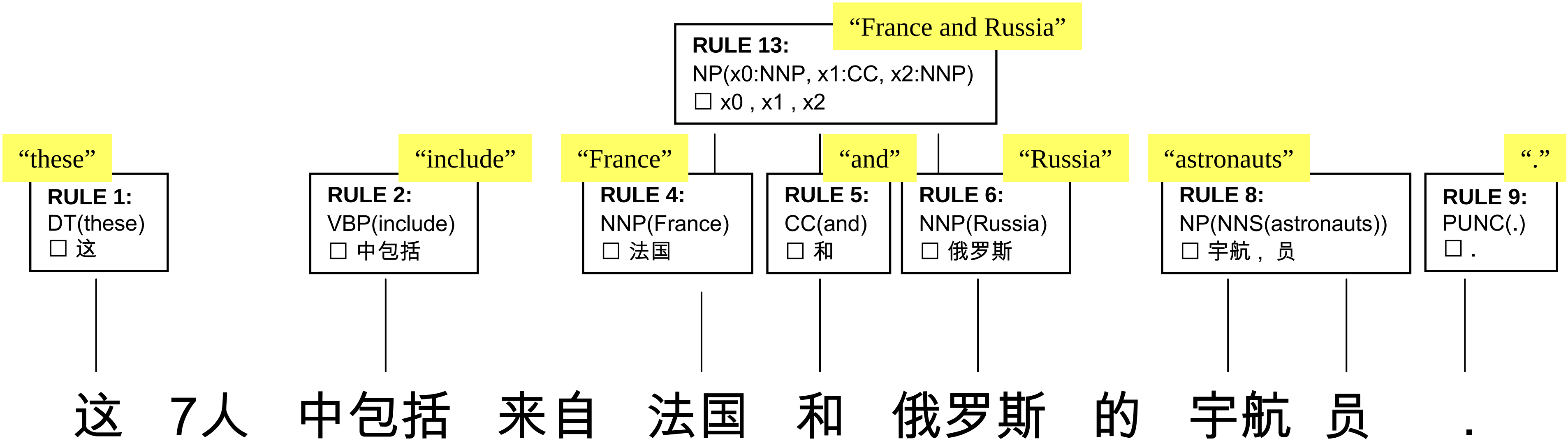
- Bottom-up CKY parser
- Builds English constituents on top of Chinese spans
- Record of rule applications (the derivation) provides information to construct English tree
- Returns k-best trees
- Same decoder can handle syntax translation rules and Hiero rules

Syntax-Based Decoding

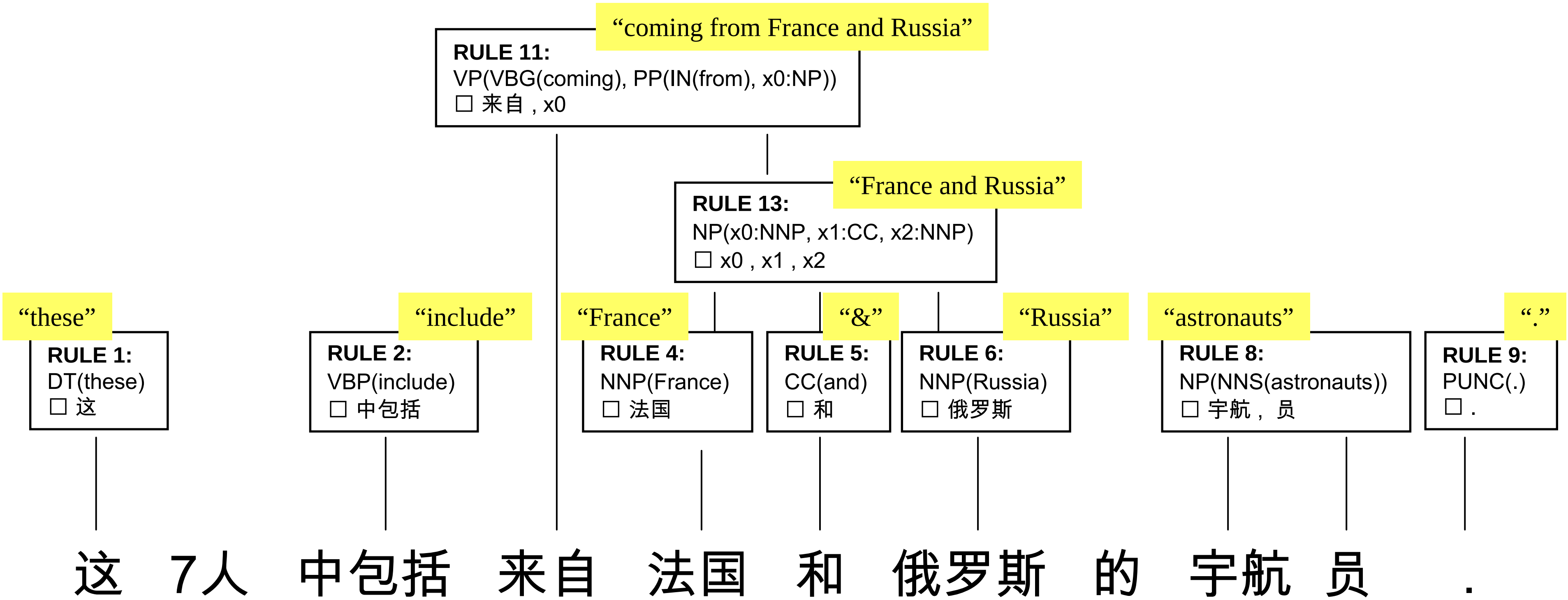
Rules apply when their right-hand sides (RHS) match some portion of the input.



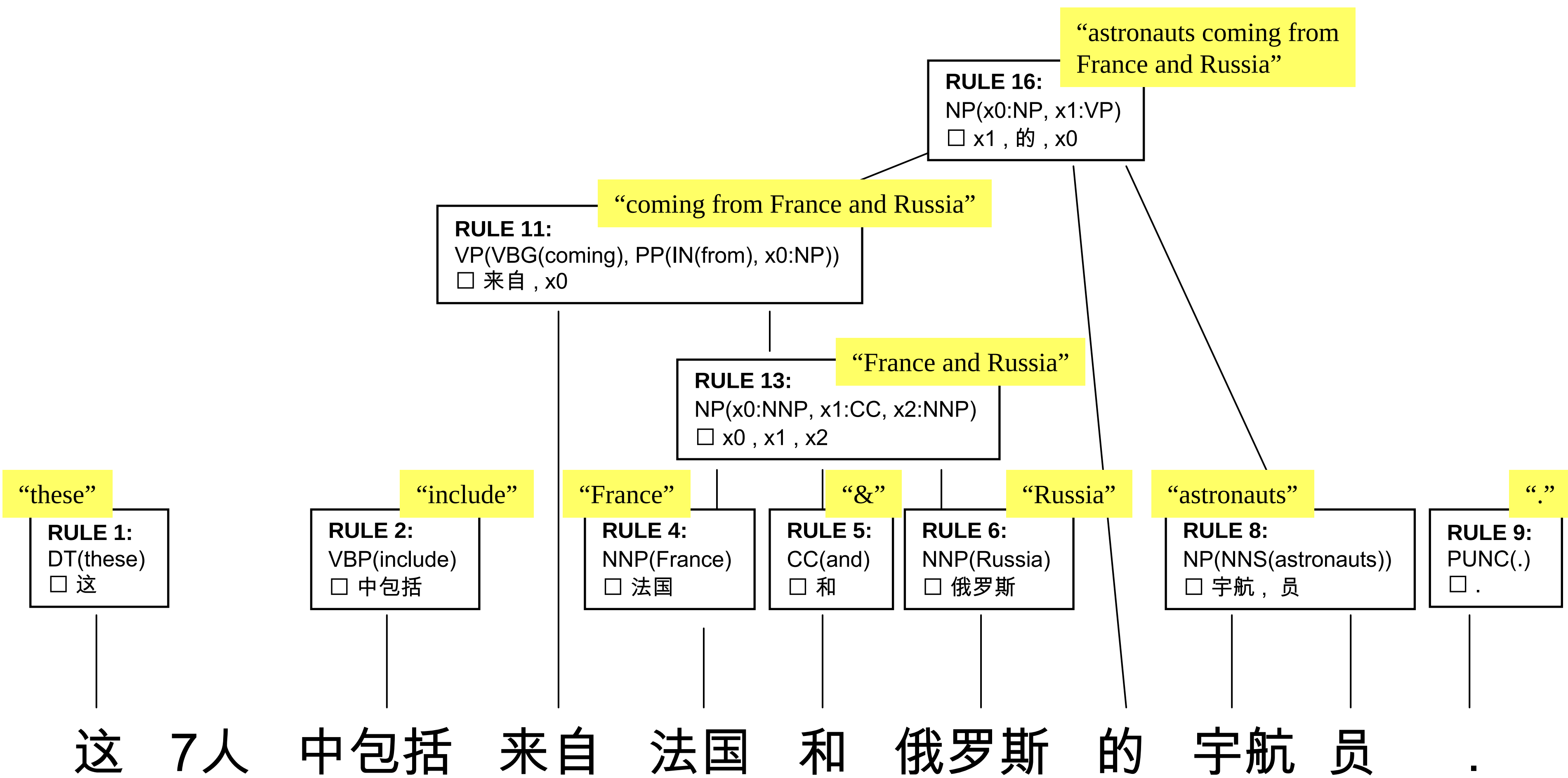
Syntax-Based Decoding

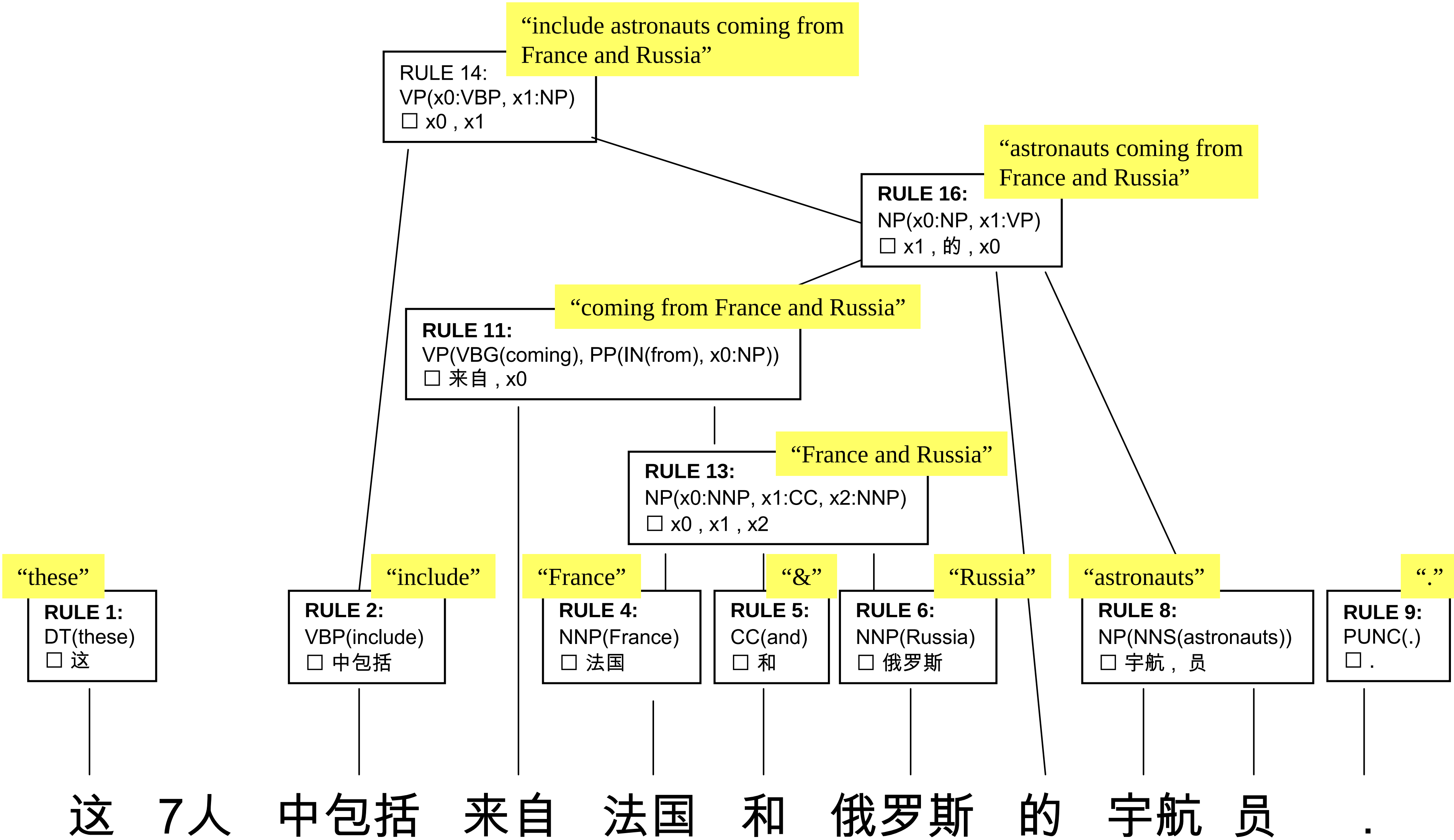


Syntax-Based Decoding



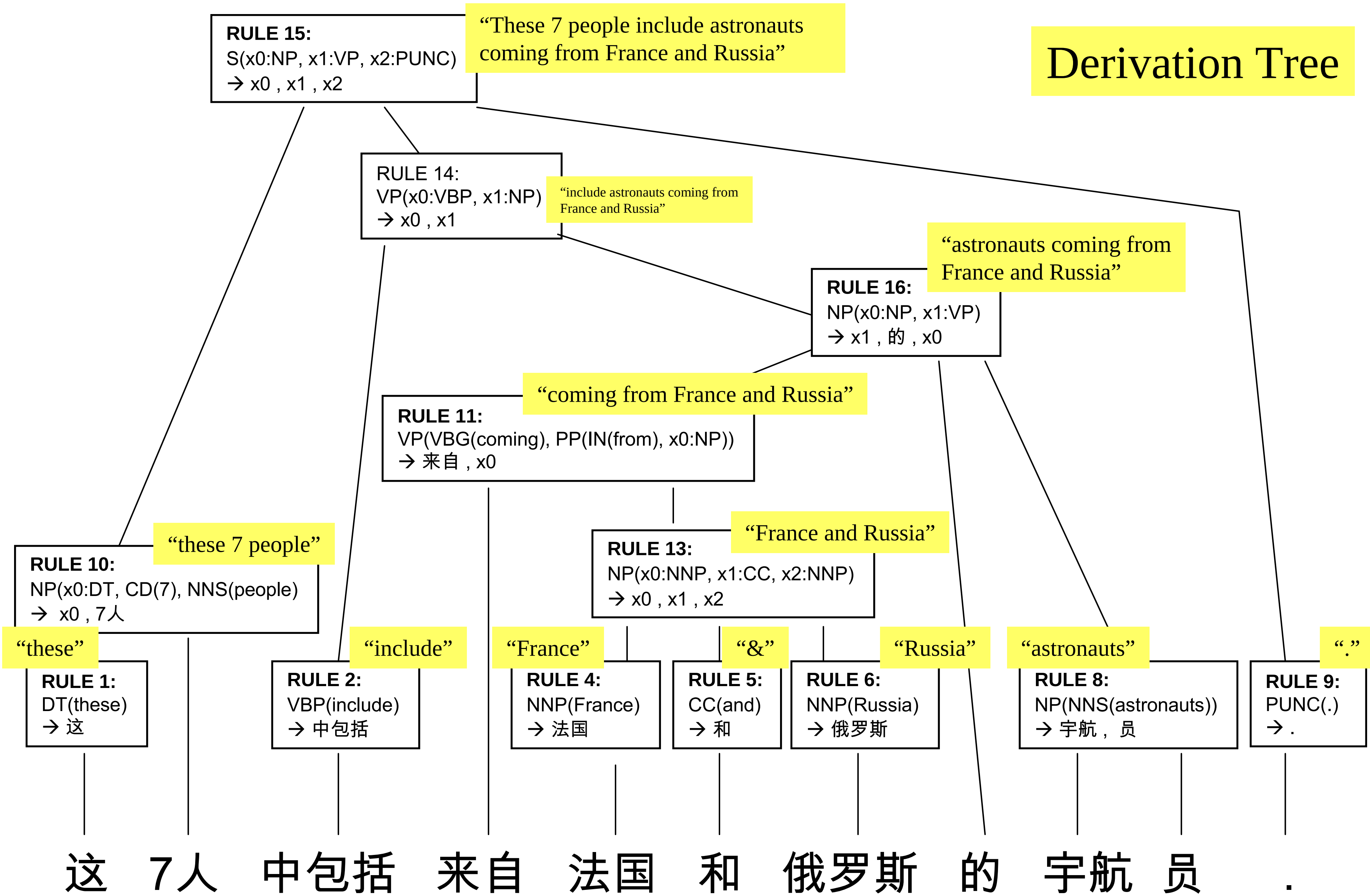
Syntax-Based Decoding



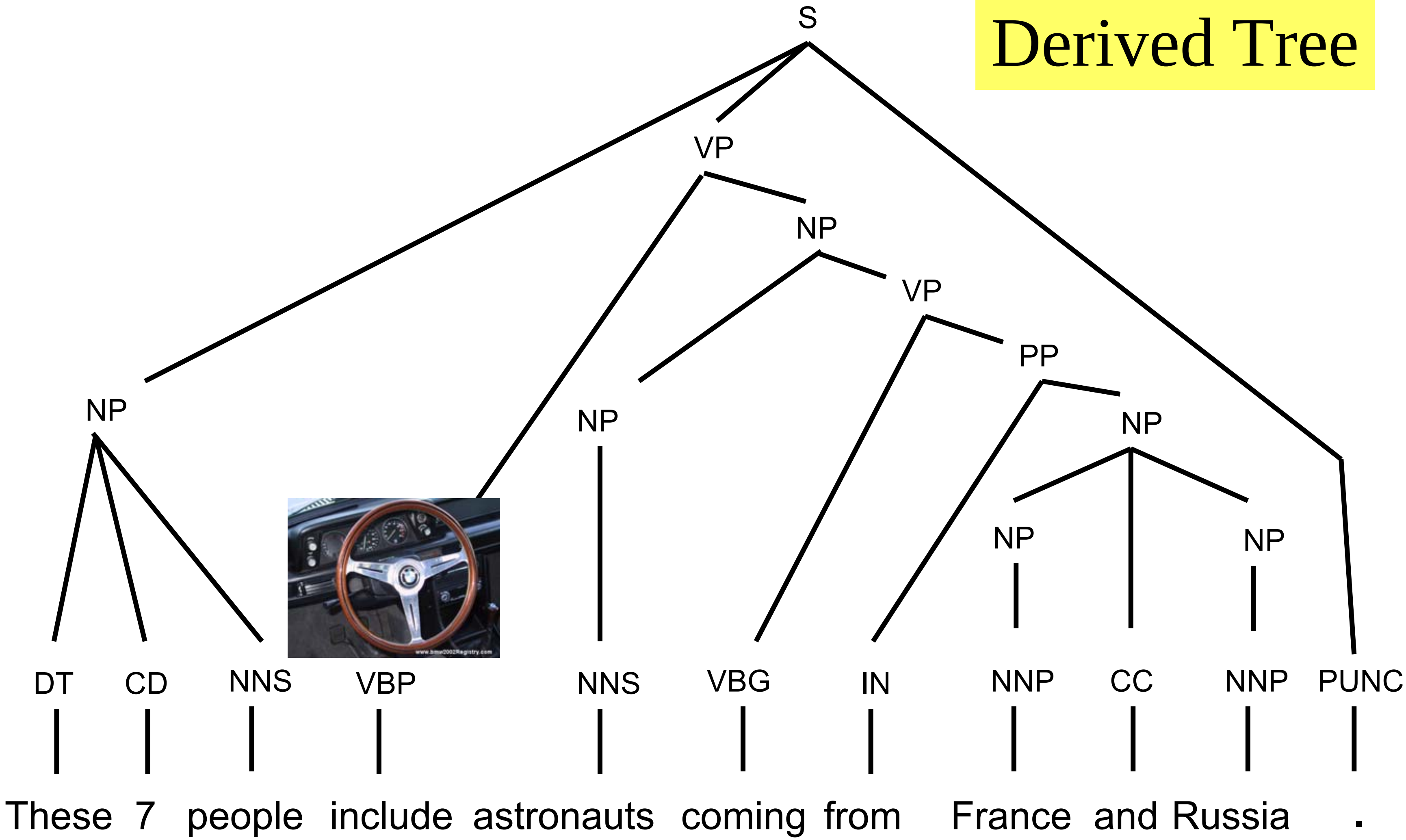


"include astronauts coming from France and Russia"
RULE 14:
VP(x0:VBP, x1:NP)
□ x0 , x1
"astronauts coming from France and Russia"
RULE 16:
NP(x0:NP, x1:VP)
□ x1 , 的 , x0
"coming from France and Russia"
RULE 11:
VP(VBG(coming), PP(IN(from), x0:NP))
□ 来自 , x0
"France and Russia"
RULE 13:
NP(x0:NNP, x1:CC,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

