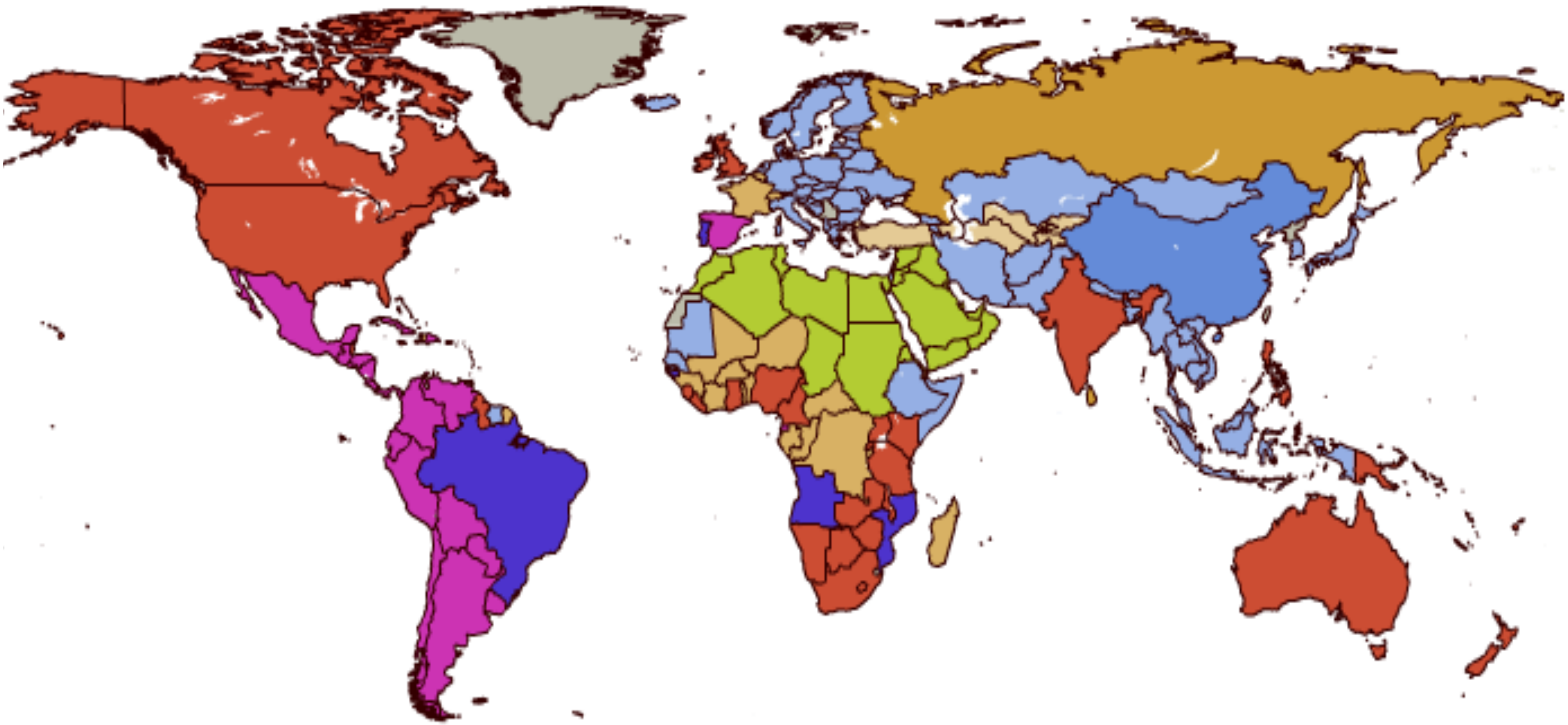
Statistical Machine Translation in the Big Data Era

Yang Liu
Tsinghua University

Part 1: Introduction

Natural Languages are Different



Natural Languages are Different

I love you

Я люблю тебя

我爱你

당신을 사랑합니다

Eu te amo

Je t'aime

אני אוהב אותך

من شما را دوست دارم

Tôi yêu bạn

Ich liebe dich

Te quiero

Miluji tě

Ti amo

ผมรักคุณ

わたしは、あなたを愛しています

Ik hou van je

Jag älskar dig

by Google Translate

Machine Translation

布什	与	沙龙	举行	了	会谈
bushi	yu	shalong	juxing	le	huitan

Bush held a talk with Sharon

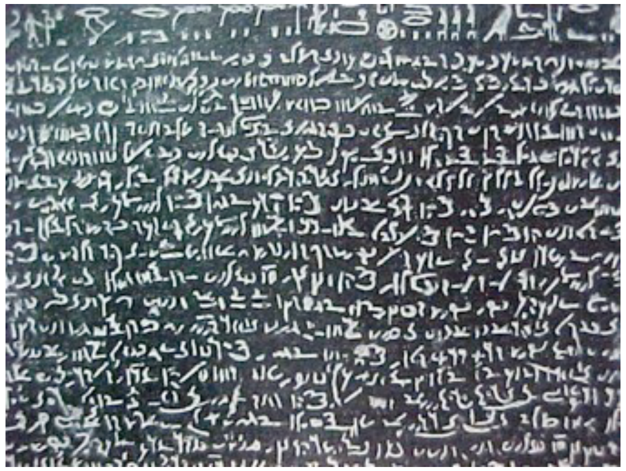
Rosetta Stone

adapted from Adam Lopez's slides

Rosetta Stone