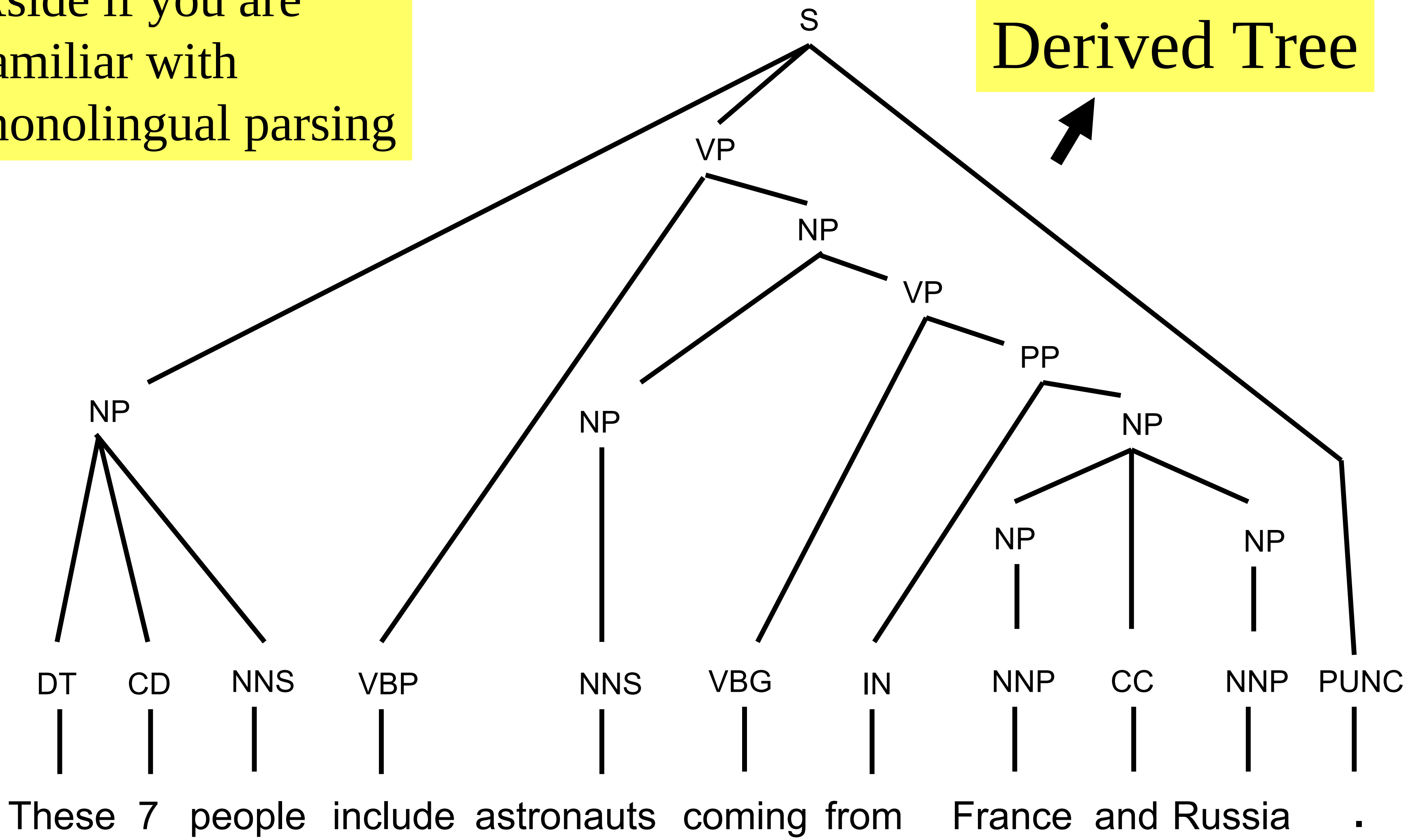


Derived Tree



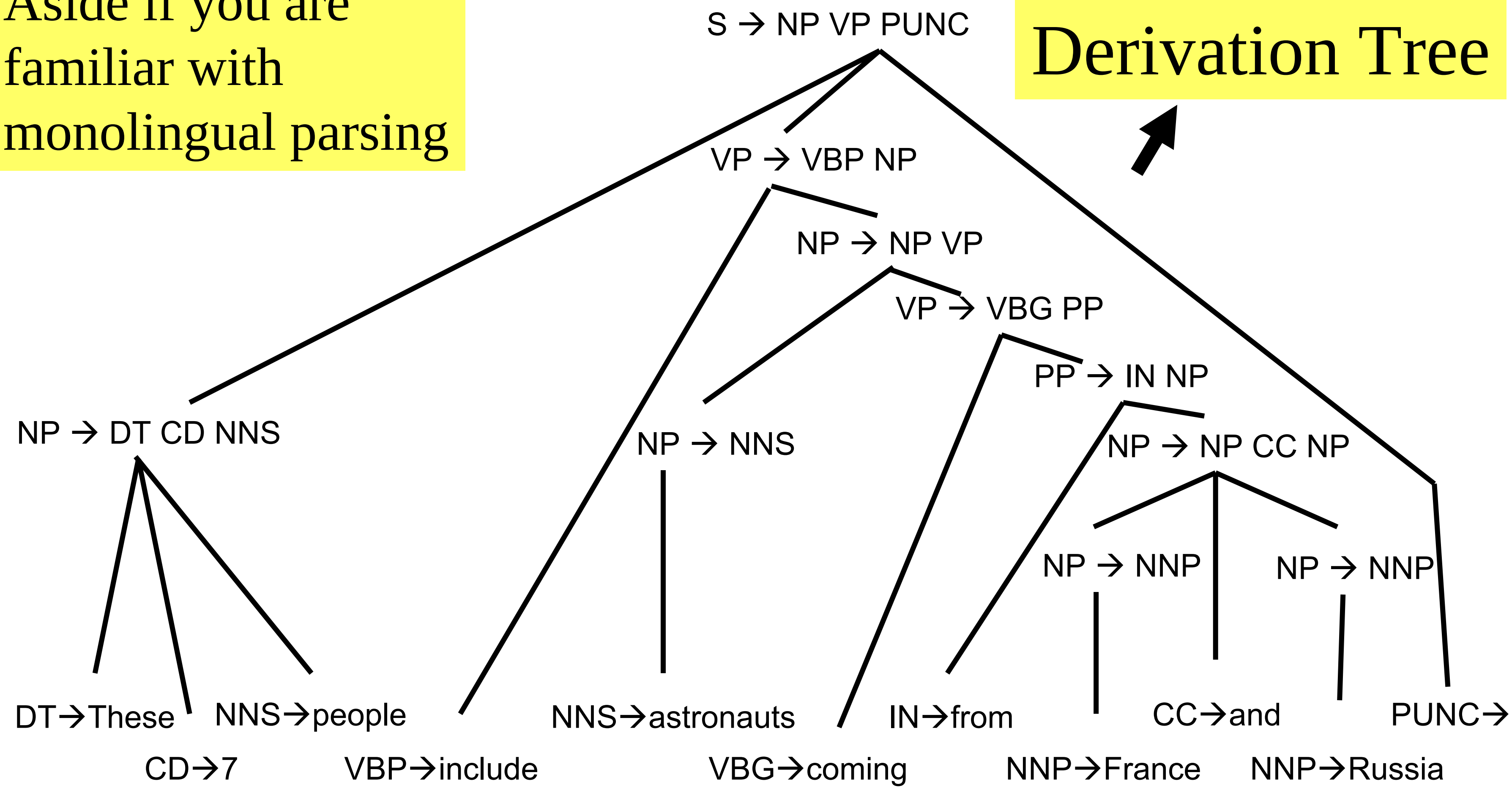
Aside if you are familiar with monolingual parsing

Derived Tree



Aside if you are familiar with monolingual parsing

Derivation Tree



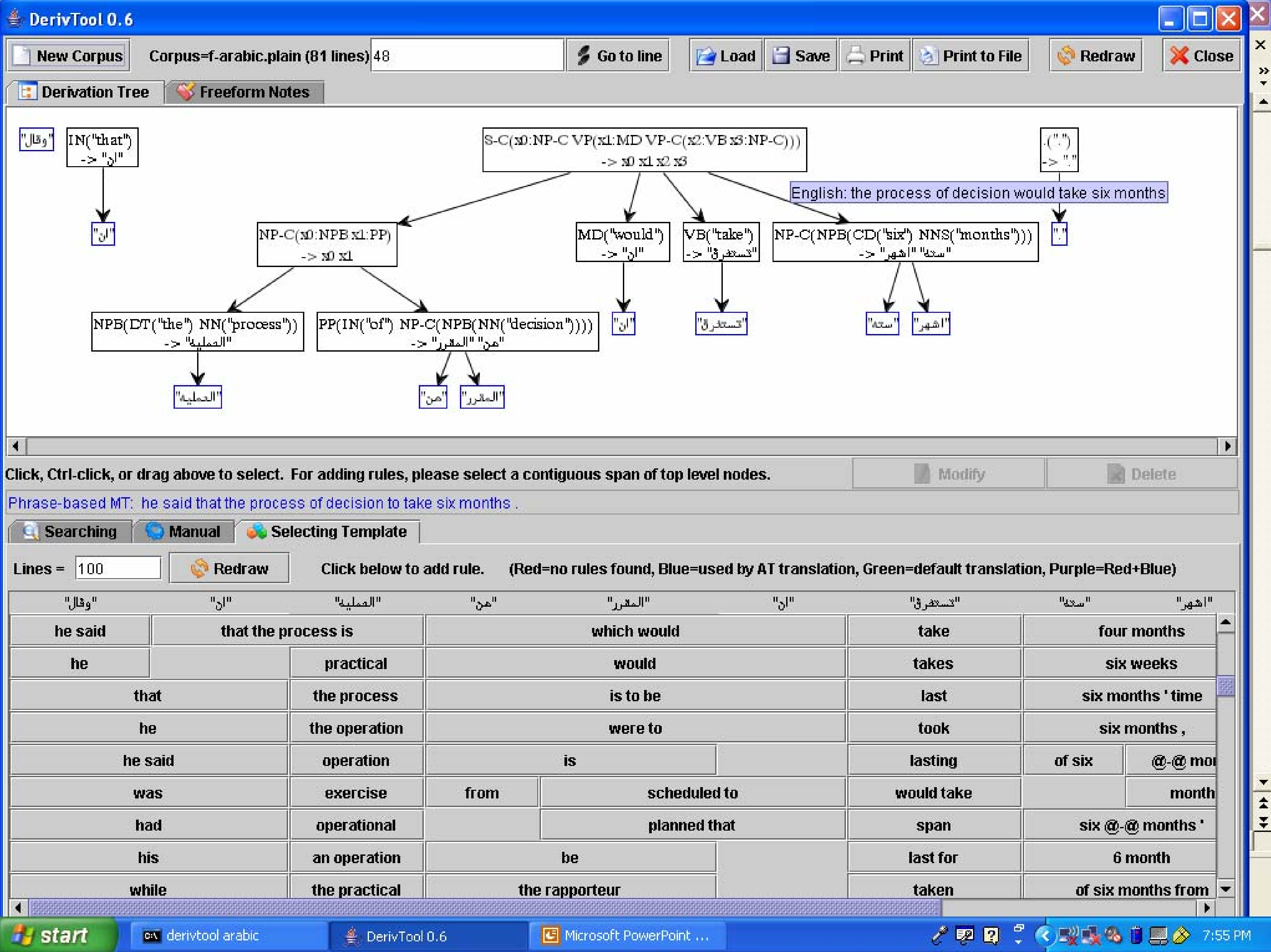
Binarization for Decoding

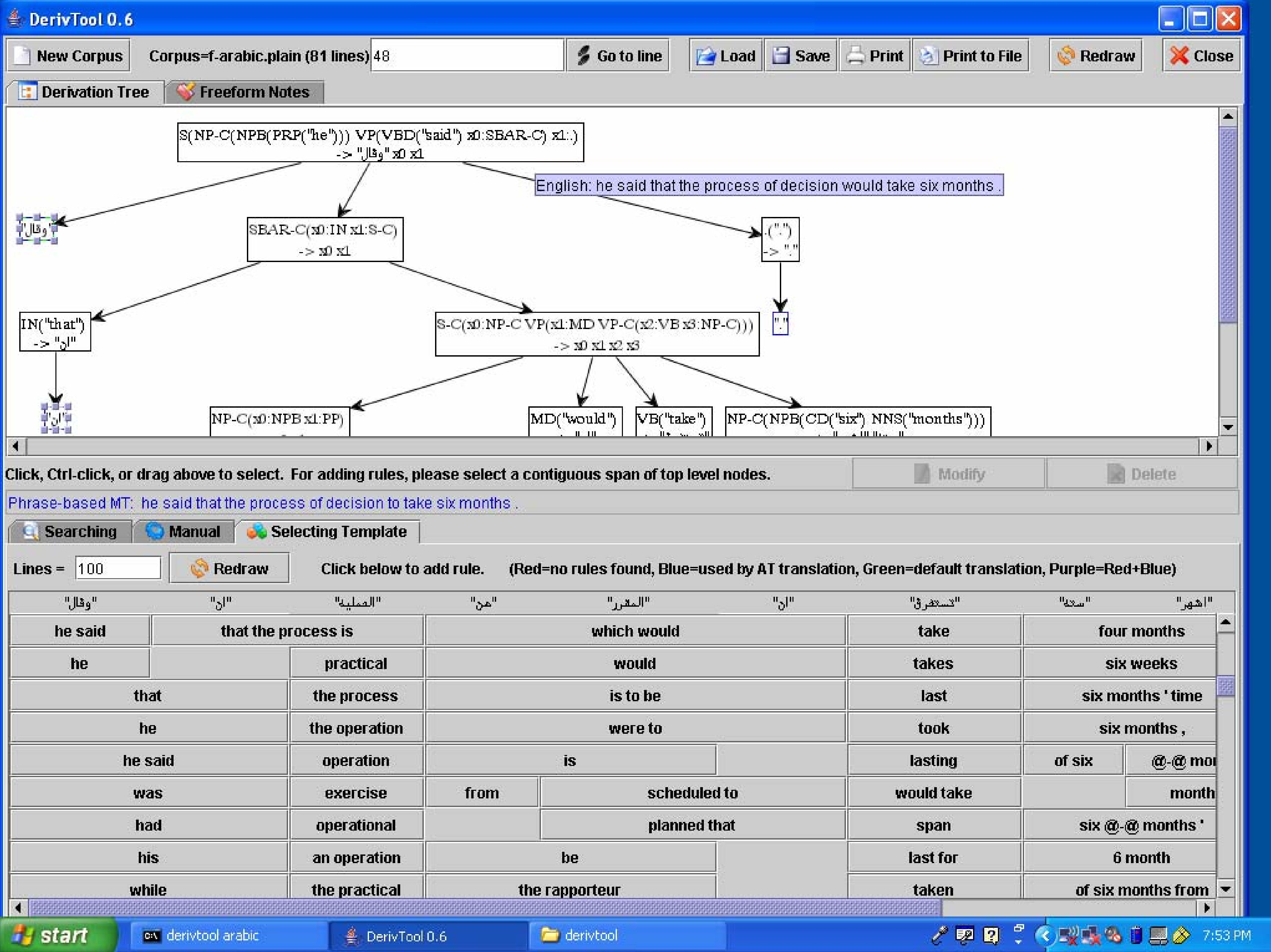
- For CKY decoding, all rules must be *binarized.*
- Rule with |RHS| > 2 must be split into rules with |RHS| = 2
  - S(x0:NP VP(x1:VBD x2:NP)) → x1 x0 x2

    Z(x0:NP x1:VBD) → x1 x0

    S(x0:Z x1:NP) → x0 x1
- Similar to putting a CFG into Chomsky normal form.
- A rule can be binarized in different ways: must pick best!
- Some translation rules cannot be binarized at all…
  - A(x0:B x1:C x2:D x3:E) → x1 x3 x0 x2 [Wu 96]
- We just delete these.

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


DerivTool 0.6
New Corpus
Corpus=f-arabic.plain (81 lines) 48
Go to line
Load
Save
Print
Print to File
Redraw
Close
Derivation Tree
Freeform Notes
"وقال"
IN("that") -> "ان"
"ان"
S-C(x0:NP-C VP(x1:MD VP-C(x2:VB x3:NP-C))) -> x0 x1 x2 x3
.(".") -> "."
"."
English: the process of decision would take six months
NP-C(x0:NPB x1:PP) -> x0 x1
MD("would") -> "ان"
VB("take") -> "تستغرق"
NP-C(NPB(CD("six") NNS("months"))) -> "ستة" "اشهر"
NPB(DT("the") NN("process")) -> "العملية"
PP(IN("of") NP-C(NPB(NN("decision")))) -> "من" "المقرر"
"ان"
"تستغرق"
"ستة"
"اشهر"
"العملية"
"من"
"المقرر"
Click, Ctrl-click, or drag above to select. For adding rules, please select a contiguous span of top level nodes.
Modify
Delete
Phrase-based MT: he said that the process of decision to take six months .
Searching
Manual
Selecting Template
Lines = 100
Redraw
Click below to add rule.
(Red=no rules found, Blue=used by AT translation, Green=default translation, Purple=Red+Blue)
"وقال" "ان" "العملية" "من" "المقرر" "ان" "تستغرق" "ستة" "اشهر"
he said | that the process is | which would | take | four months
he | practical | would | takes | six weeks
that | the process | is to be | last | six months ' time
he | the operation | were to | took | six months ,
he said | operation | is | lasting | of six | @-@ mo
was | exercise | from | scheduled to | would take | month
had | operational | planned that | span | six @-@ months '
his | an operation | be | last for | 6 month
while | the practical | the rapporteur | taken | of six months from
start
derivtool arabic
DerivTool 0.6
Microsoft PowerPoint ...
7:55 PM


DerivTool 0.6
New Corpus
Corpus=f-arabic.plain (81 lines) 48
Go to line
Load
Save
Print
Print to File
Redraw
Close
Derivation Tree
Freeform Notes
S(NP-C(NPB(PRP("he"))) VP(VBD("said") x0:SBAR-C) x1:.)
-> "وقال" x0 x1
English: he said that the process of decision would take six months .
"وقال"
SBAR-C(x0:IN x1:S-C)
-> x0 x1
.(".")
-> "."
"."
IN("that")
-> "أن"
S-C(x0:NP-C VP(x1:MD VP-C(x2:VB x3:NP-C)))
-> x0 x1 x2 x3
"أن"
NP-C(x0:NPB x1:PP)
MD("would")
VB("take")
NP-C(NPB(CD("six") NNS("months")))
Click, Ctrl-click, or drag above to select. For adding rules, please select a contiguous span of top level nodes.
Modify
Delete
Phrase-based MT: he said that the process of decision to take six months .
Searching
Manual
Selecting Template
Lines = 100
Redraw
Click below to add rule.
(Red=no rules found, Blue=used by AT translation, Green=default translation, Purple=Red+Blue)
"وقال" "أن" "العملية" "من" "المقرر" "أن" "تستغرق" "ستة" "أشهر"
he said
that the process is
which would
take
four months
he
practical
would
takes
six weeks
that
the process
is to be
last
six months ' time
he
the operation
were to
took
six months ,
he said
operation
is
lasting
of six
was
exercise
from
scheduled to
would take
month
had
operational
planned that
span
six @-@ months '
his
an operation
be
last for
6 month
while
the practical
the rapporteur
taken
of six months from
start
derivtool arabic
DerivTool 0.6
derivtool
7:53 PM

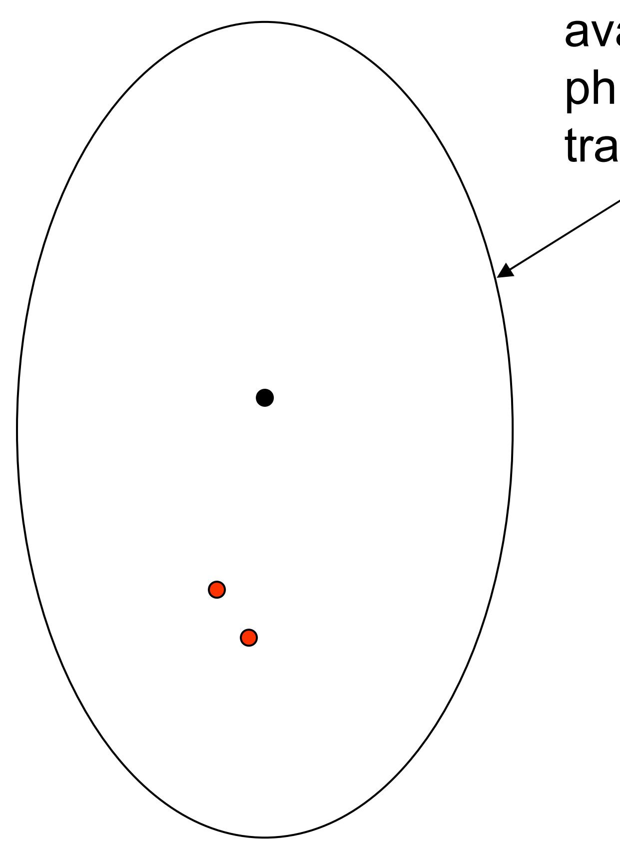
Why Might Syntax Help?

- Phrase-based MT output is “n-grammatical”, not grammatical
  - Every sentence needs a subject and a verb
- Re-ordering is poorly explained as “distortion” -- better explained as syntactic transformation
  - Arabic to English, VSO → SVO
- Function words have syntactic effects even if they are not themselves translated

Why Might Syntax Hurt?

- Less freedom to glue pieces of output together -- search space has fewer output strings
- Search space is more difficult to navigate
- Rule extraction from bilingual text has limitations

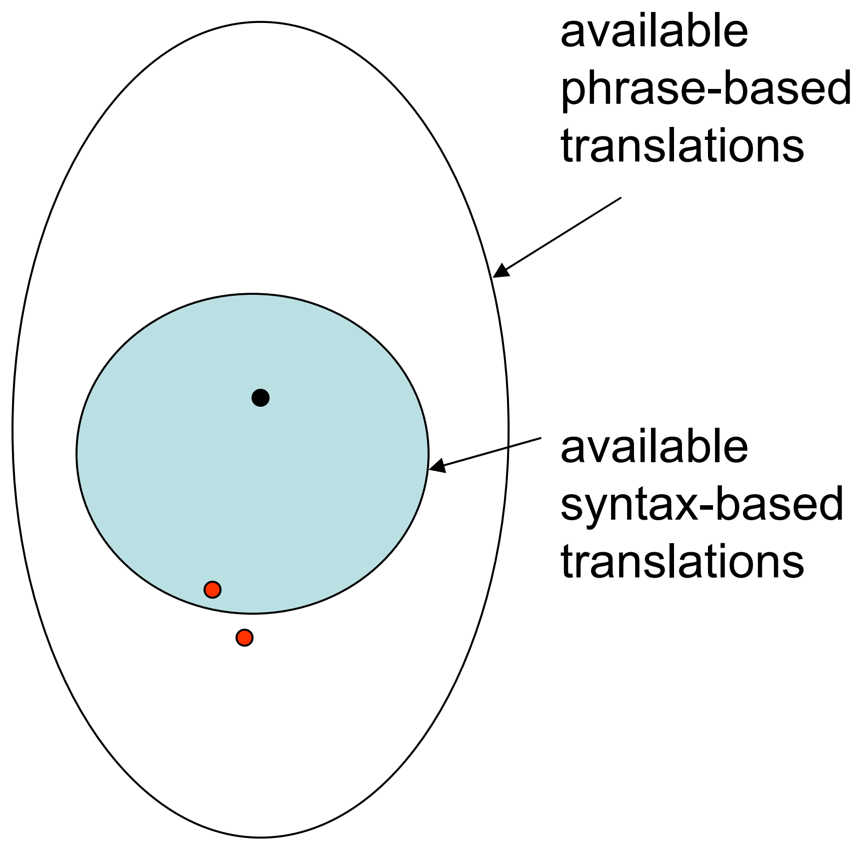
this section



Why Might Syntax Hurt?

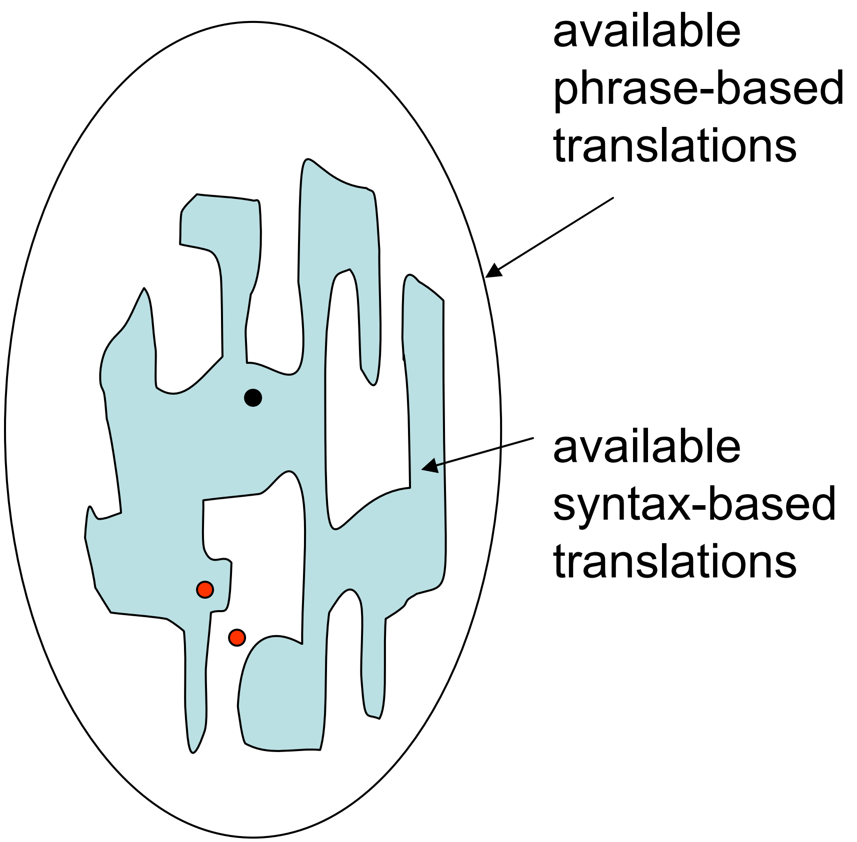
- Less freedom to glue pieces of output together -- search space has fewer output strings
- Search space is more difficult to navigate
- Rule extraction from bilingual text has limitations

this section



Why Might Syntax Hurt?

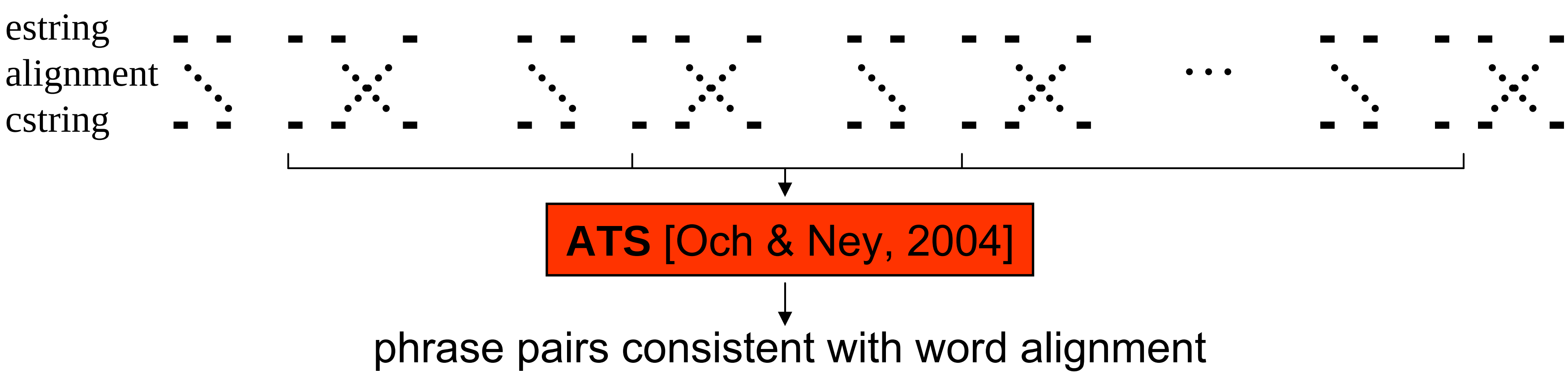
- Less freedom to glue pieces of output together -- search space has fewer output strings
- Search space is more difficult to navigate
- Rule extraction from bilingual text has limitations

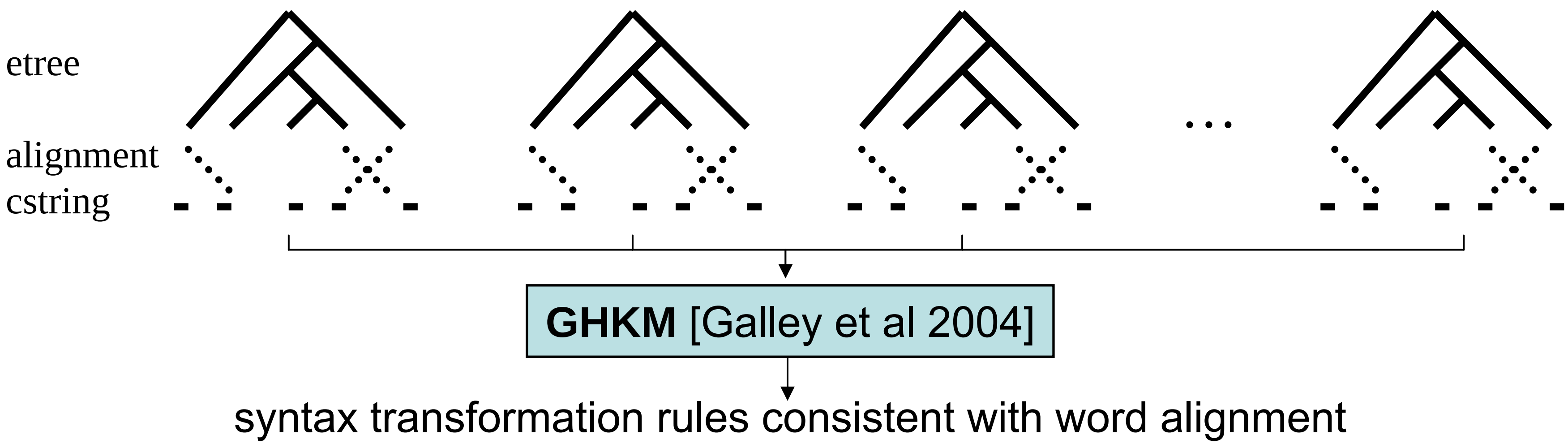


Comparing Phrase-Based Extraction with Syntax-Based Extraction

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

Phrase-Based and Syntax-Based Pattern Extraction





ATS (Och & Ney, 2004)

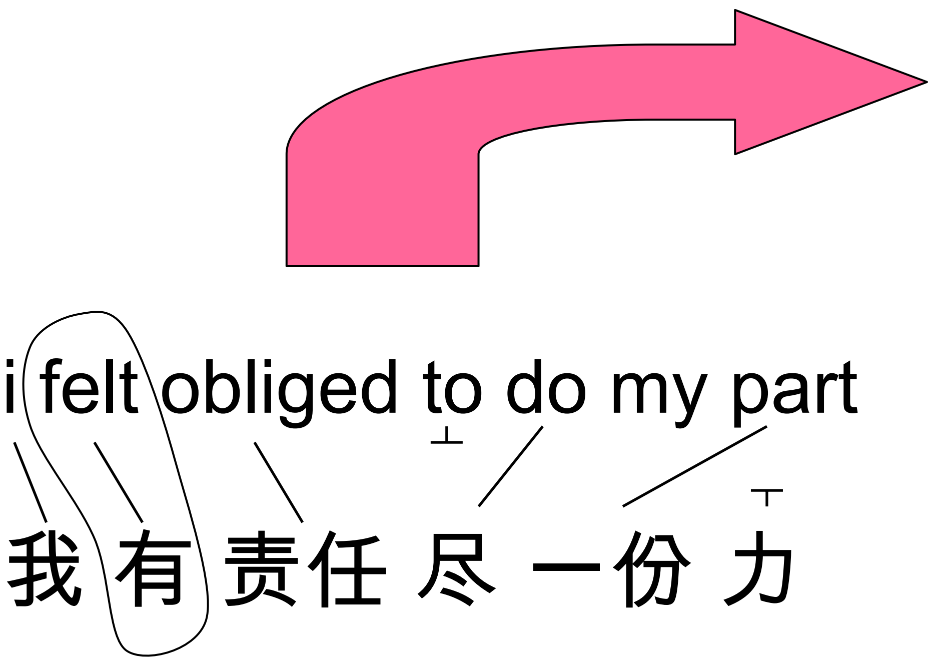
i felt obliged to do my part

我 有 责任 尽 一份 力

PHRASE PAIRS ACQUIRED:

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

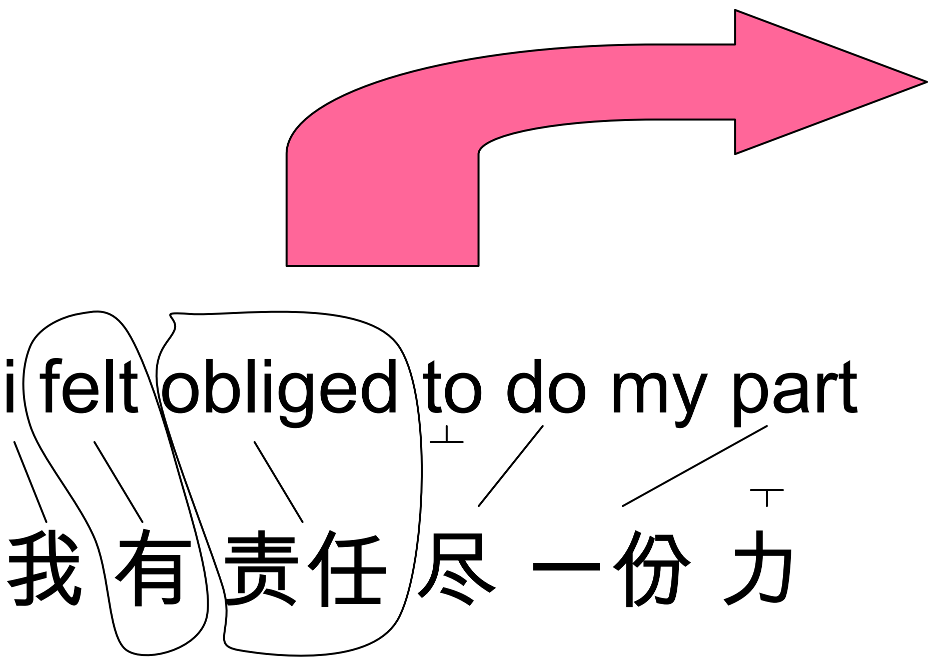
ATS (Och & Ney, 2004)



PHRASE PAIRS ACQUIRED:

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

ATS (Och & Ney, 2004)



PHRASE PAIRS ACQUIRED:

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

ATS (Och & Ney, 2004)

PHRASE PAIRS ACQUIRED:

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

GHKM (Galley et al, 2004)

RULES ACQUIRED:

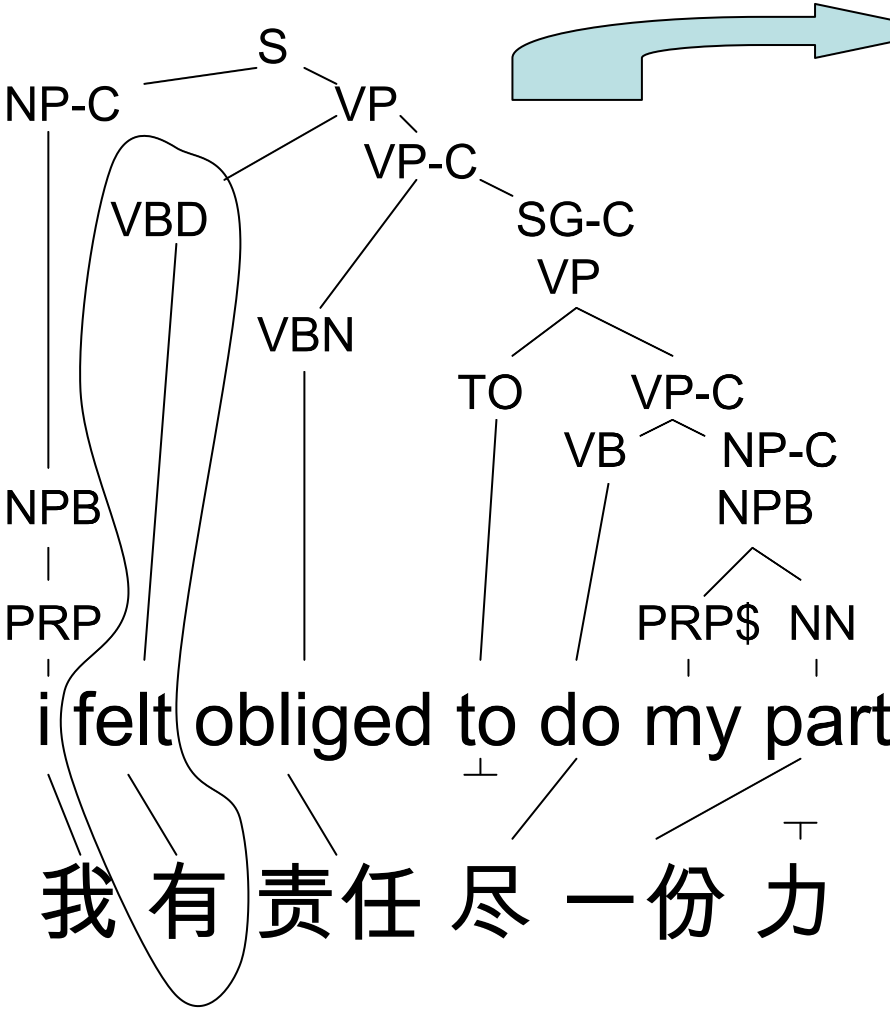
VBD(felt) → 有

VBN(obliged) → 责任

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

S(x0:NP-C x1:VP) → x0 x1

GHKM (Galley et al, 2004)



RULES ACQUIRED:

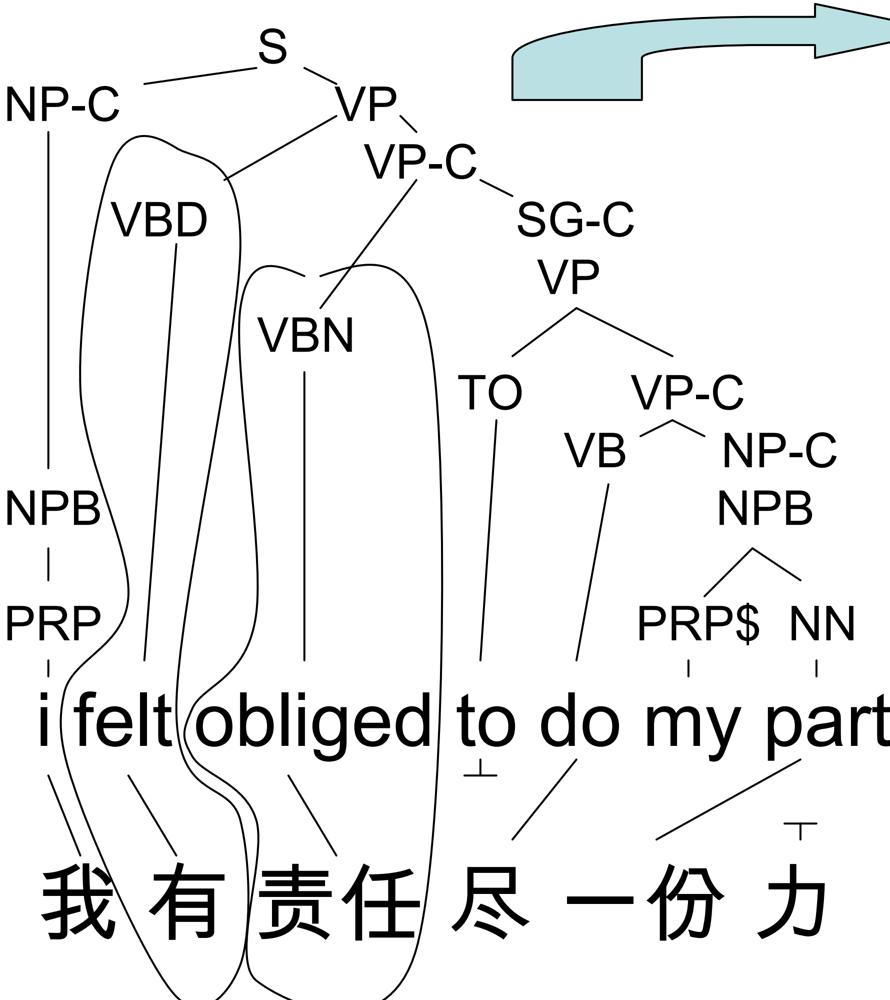
VBD(felt) → 有

VBN(obliged) → 责任

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

S(x0:NP-C x1:VP) → x0 x1

GHKM (Galley et al, 2004)



RULES ACQUIRED:

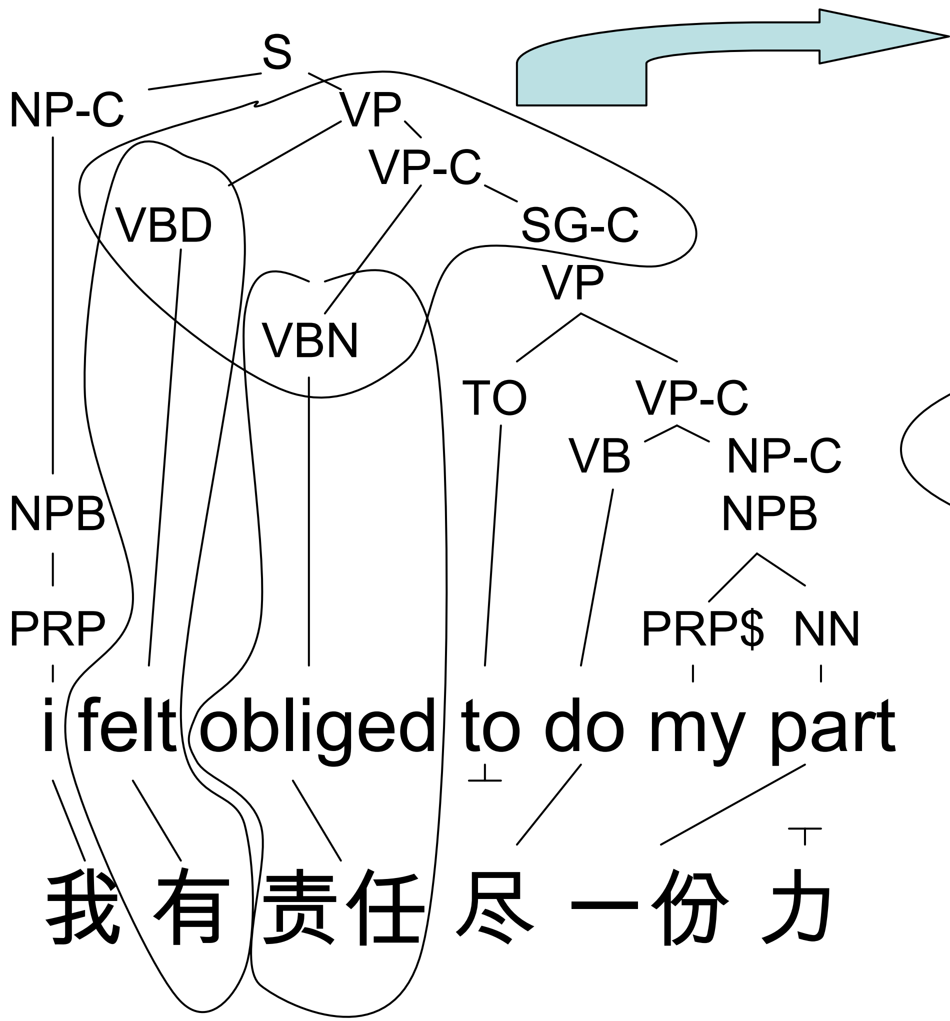
VBD(felt) → 有

VBN(obliged) → 责任

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

S(x0:NP-C x1:VP) → x0 x1

GHKM (Galley et al, 2004)



RULES ACQUIRED:

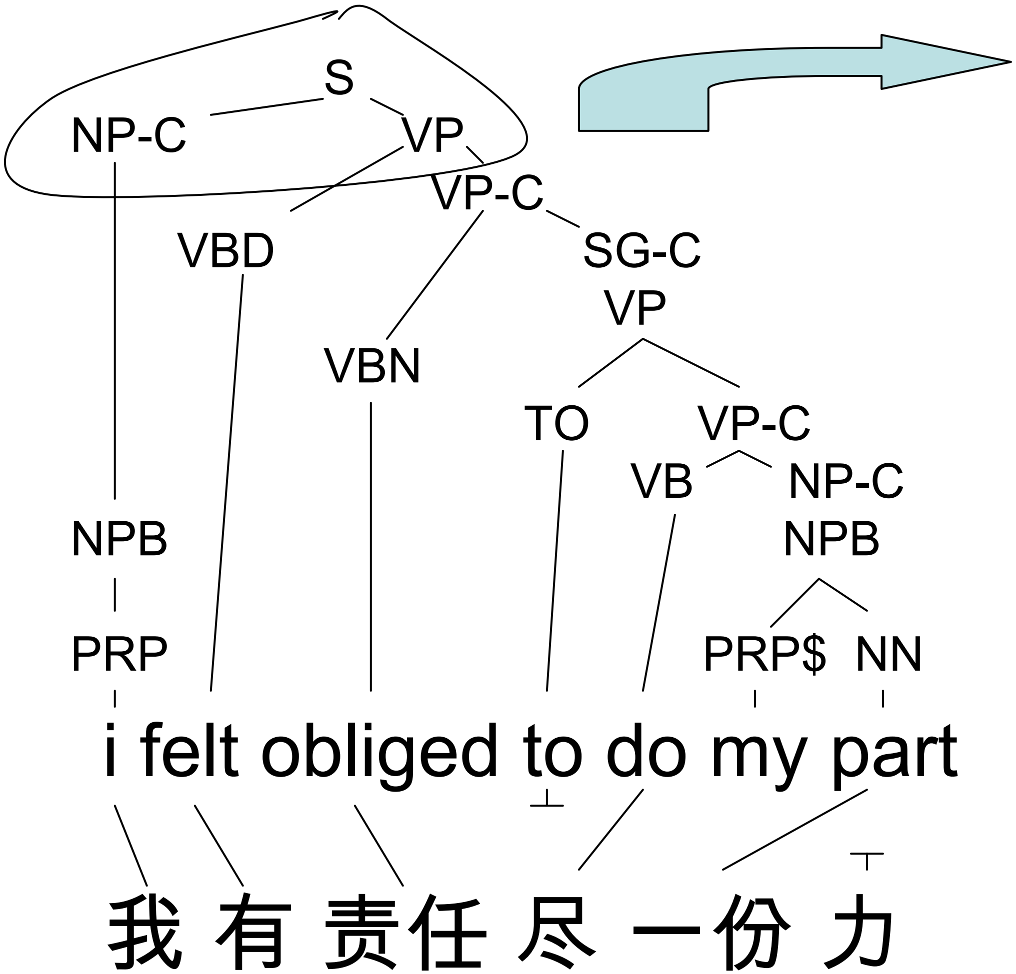
VBD(felt) → 有

VBN(obliged) → 责任

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

S(x0:NP-C x1:VP) → x0 x1

GHKM (Galley et al, 2004)



RULES ACQUIRED:

VBD(felt) → 有

VBN(obliged) → 责任

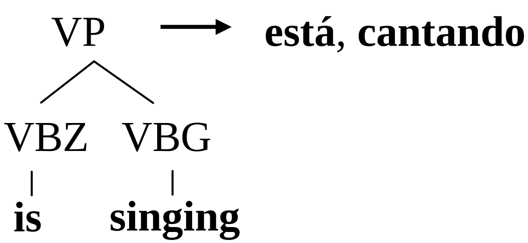
VP(x0:VBD
VP-C(x1:VBN
x2:SG-C) → x0 x1 x2

S(x0:NP-C x1:VP) → x0 x1

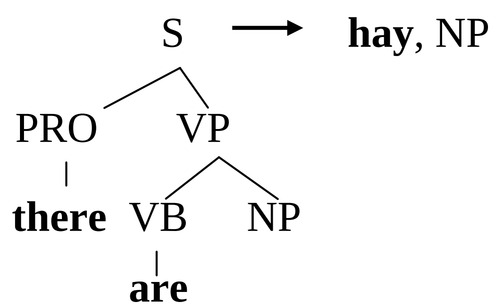
There is a unique tiling that identifies minimal translation units.

GHKM Syntax Rules

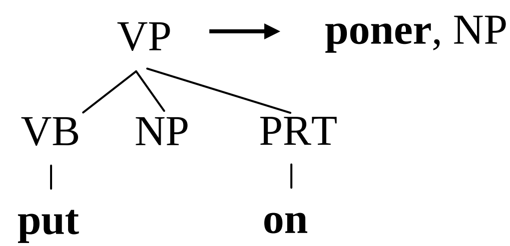
Phrasal Translation



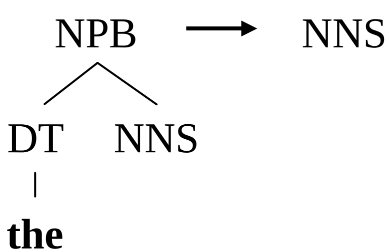
Non-constituent Phrases



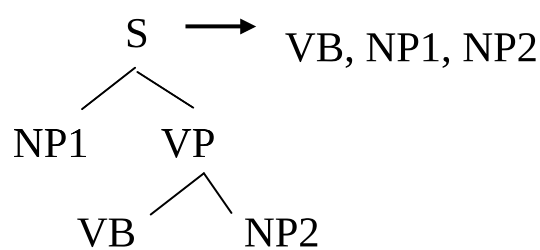
Non-contiguous Phrases



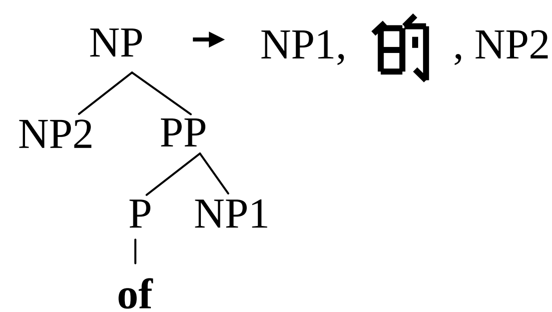
Context-Sensitive Word Insertion



Multilevel Re-Ordering

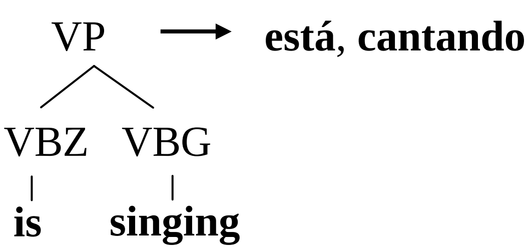


Lexicalized Re-Ordering

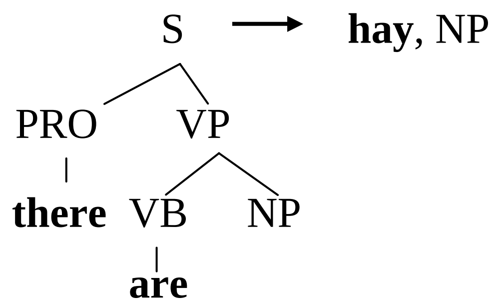


GHKM Syntax Rules

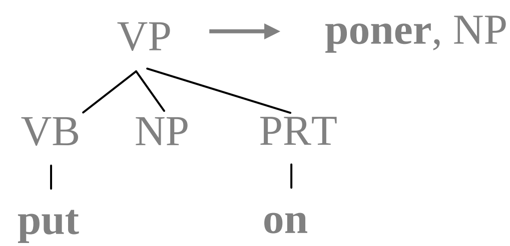
Phrasal Translation



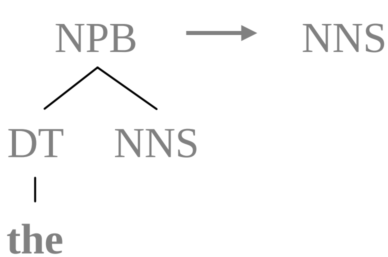
Non-constituent Phrases



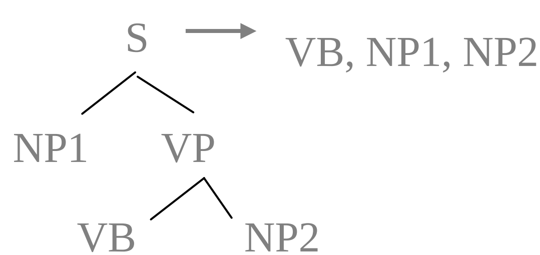
Non-contiguous Phrases



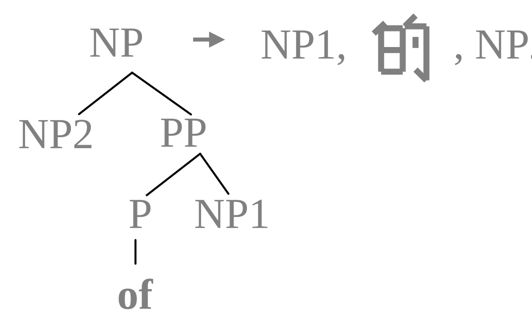
Context-Sensitive Word Insertion



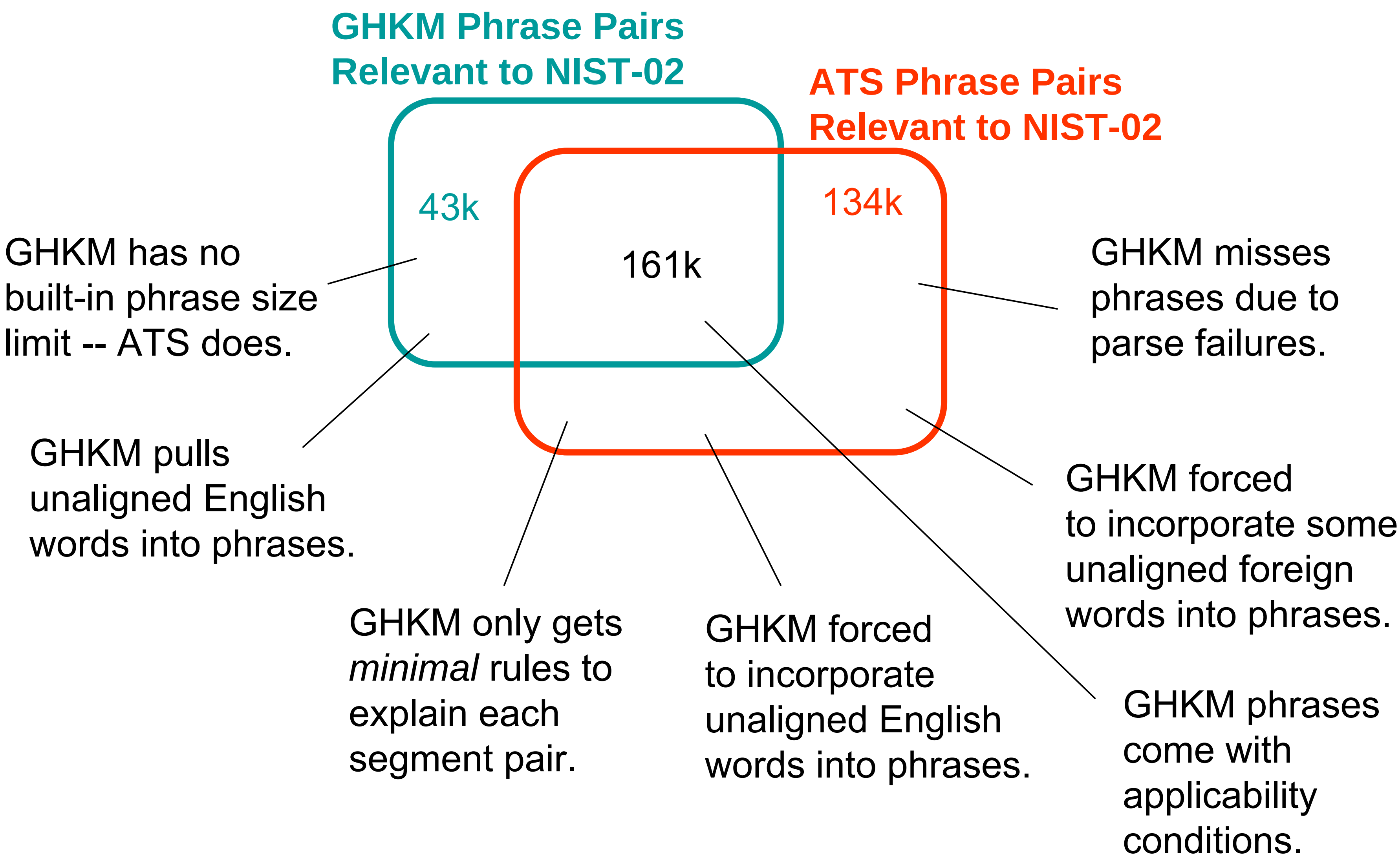
Multilevel Re-Ordering



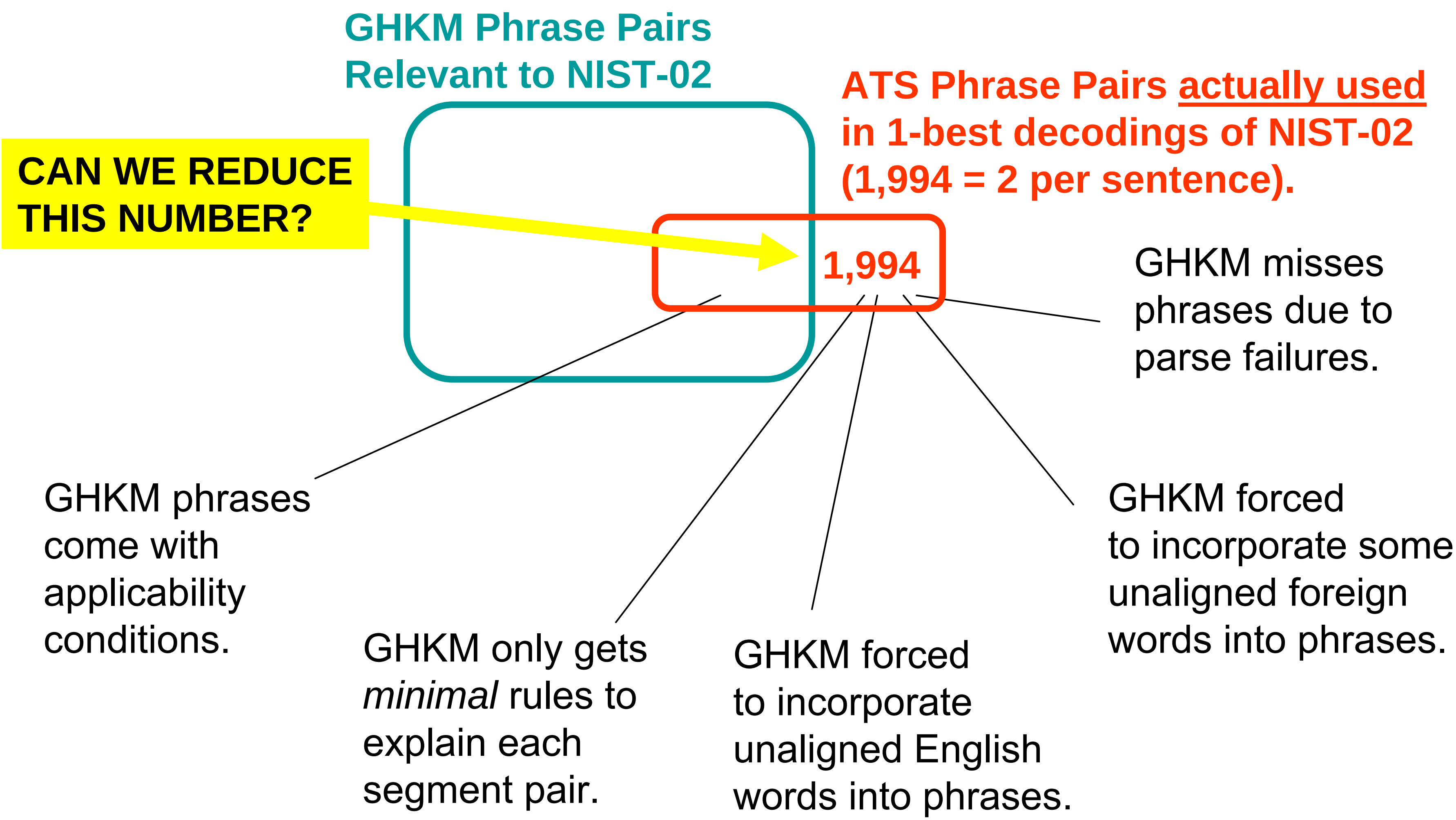
Lexicalized Re-Ordering



ATS and GHKM Methods Do Not Coincide



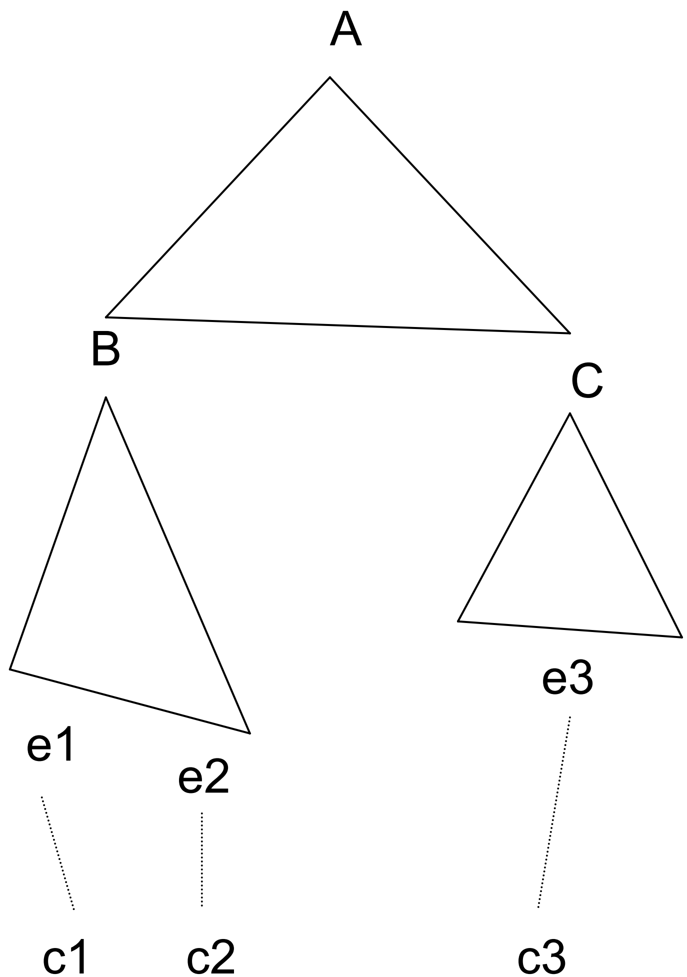
ATS and GHKM Methods Overlap



Some Methods for Improving Syntax-Based Rule Extraction

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

Larger, Composed Rules



Minimal GHKM Rules:

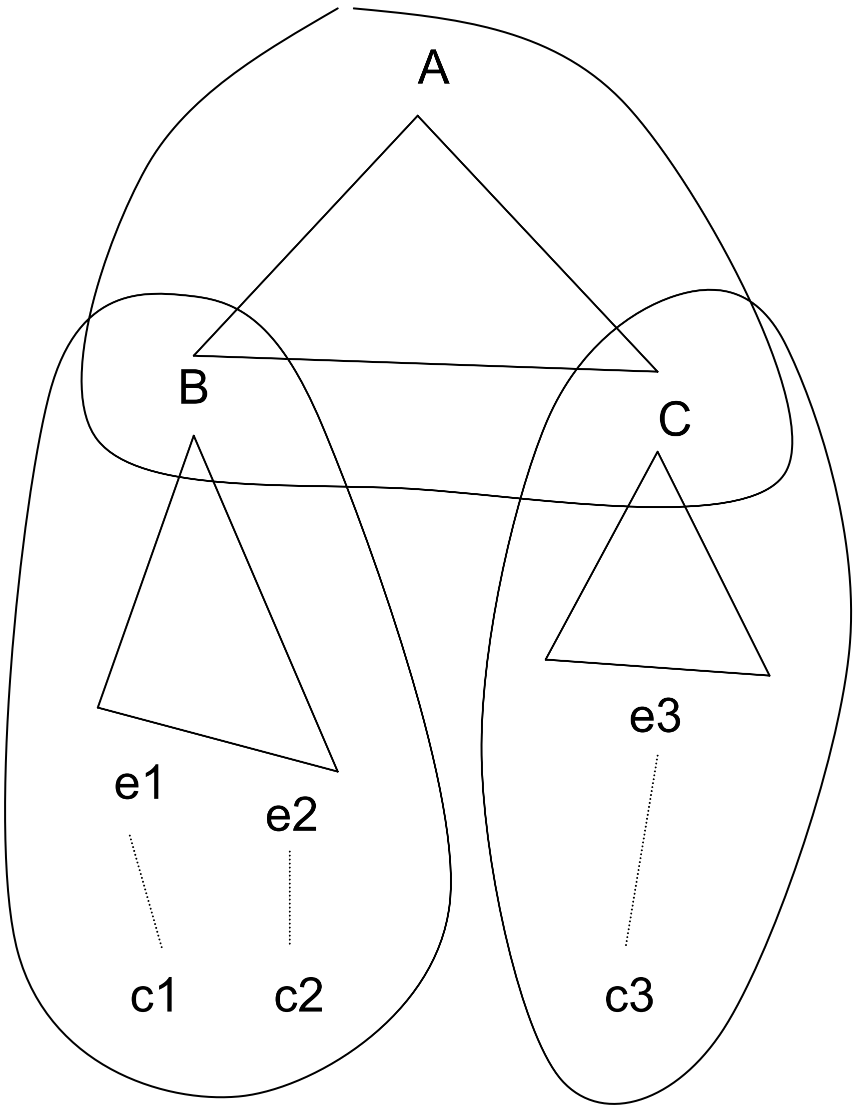
B(e1 e2) → c1 c2
C(e3) → c3
A(x0:B x1:C) → x0 x1

Additional Composed Rules:

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

"big phrasal rule"

Larger, Composed Rules



Minimal GHKM Rules:

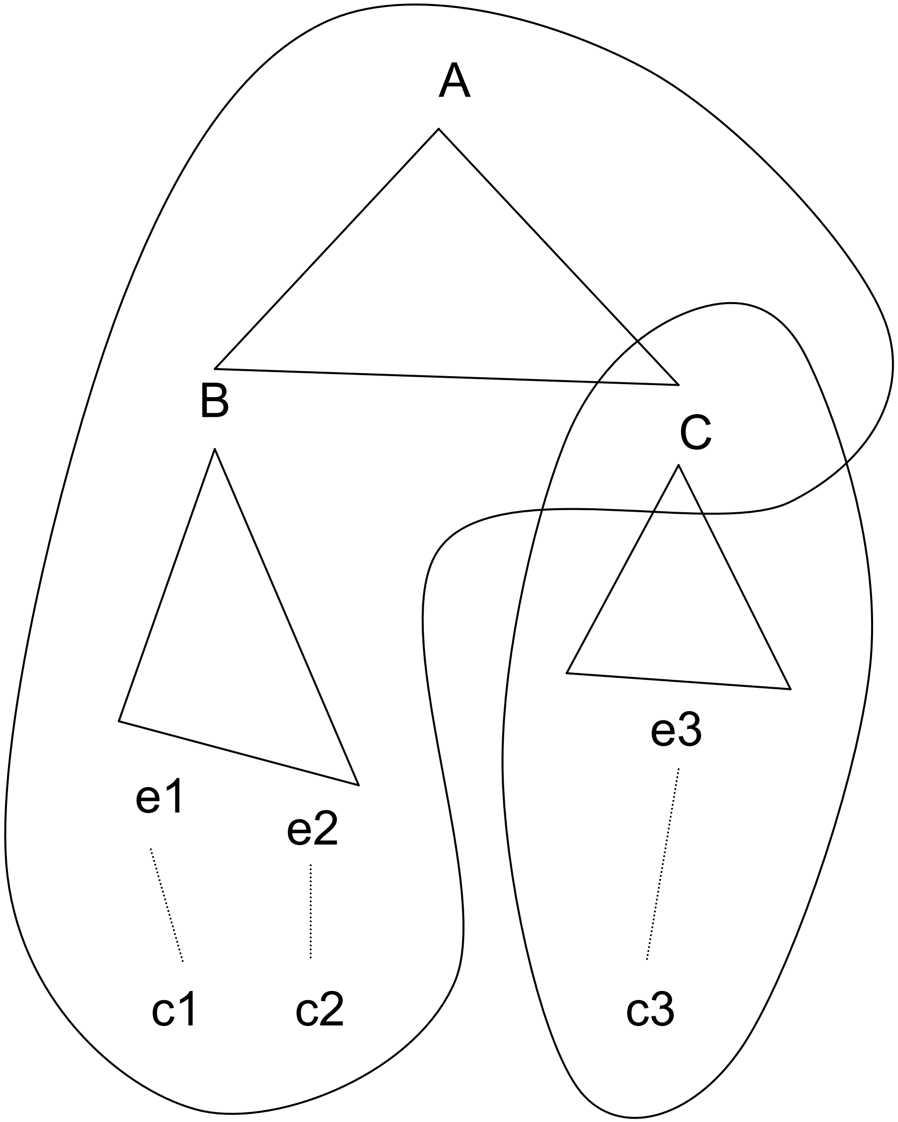
B(e1 e2) → c1 c2
C(e3) → c3
A(x0:B x1:C) → x0 x1

Additional Composed Rules:

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

"big phrasal rule"

Larger, Composed Rules



Minimal GHKM Rules:

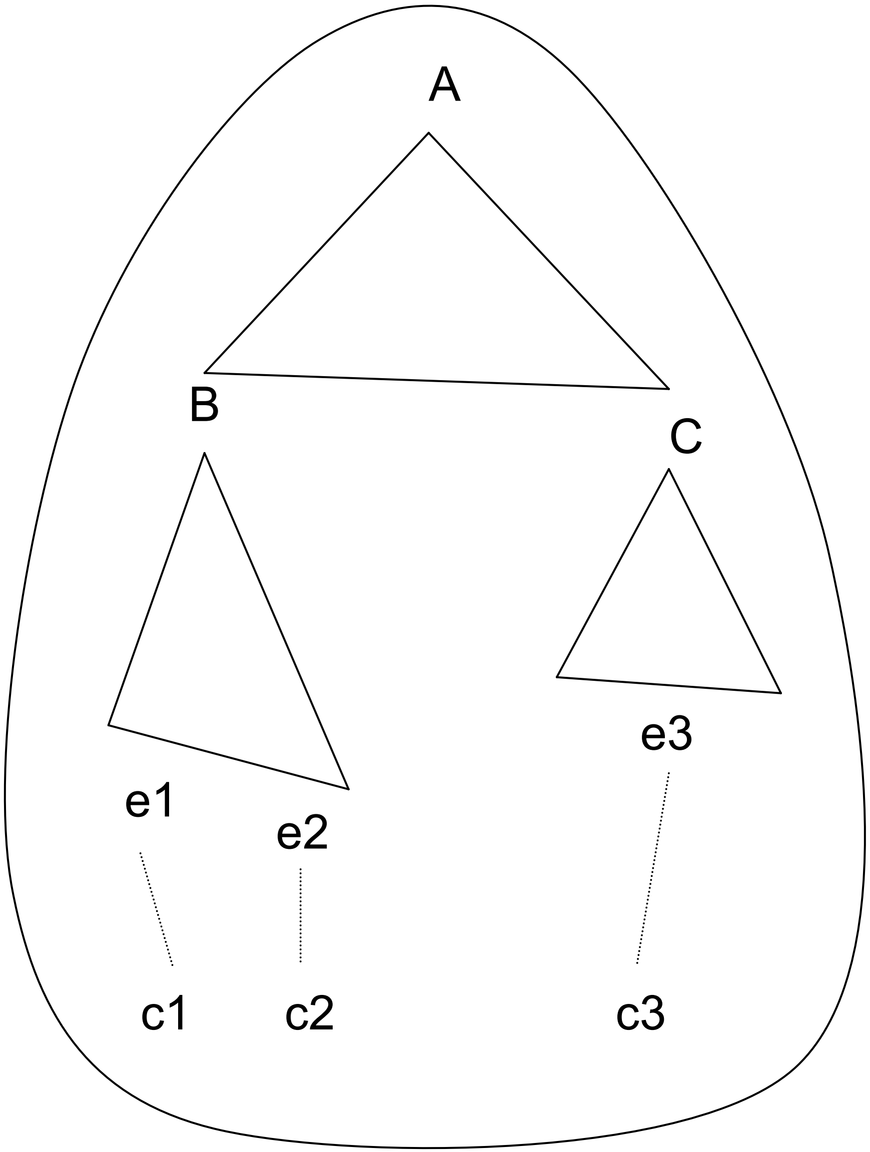
B(e1 e2) → c1 c2
C(e3) → c3
A(x0:B x1:C) → x0 x1

Additional Composed Rules:

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

"big phrasal rule"

Larger, Composed Rules



Minimal GHKM Rules:

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

Additional Composed Rules:

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

"big phrasal rule"

GHKM (Galley et al, 2006)

RULES ACQUIRED:

VBD(felt) → 有

VBN(obliged) → 责任

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

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

S(x0:NP-C x1:VP) → x0 x1

GHKM (Galley et al, 2006)

RULES ACQUIRED:

VBD(felt) → 有

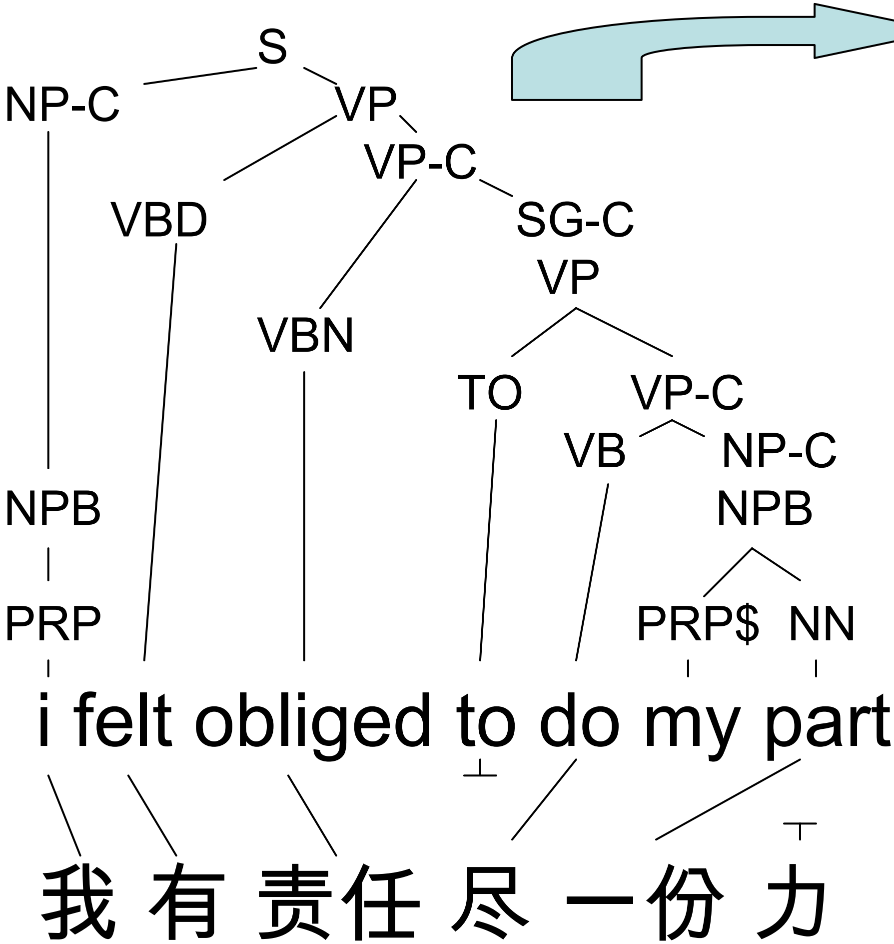
VBN(obliged) → 责任

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

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

S(x0:NP-C x1:VP) → x0 x1

GHKM (Galley et al, 2006)



RULES ACQUIRED:

VBD(felt) → 有

VBN(obliged) → 责任

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

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

S(x0:NP-C x1:VP) → x0 x1

minimal rules tile the tree/string/alignment triple.

composed rules are made by combining those tiles.

Larger, Composed Rules

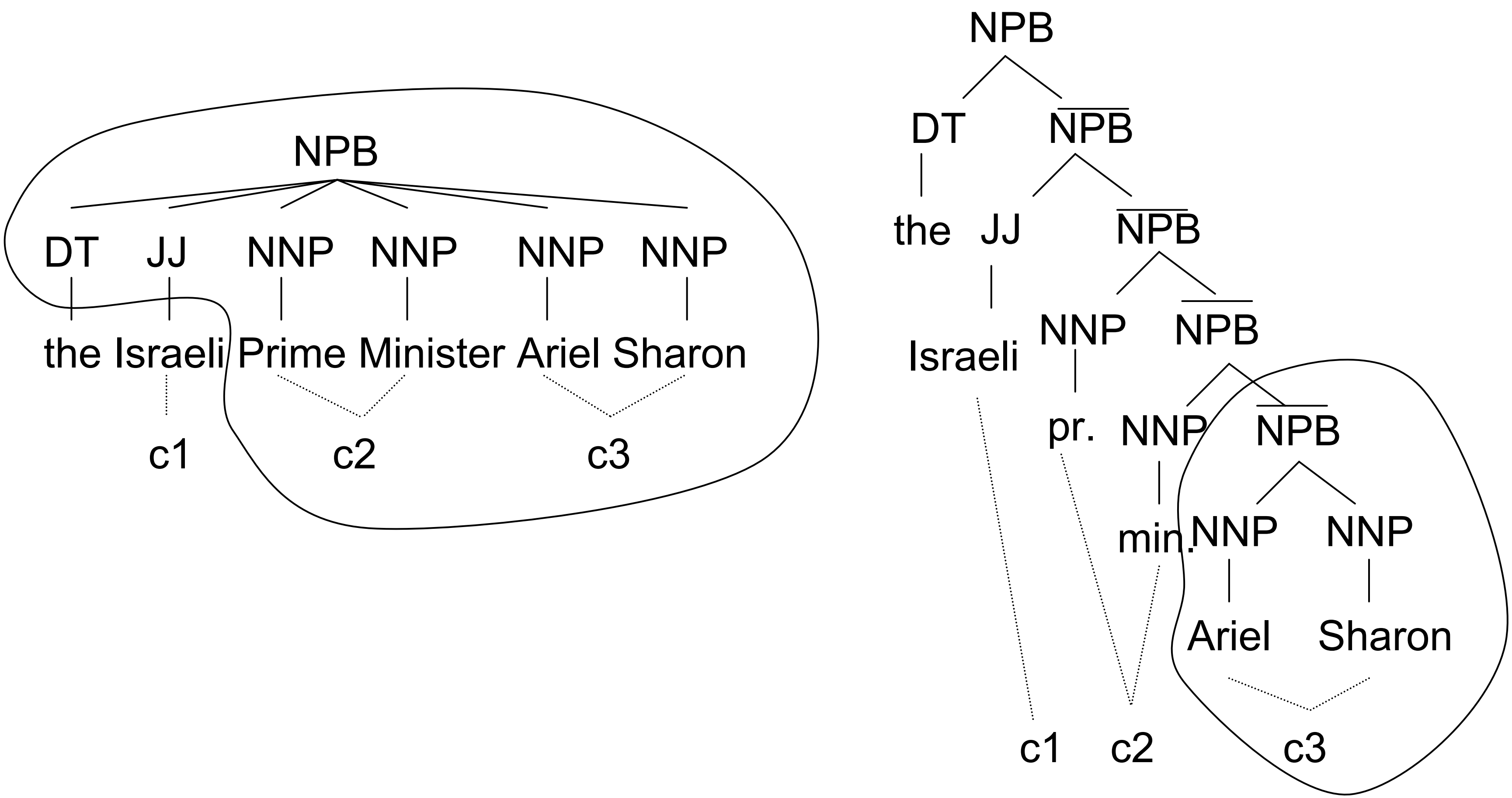
Composed limit (internal nodes in composed rule)	# of rules acquired	Unacquired phrase pairs used in ATS 1-best decodings
0 = minimal	2.5m	1994
2	12.4m	1478
3	26.9m	1096
4	**55.8m**	**900**

“Phrasal” Syntax Rules

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

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

Restructuring English Training Trees



Restructuring English Training Trees

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

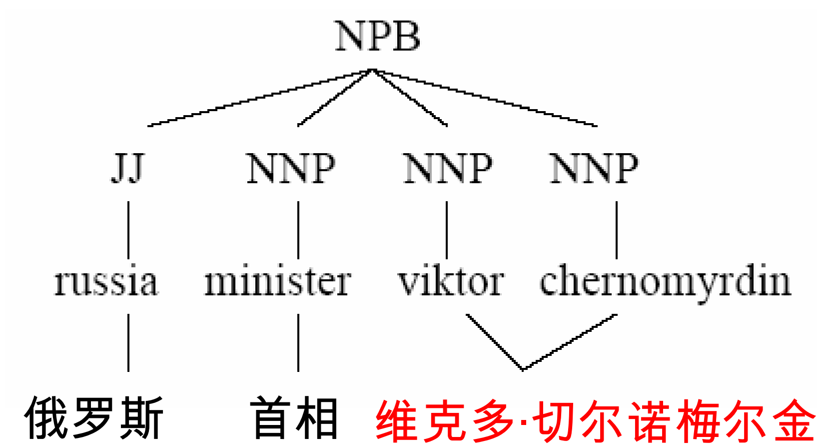
Effects of Coverage Improvements on Syntax-Based MT Accuracy

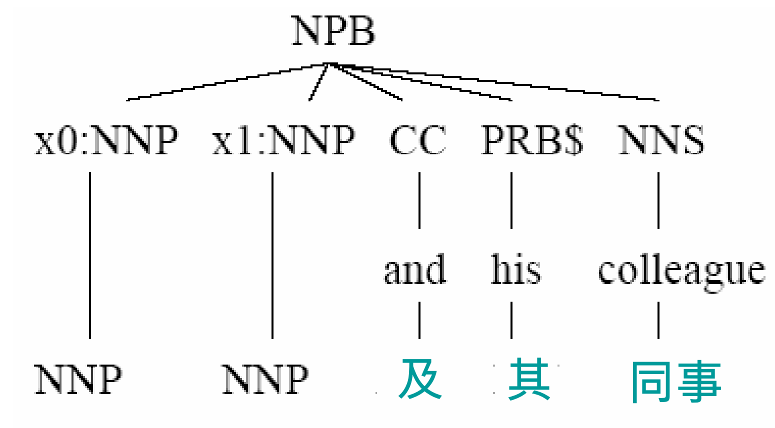
	Chinese/English Trained on 9.8m words		Arabic/English Trained on 4.1m words	
	Dev-02	Test-03	Dev-02	Test-03
ATS	36.00	34.31	50.88	51.04
GHKM minimal	39.11	38.85	49.81	50.46
GHKM composed 2	41.59	40.90	51.18	51.52
GHKM composed 3	42.28	41.62	51.96	52.04
GHKM composed 4	42.63	41.82	52.05	52.26
GHKM minimal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

NIST Bleu r4n4

Improved English Binarization

Why are Penn Treebank Trees Problematic for Translation?





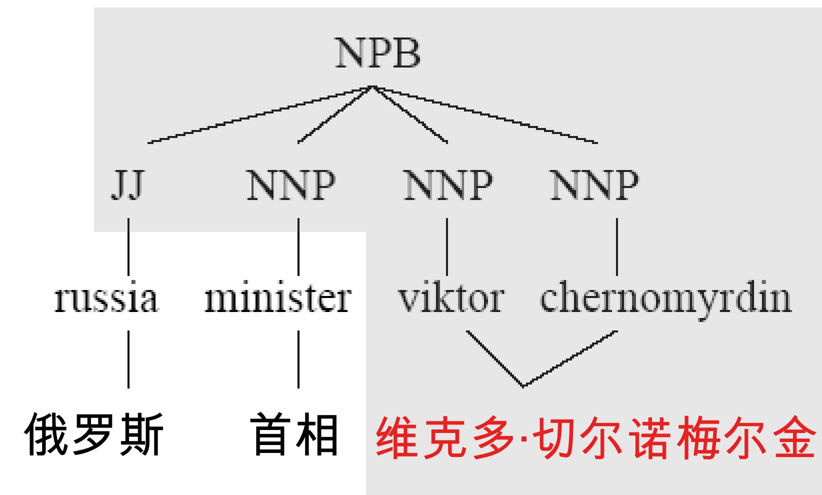
?

维克多·切尔诺梅尔金 及 其 同事

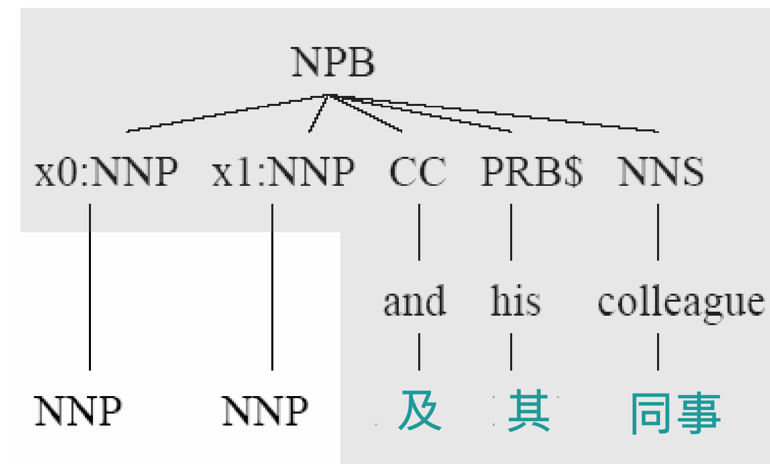
Improved English Binarization

Why are Penn Treebank Trees Problematic for Translation?

R1



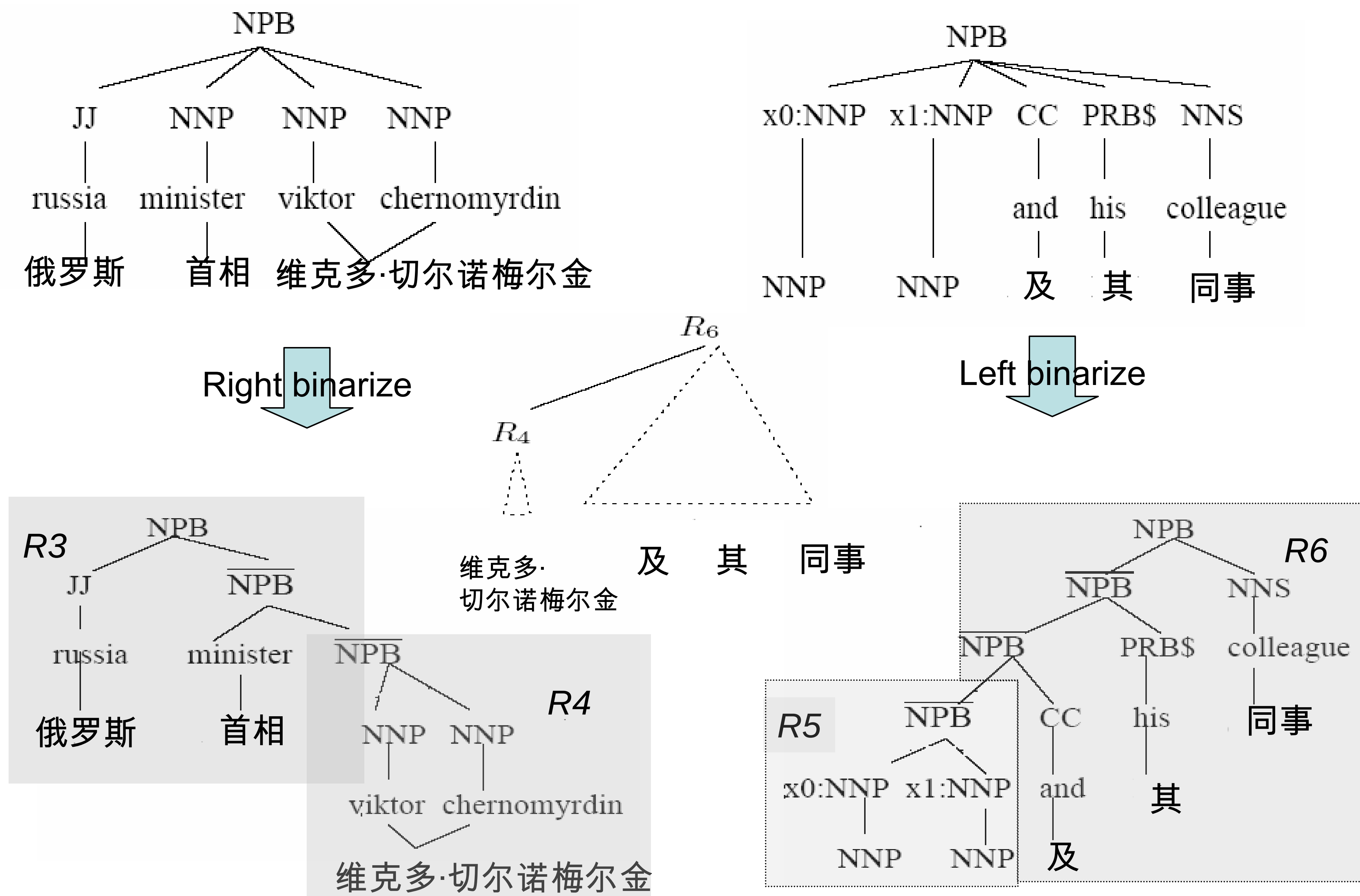
R2



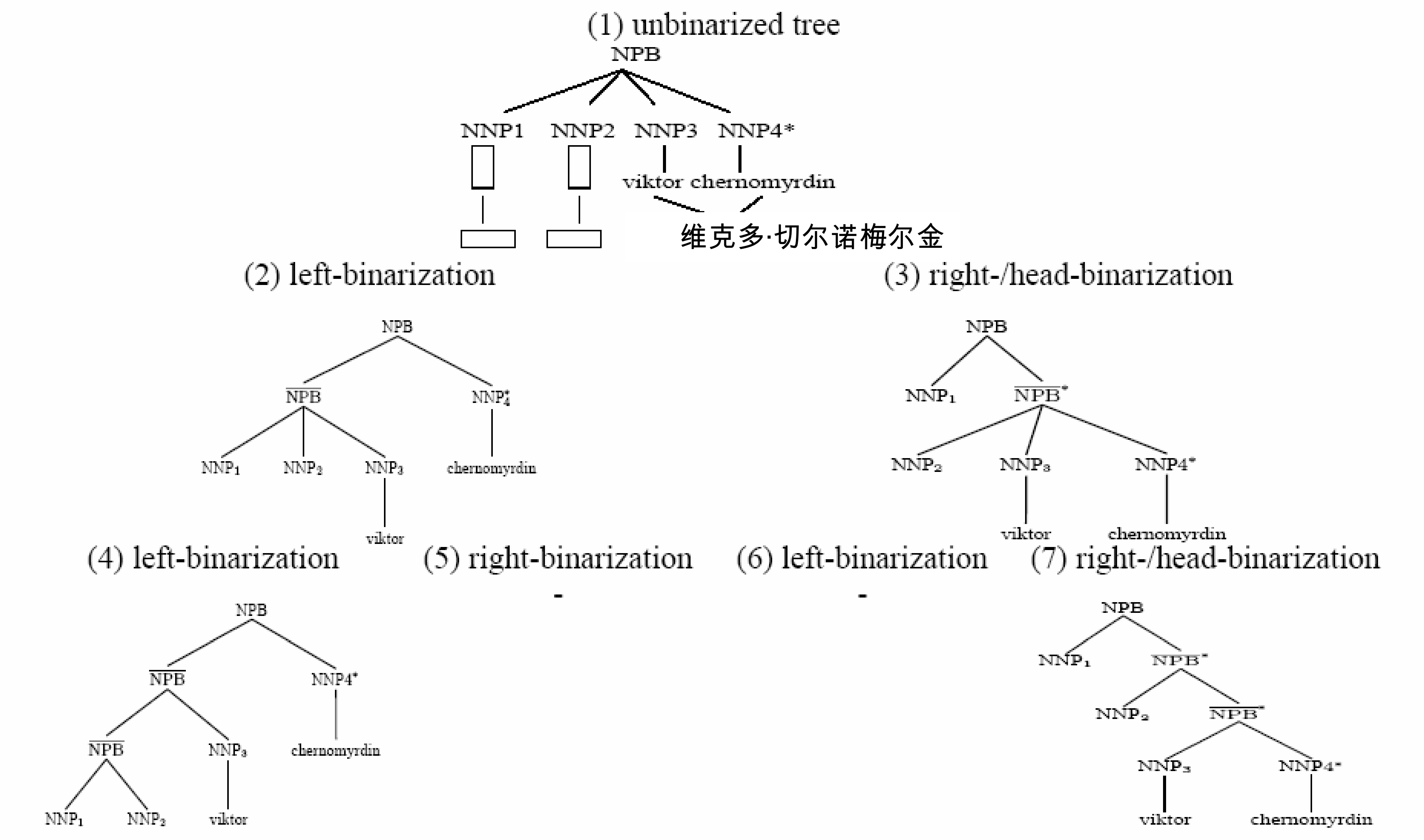
?

维克多·切尔诺梅尔金 及 其 同事

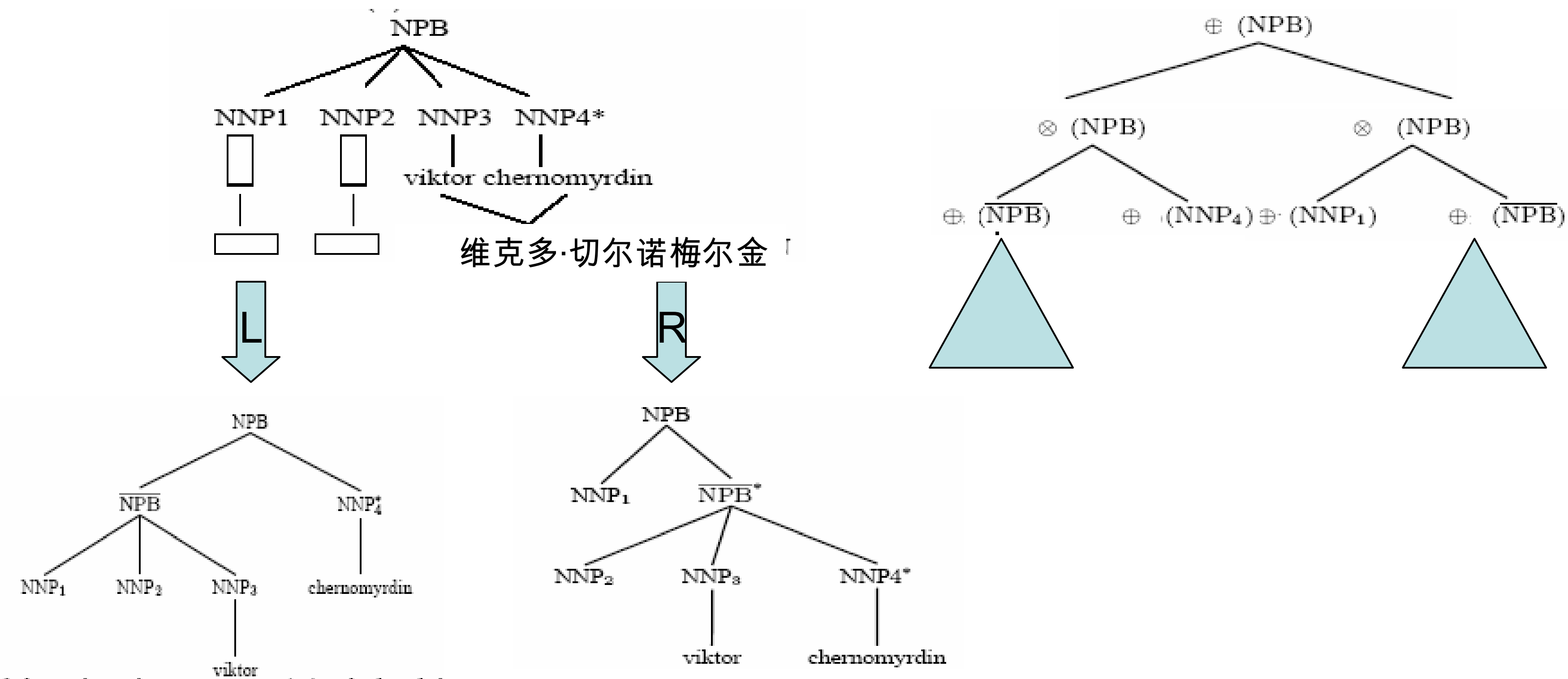
Binarizing English Trees



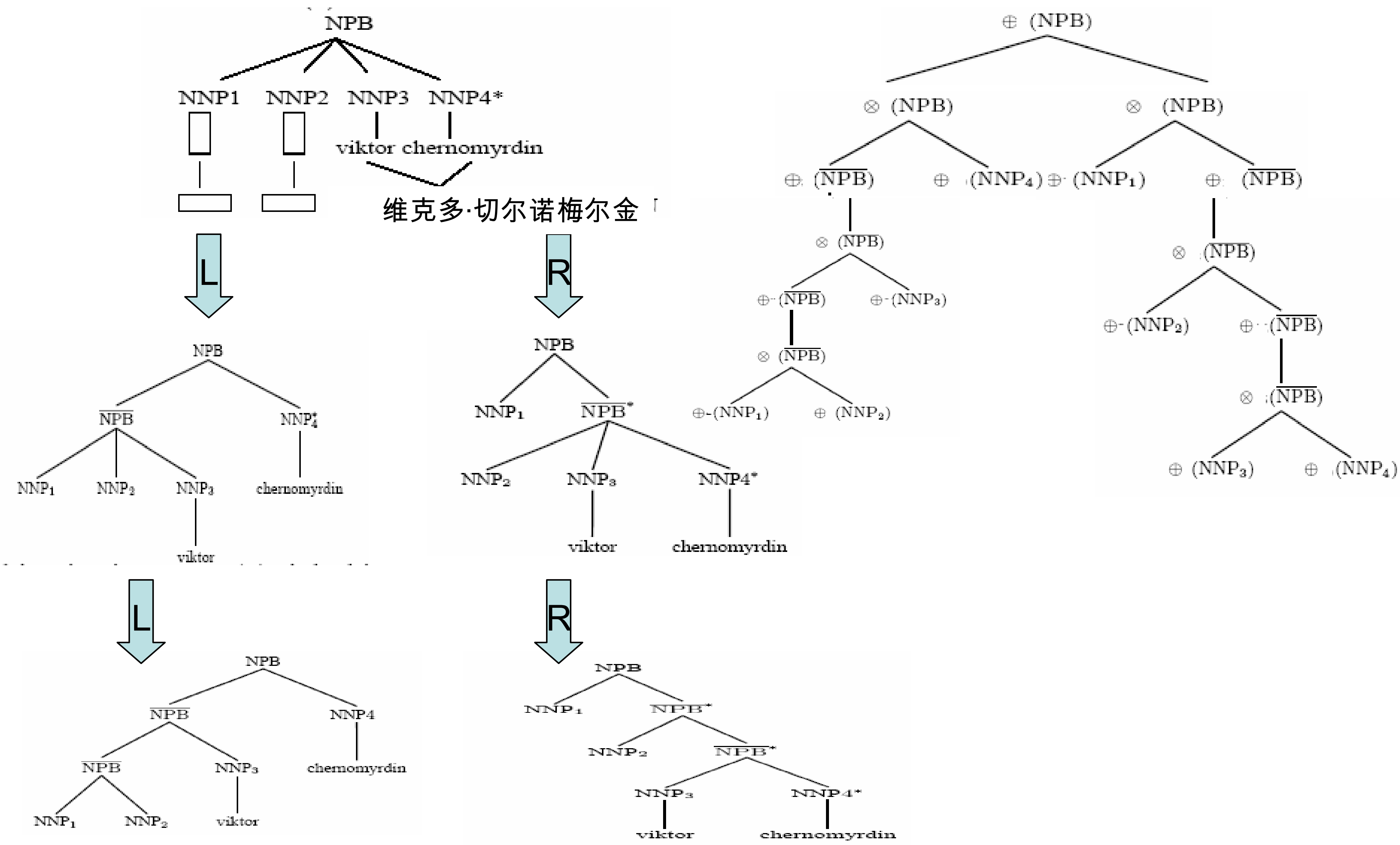
Simple Binarizations



Parallel Binarization



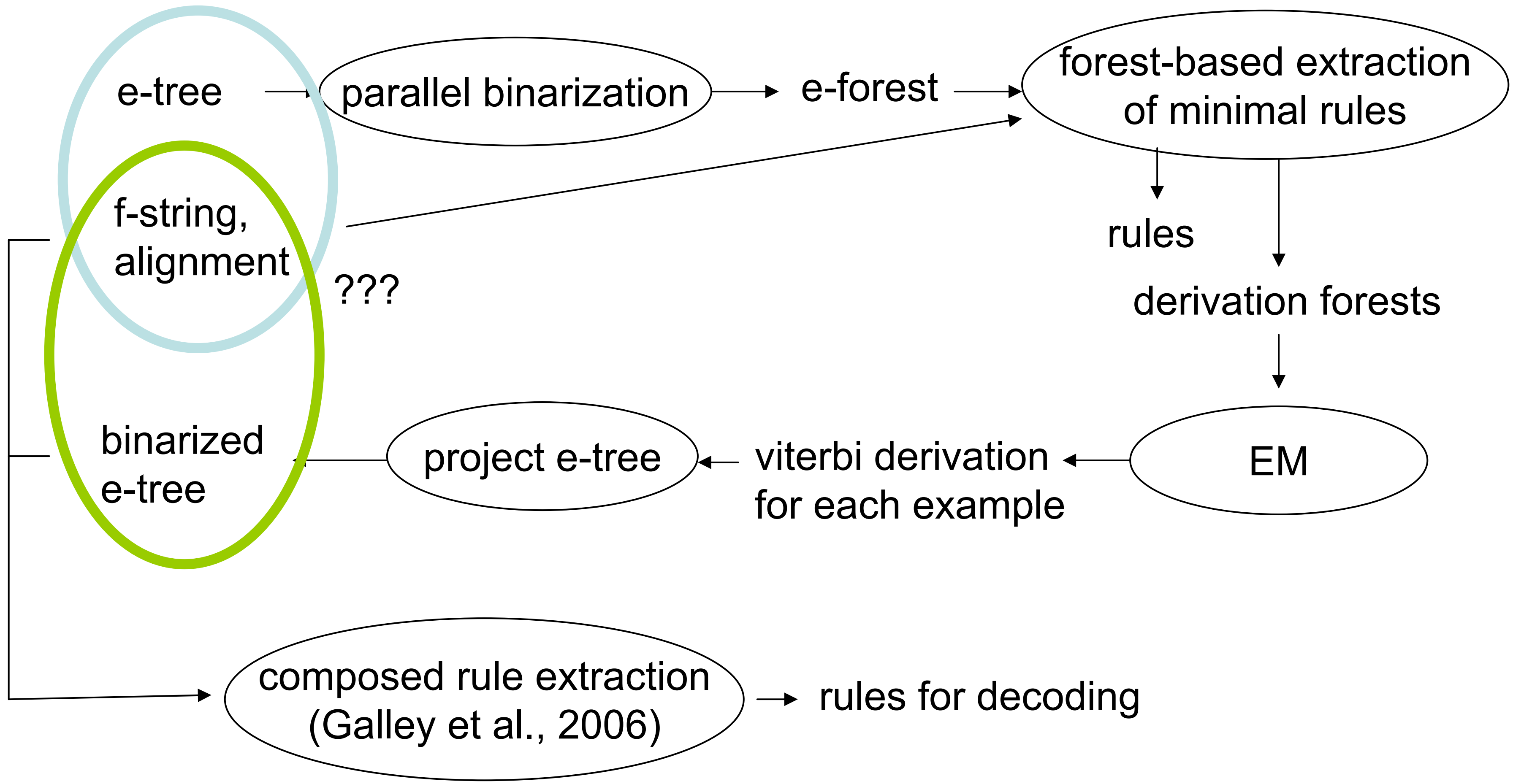
Parallel Binarization



Forest-Based Rule Extraction

- Gets **all** minimal rules consistent with word alignment and **some** binarization
- Run EM algorithm to determine best binarization of each node in each tree

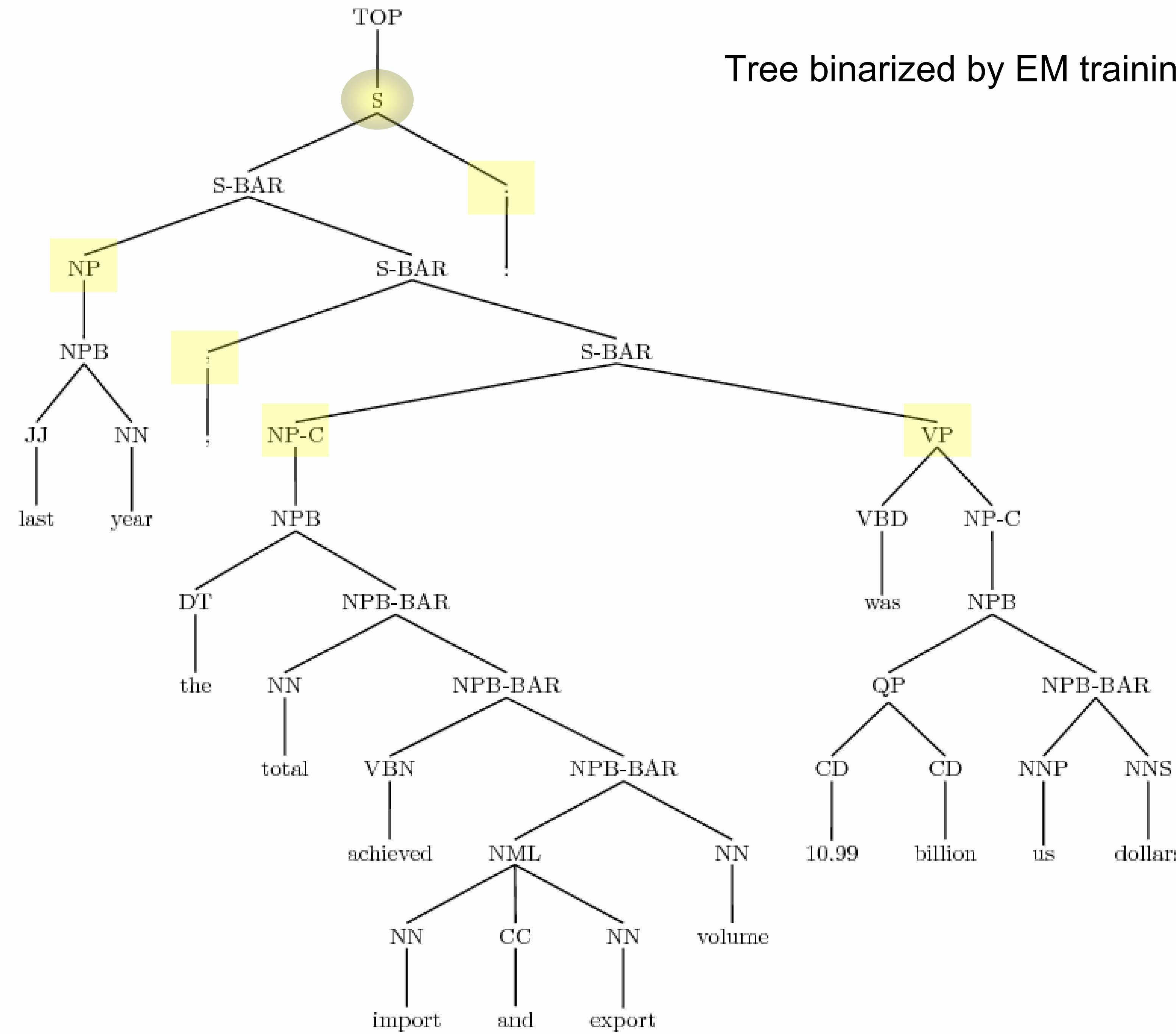
Binarization Using EM



Experimental Results
(Wang, Knight, Marcu 2007)

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

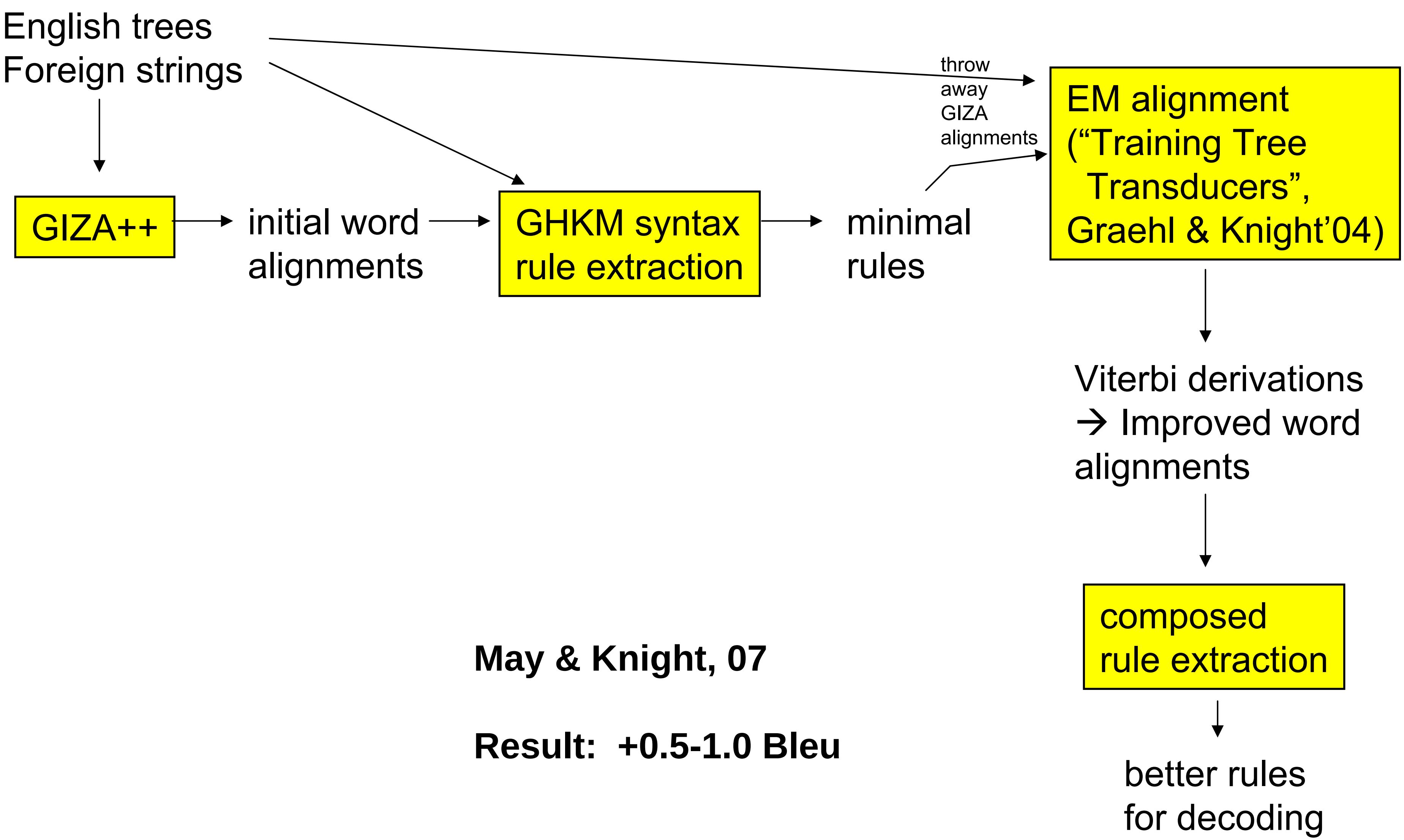
Tree binarized by EM training



Syntax-Based Word Alignment

- GIZA++ string-based alignments
  - are errorful
  - don’t match our syntax-based MT system
- We would like to use our tree-based translation model to align data

Syntax-Based Word Alignment



May & Knight, 07

Result: +0.5-1.0 Bleu

Remarks

- Phrase-based and syntax-based extraction algorithms have different coverage.
- Syntax-based coverage can be improved:
  - composed rules
  - phrasal rules
  - binarizing English trees with EM
  - re-aligning tree/string pairs with EM
- Improvements lead to better translation accuracy.

Some Sample Outputs

input

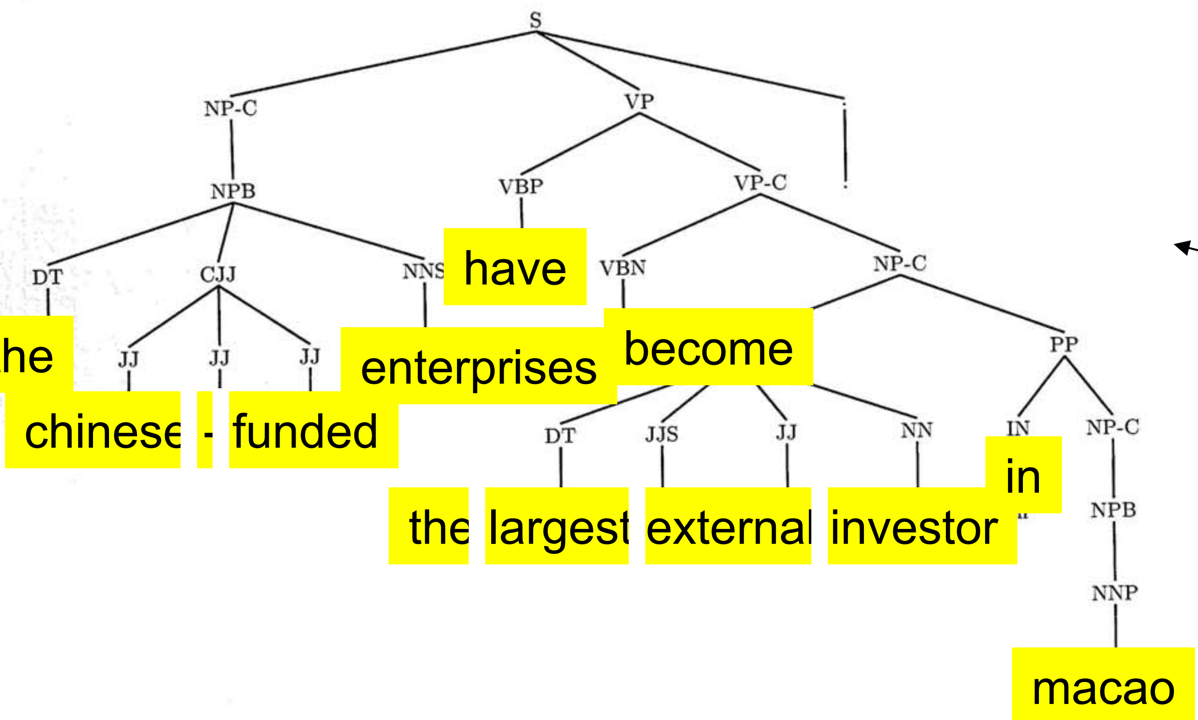
phrase-based
system output

syntax-based
system output

dev-little (line 47) - dev-little

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

[dev-little] 1-Best Tree



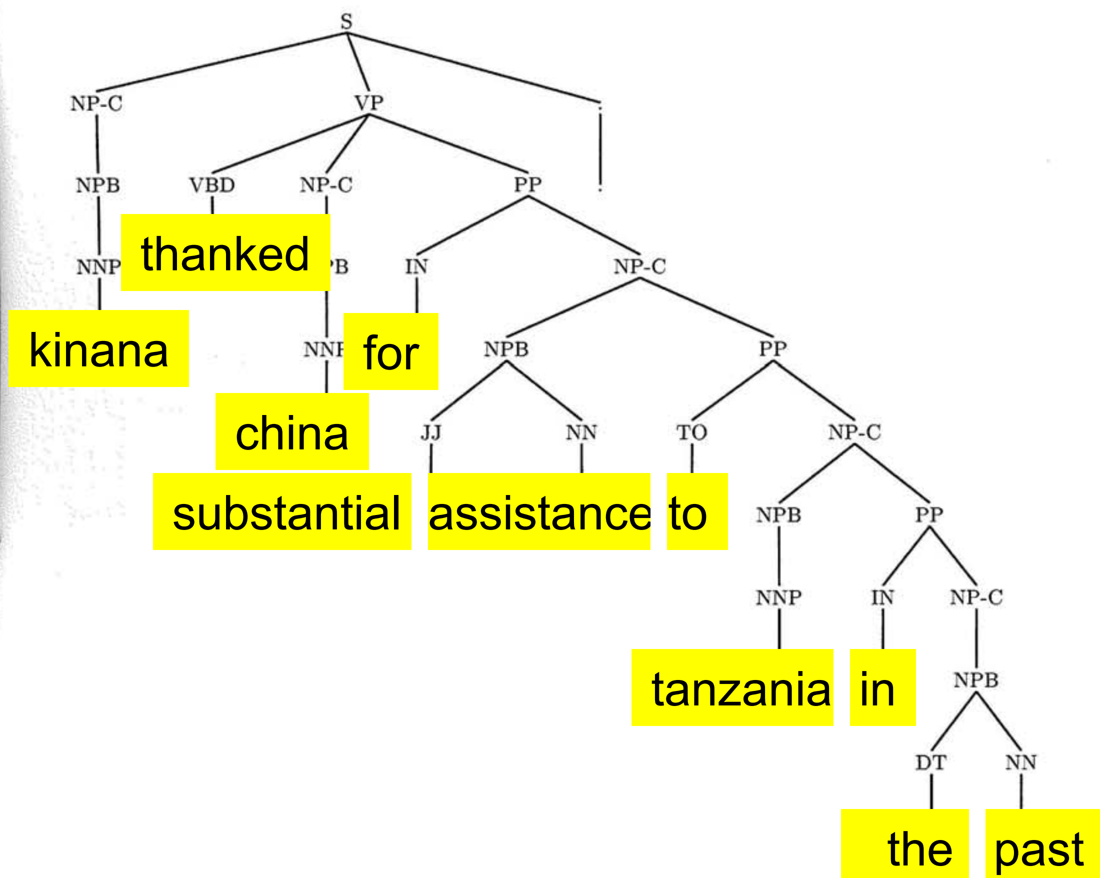
input

dev-little (line 59) - dev-little

Input: 基 纳 纳 对 中国 过去 向 坦 桑 提供 的 大量 援助 表示 感谢 。
Reference: keenana expressed gratitude to china for its great assistance to tanzania in the past .
AlTemp-e: kinana $_{0}$ | to $_{1}$ | china $_{2}$ | in the past to $_{3}$ | tanzania $_{4}$ | expressed appreciation $_{5}$ | for the substantial $_{6}$ | assistance $_{7}$ | . $_{8}$
AlTemp-f: 基 纳 纳 $_{0}$ | 对 $_{1}$ | 中国 $_{2}$ | 过去 向 $_{3}$ | 坦 桑 $_{4}$ | 提供 的 大量 $_{6}$ | 援助 $_{7}$ | 表示 感谢 $_{5}$ | 。 $_{8}$
[dev-little] 1-Best: kinana thanked china for substantial assistance to tanzania in the past .

phrase-based system output

[dev-little] 1-Best Tree

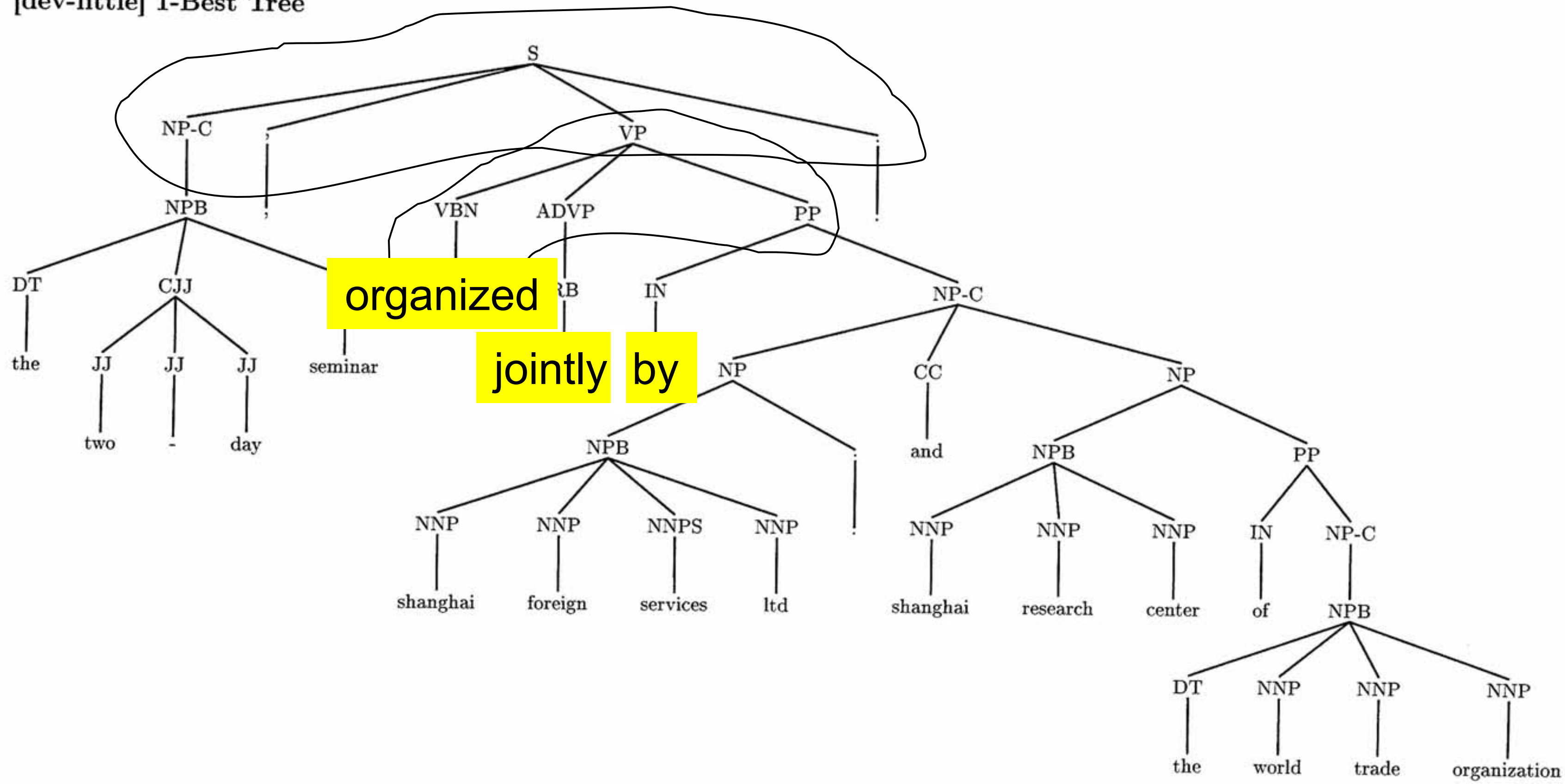


syntax-based system output

dev-little (line 38) - dev-little

Input: 此次 为期 两天 的 研讨会 ， 由 世界贸易组织 上海 研究中心 与 上海市 对外 服务 有限公司 联合 举办 。
Reference: the two - day seminar is jointly sponsored by the wto shanghai research center and shanghai foreign service company limited .
AlTemp-e: the $_0$ | two - day $_1$ | seminar $_2$ | by the world trade organization $_3$ | , $_4$ | shanghai research center $_5$ | and $_6$ | shanghai foreign service $_7$ | co . , ltd . $_8$ | jointly $_9$ | . $_{10}$
AlTemp-f: 此次 $_0$ | 为期 两天 的 $_1$ | 研讨会 $_2$ | ， $_4$ | 由 世界贸易组织 $_3$ | 上海 研究中心 $_5$ | 与 $_6$ | 上海市 对外 服务 $_7$ | 有限公司 $_8$ | 联合 举办 $_9$ | 。 $_{10}$
[dev-little] 1-Best: the two - day seminar , organized jointly by shanghai foreign services ltd . and shanghai research center of the world trade organization .

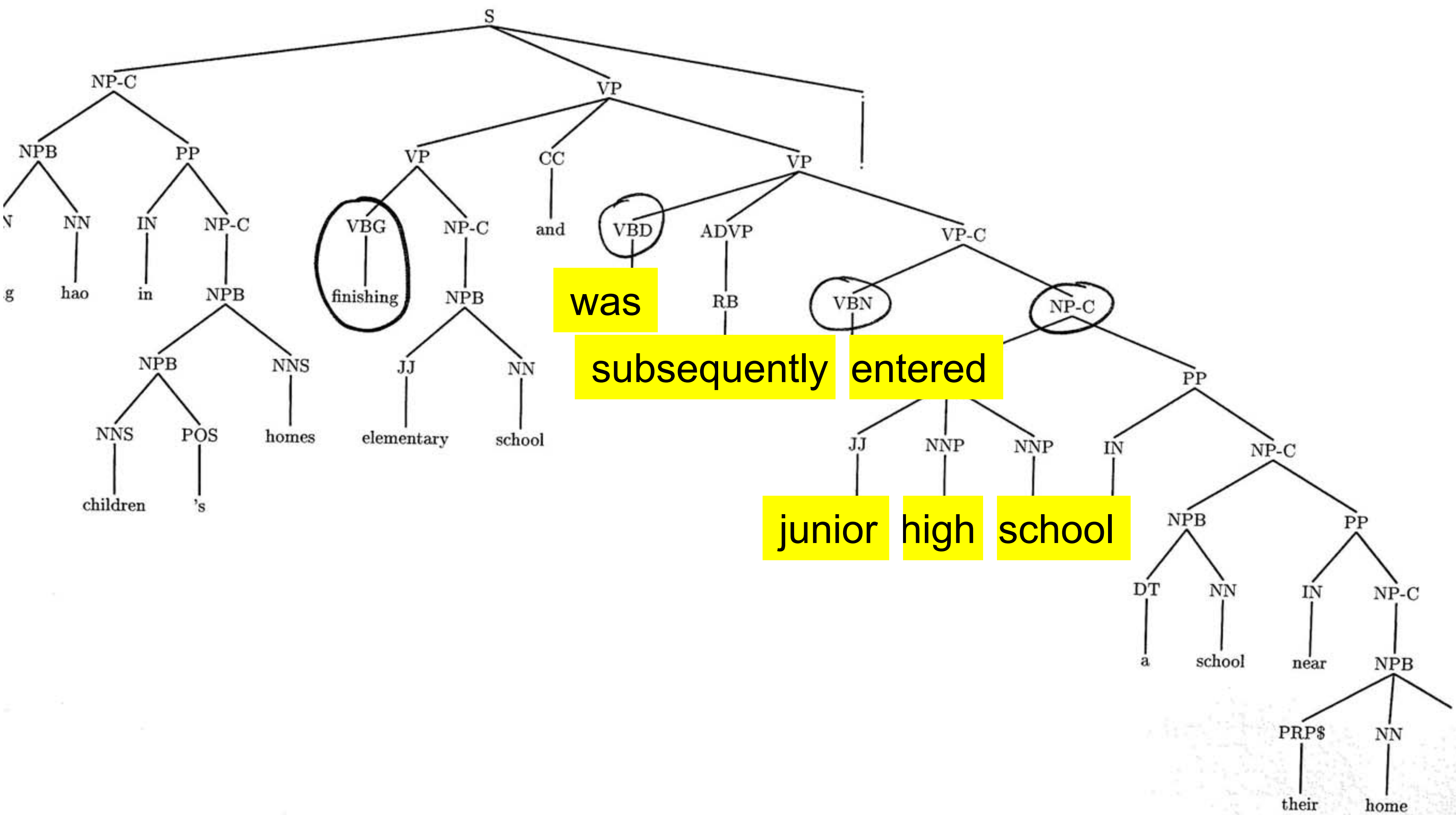
[dev-little] 1-Best Tree



lev-little (line 53) - dev-little

Input: 丁 豪 在 儿童 福利 院 读 完 小学 ， 随后 进入 附近 乡里 一所 学校 上 初中 。
Reference: ding hao completed his primary school at the children welfare school , and then went to a nearby township middle school .
AlTemp-e: ding hao $_0$ | in $_1$ | children 's welfare institute $_2$ | complete primary school $_3$ | , $_4$ | then entered $_5$ | commune $_6$ | near $_7$ | a school $_8$ | , $_9$ | junior high school level . $_{10}$
AlTemp-f: 丁 豪 $_0$ | 在 $_1$ | 儿童 福利 院 $_2$ | 读 完 小学 $_3$ | ， $_4$ | 随后 进入 $_5$ | 附近 $_7$ | 乡里 $_6$ | 一所 学校 $_8$ | 上 $_9$ | 初中 。 $_{10}$
dev-little] 1-Best: ding hao in children 's homes finishing elementary school and was subsequently entered junior high school at a school near their home towns .

lev-little] 1-Best Tree

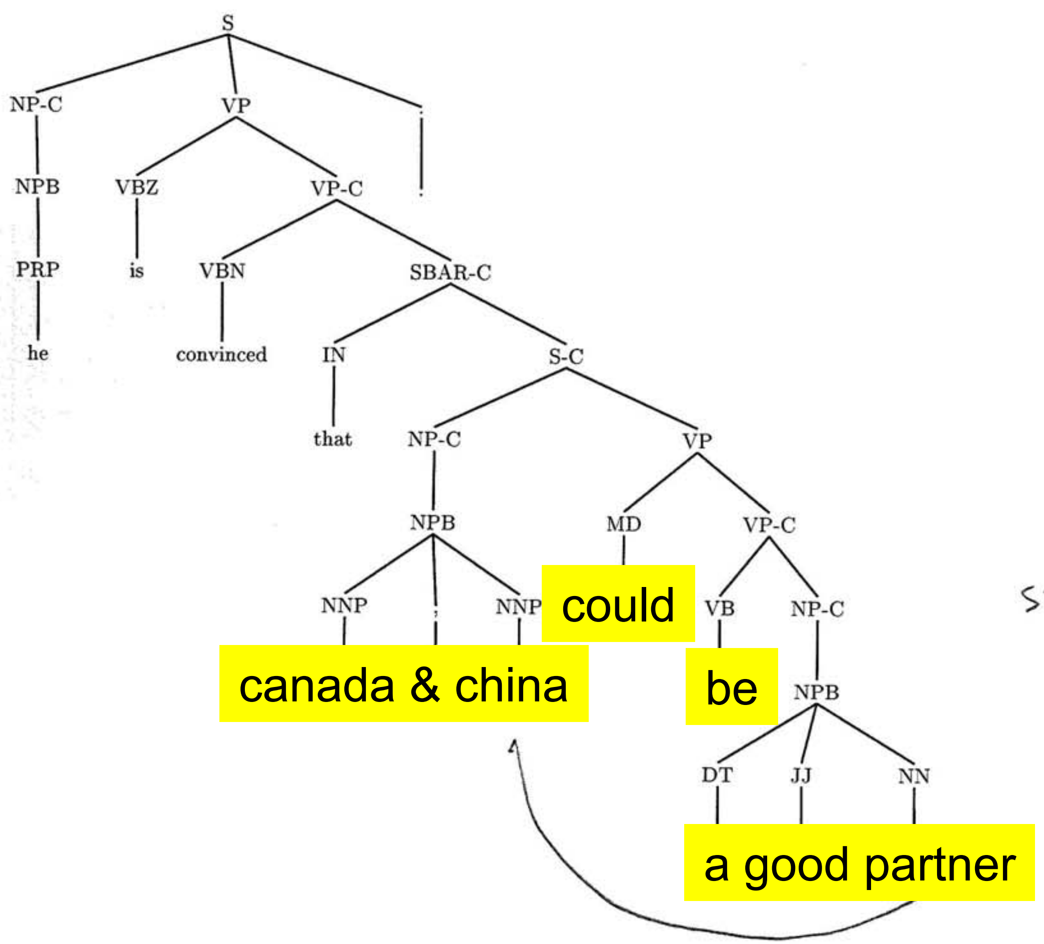


dev-little (line 64) - dev-little

can become very good partners

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

[dev-little] 1-Best Tree



Subj-obj number agreement.

dev-little (line 51) - dev-little

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

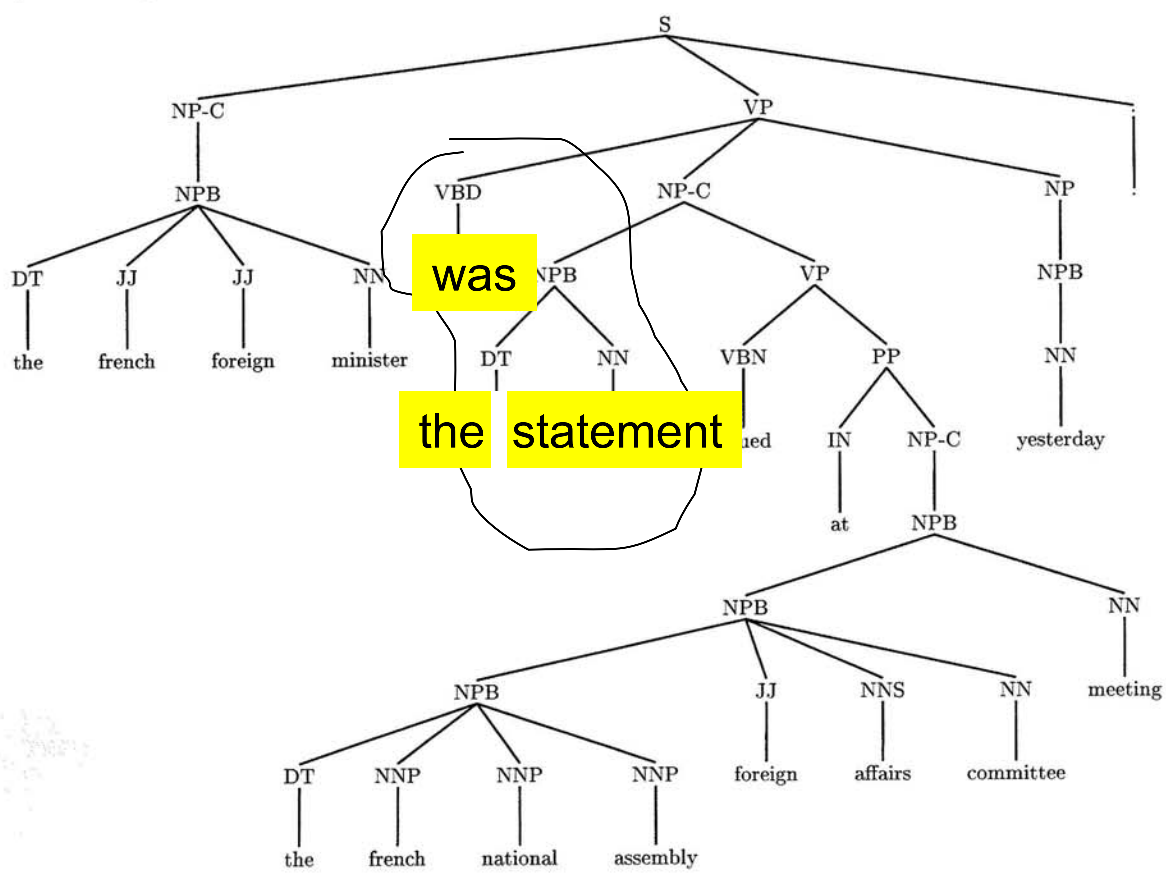
Reference: the french foreign minister made the above statement in a meeting of the foreign affairs commission of the french national congress .

AlTemp-e: french $_{0}$ | foreign minister $_{1}$ | in the french national assembly $_{2}$ | yesterday , $_{3}$ | the statement delivered by $_{4}$ | foreign affairs $_{5}$ | committee meeting $_{6}$ | . $_{7}$

AlTemp-f: 法国 $_{0}$ | 外长 $_{1}$ | 昨天 是 $_{3}$ | 在 法国 国民议会 $_{2}$ | 外事 $_{5}$ | 委员会 会议 上 $_{6}$ | 发表 上述 声明 $_{4}$ | 的 。 $_{7}$

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

[dev-little] 1-Best Tree



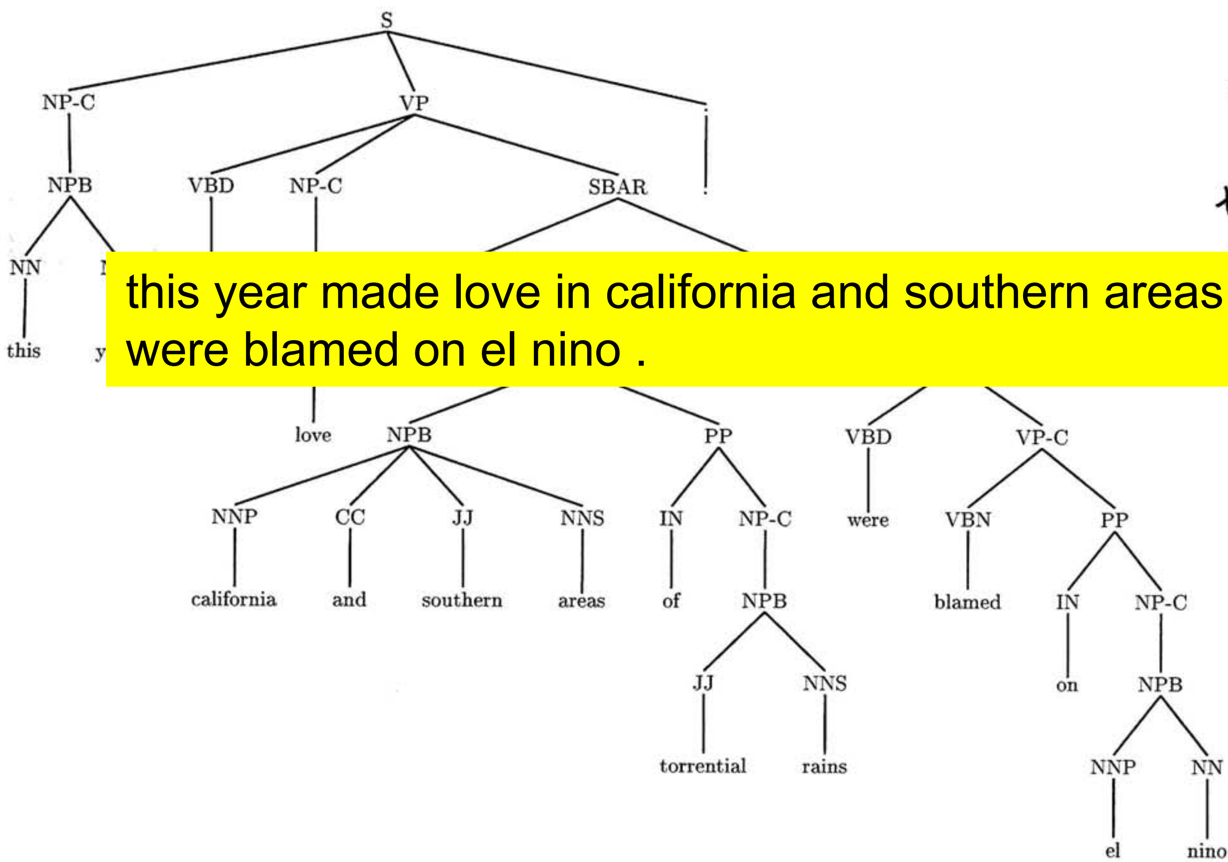
dev-little (line 125) - dev-little

rain

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

[dev-little] 1-Best Tree

Scope.



funny.

this year rarely NP-C

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

Lots of Open Problems

Chomsky’s Program [1957]

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

This Research Program has Contributed Powerful Ideas

Context-free grammar

Formal language hierarchy

Syntax, Phonology…

This Research Program is Really Unfinished

Type in your English sentence here:

Is this grammatical?

Is this sensible?

Lots of Open Problems

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

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

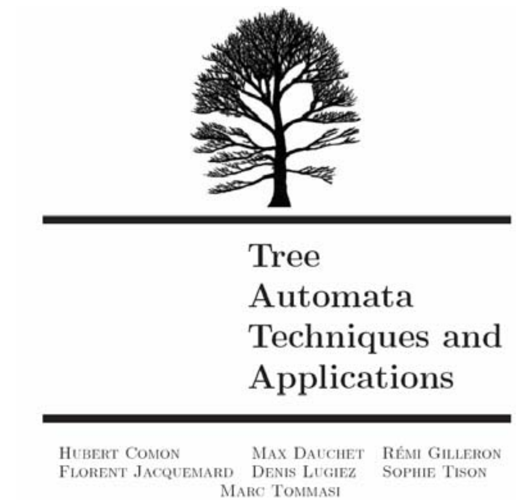
- Decoder search in runtime translation
  - Search errors hurt MT accuracy
  - Faster speed is needed to support experimentation
  - Some key ideas to date:
    - cube pruning [Chiang, 2007]
    - rule binarization [Zhang, Huang, Knight, Gildea, 2006]

Lots of Open Problems

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

Tree Automata

Doner (1968), Rounds (1970), Thatcher (1970), Engelfriet PhD thesis (1975), Gecseg & Steinby textbook (1984), …



CIAA

Conference on Implementation and Application of Automata

Tiburon: A Tree Automata Toolkit

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

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

Tiburon: A Tree Automata Toolkit

Towards simplifying system ideas:

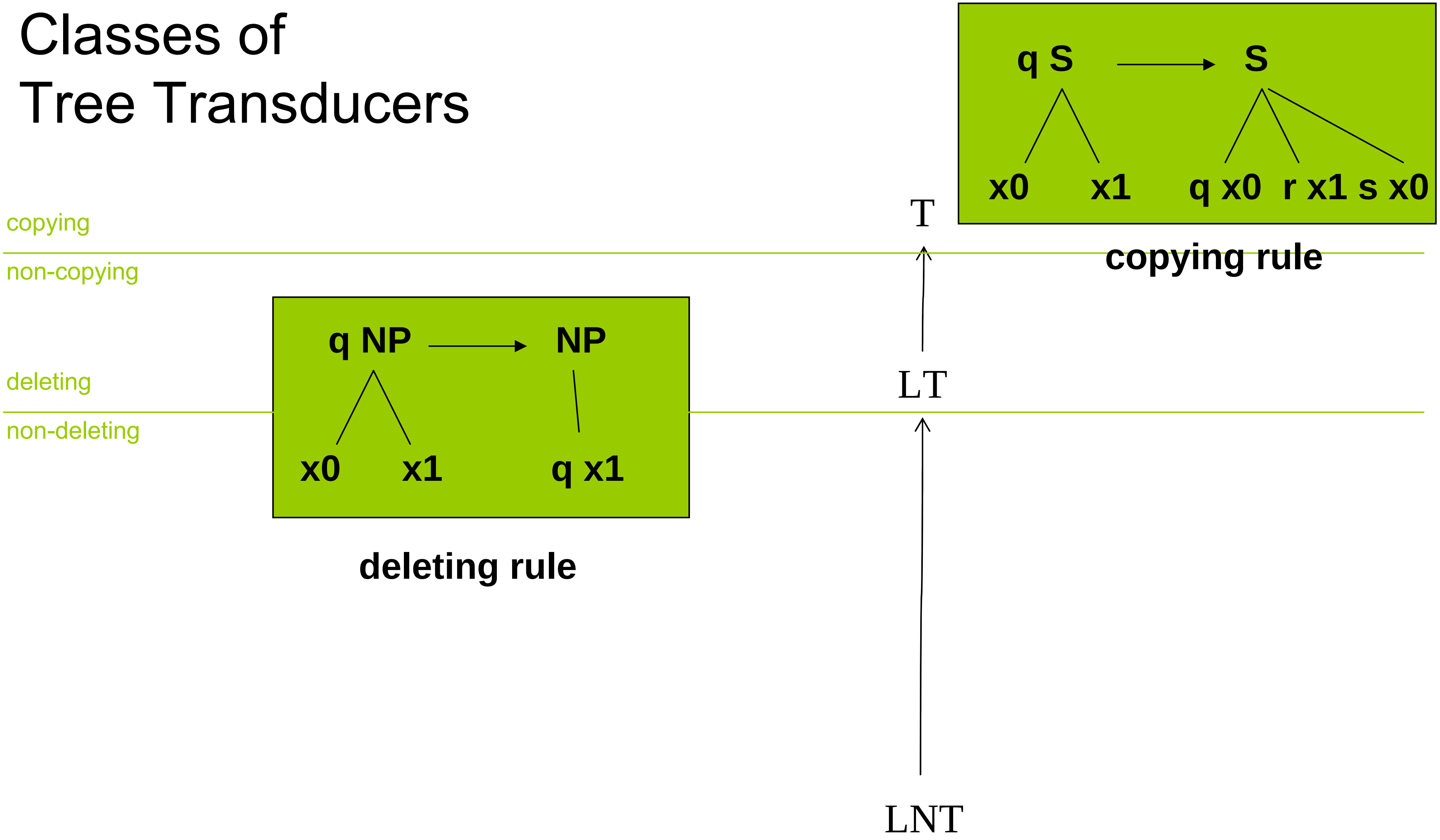
e = yield(best-tree(intersect(**lm.rtg**,
b-apply(cstring, **tm.tt**))))

What tree automata operations are needed/supported?

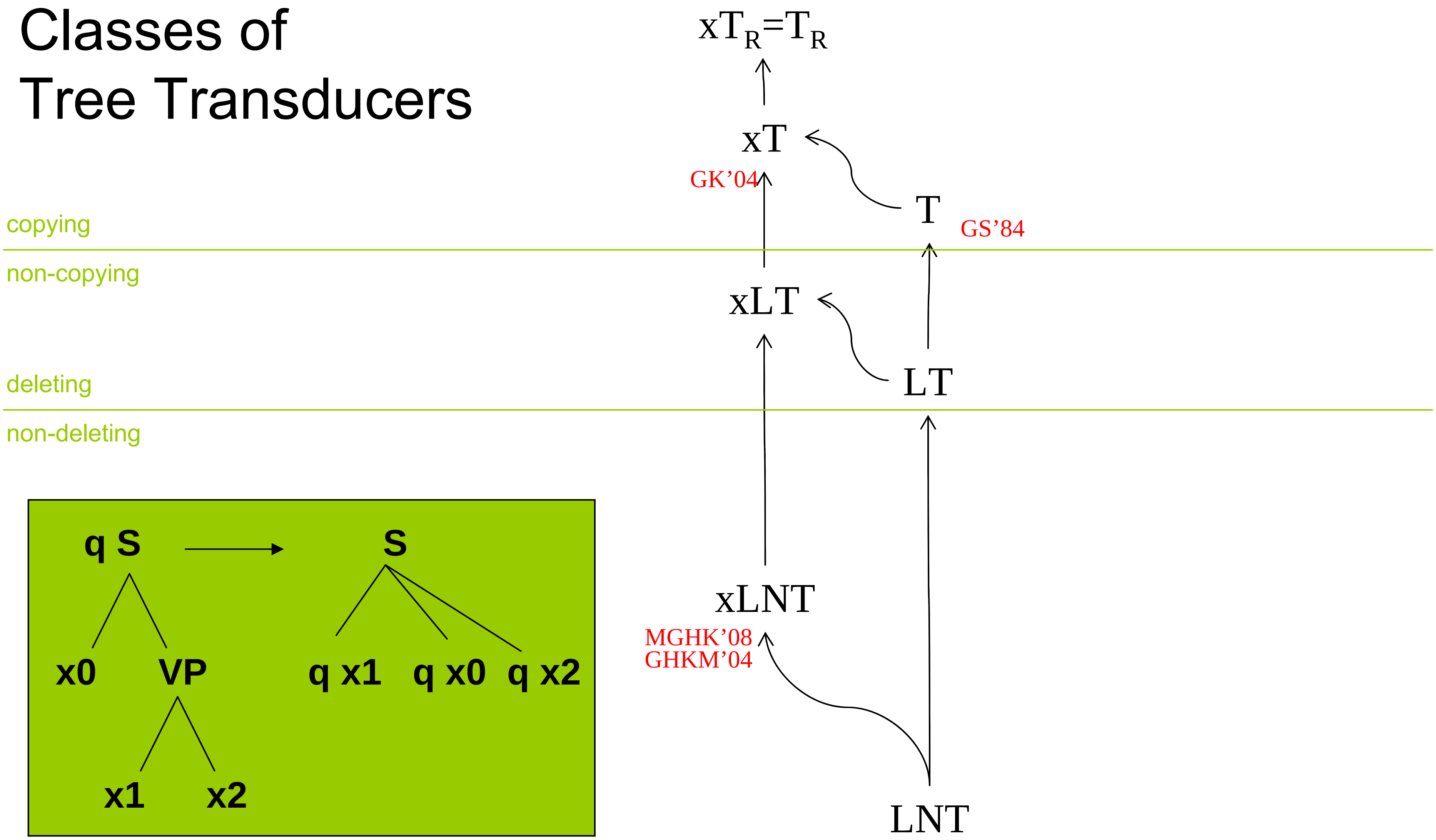
String World & Tree World

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

Classes of Tree Transducers



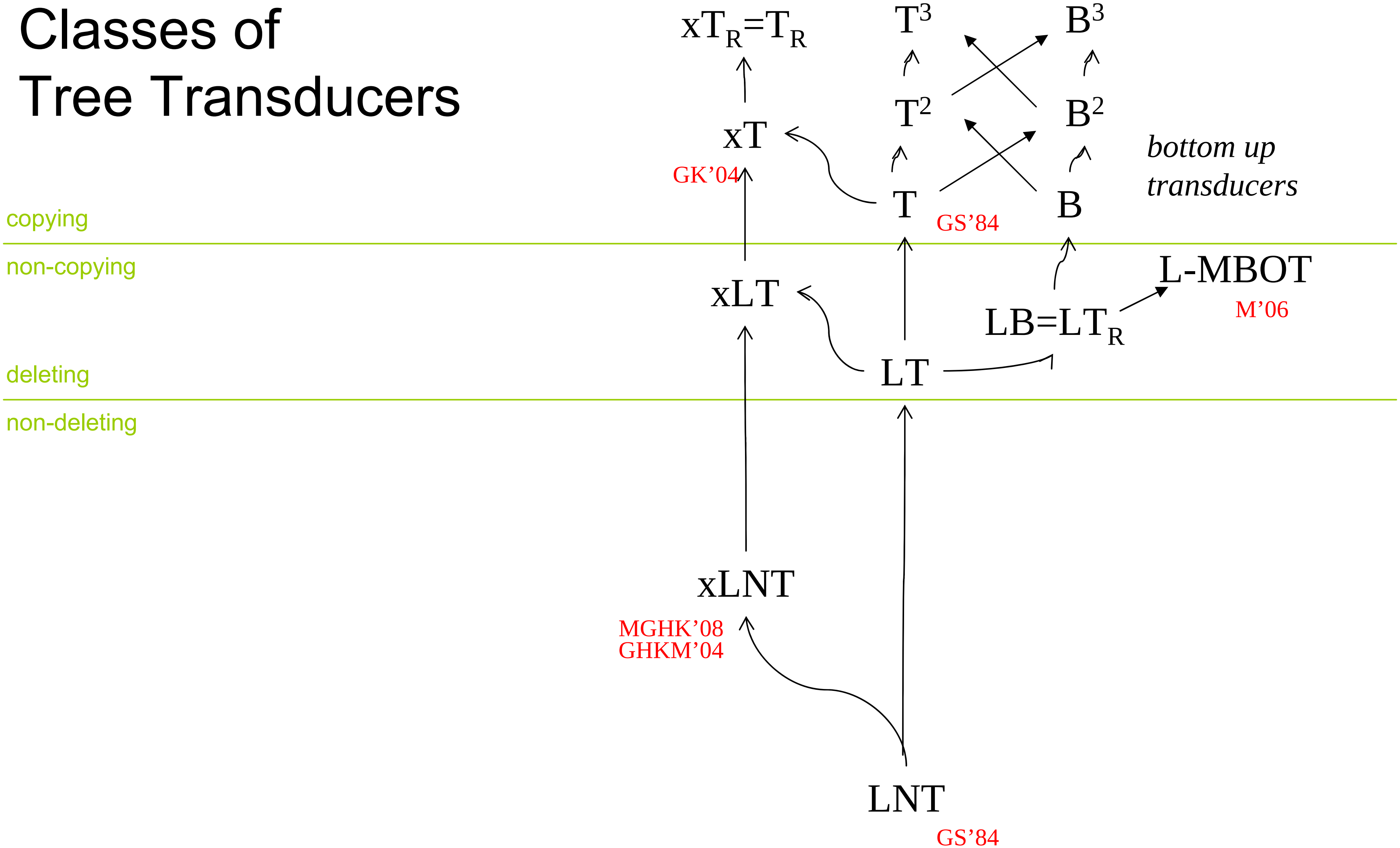
Classes of Tree Transducers



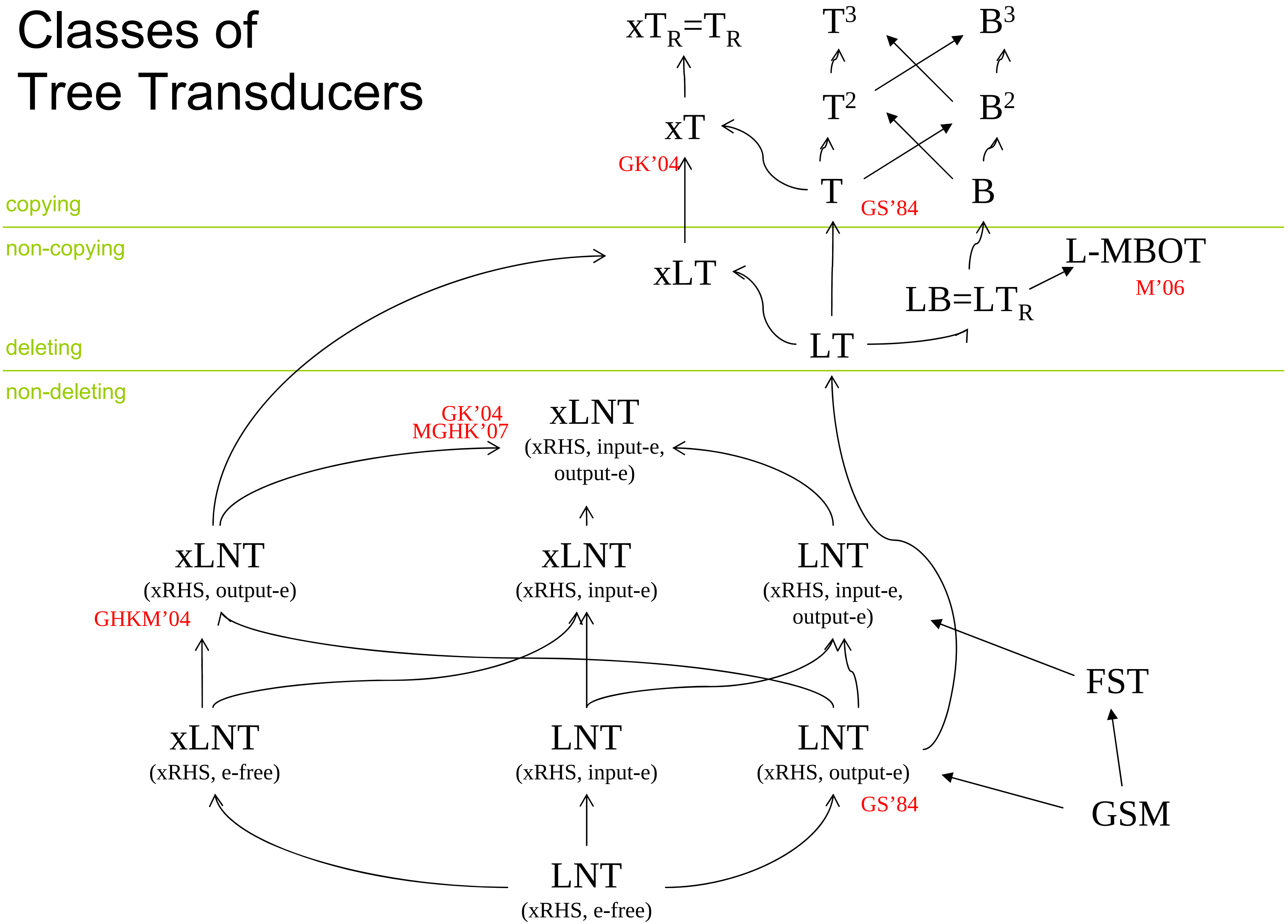
x-tended rule

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

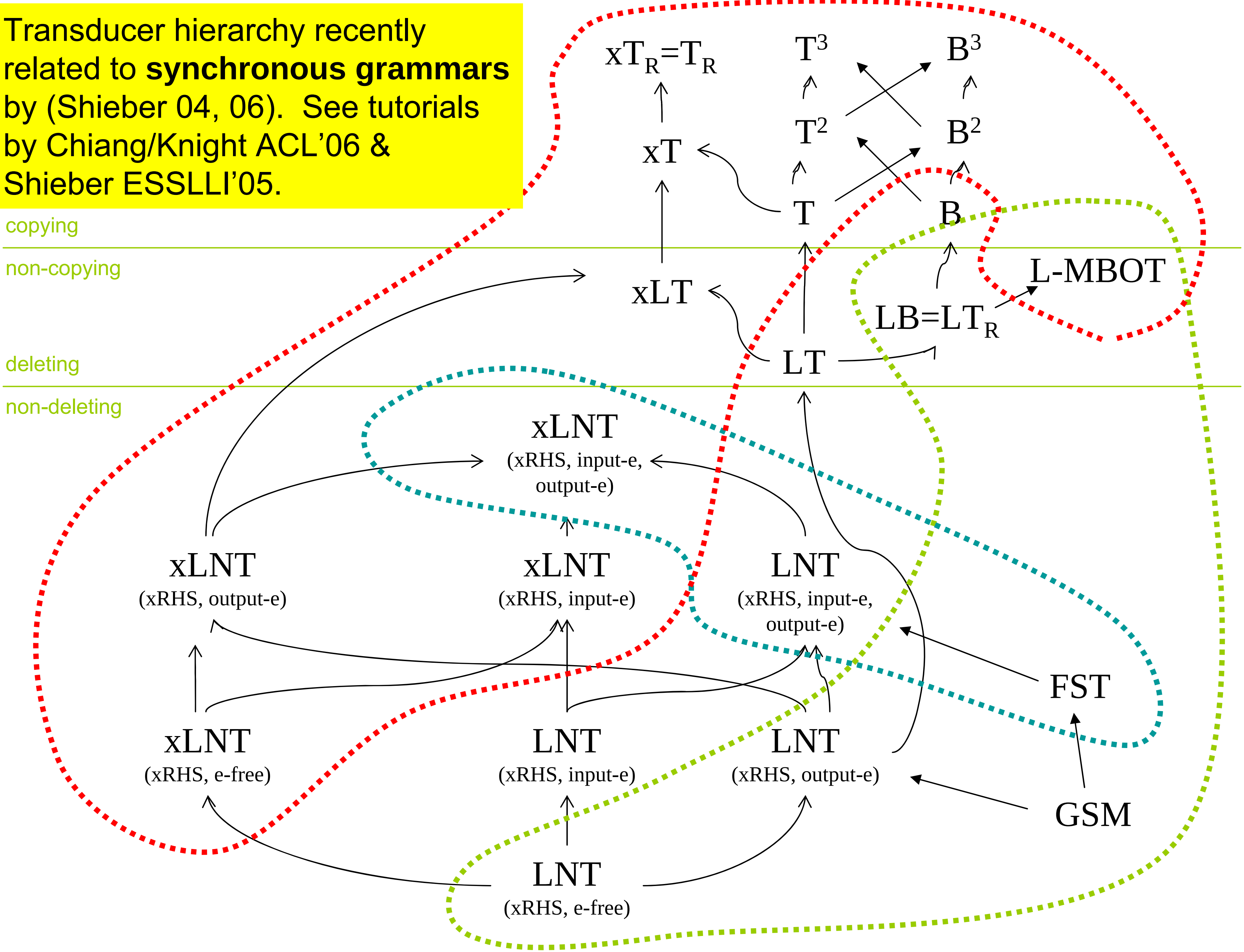
Classes of Tree Transducers



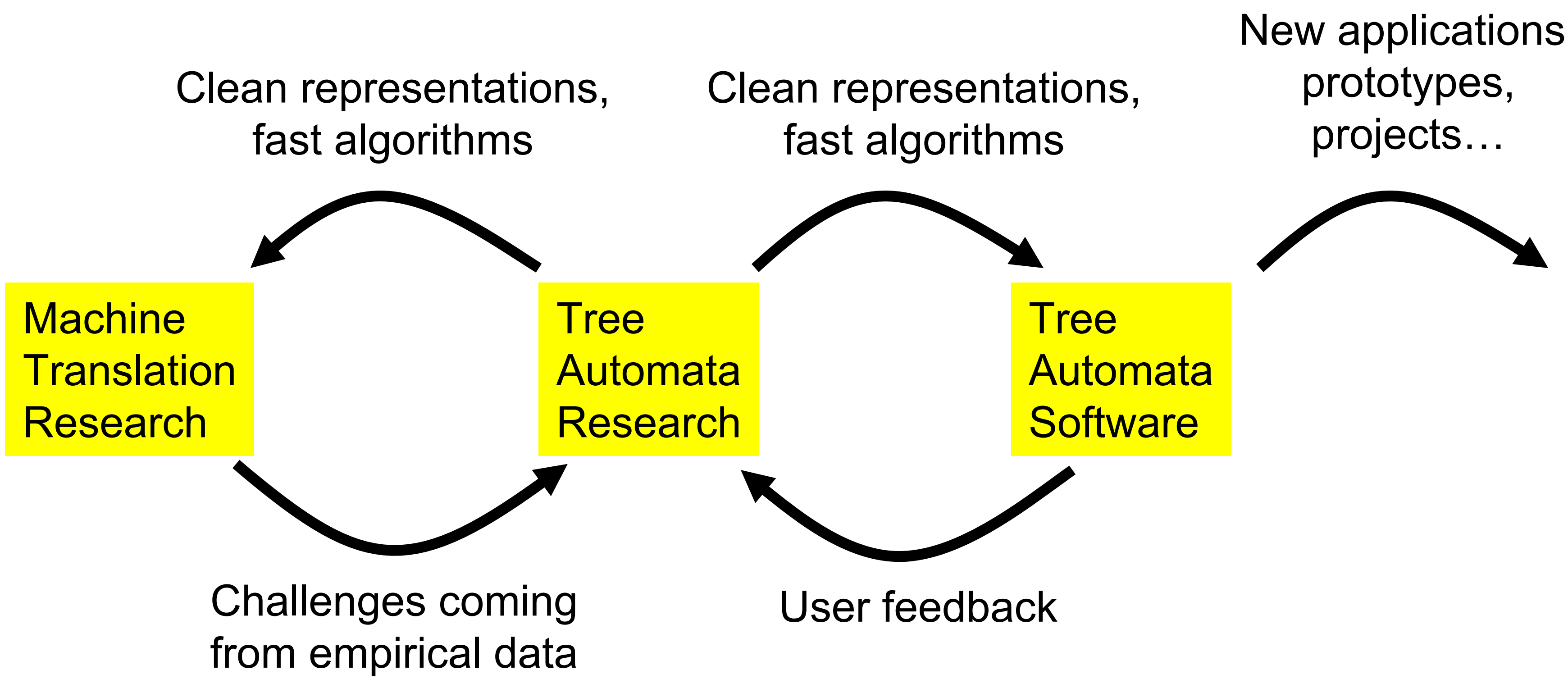
Classes of Tree Transducers



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



Research Synergy



This Is Interdisciplinary Research

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

…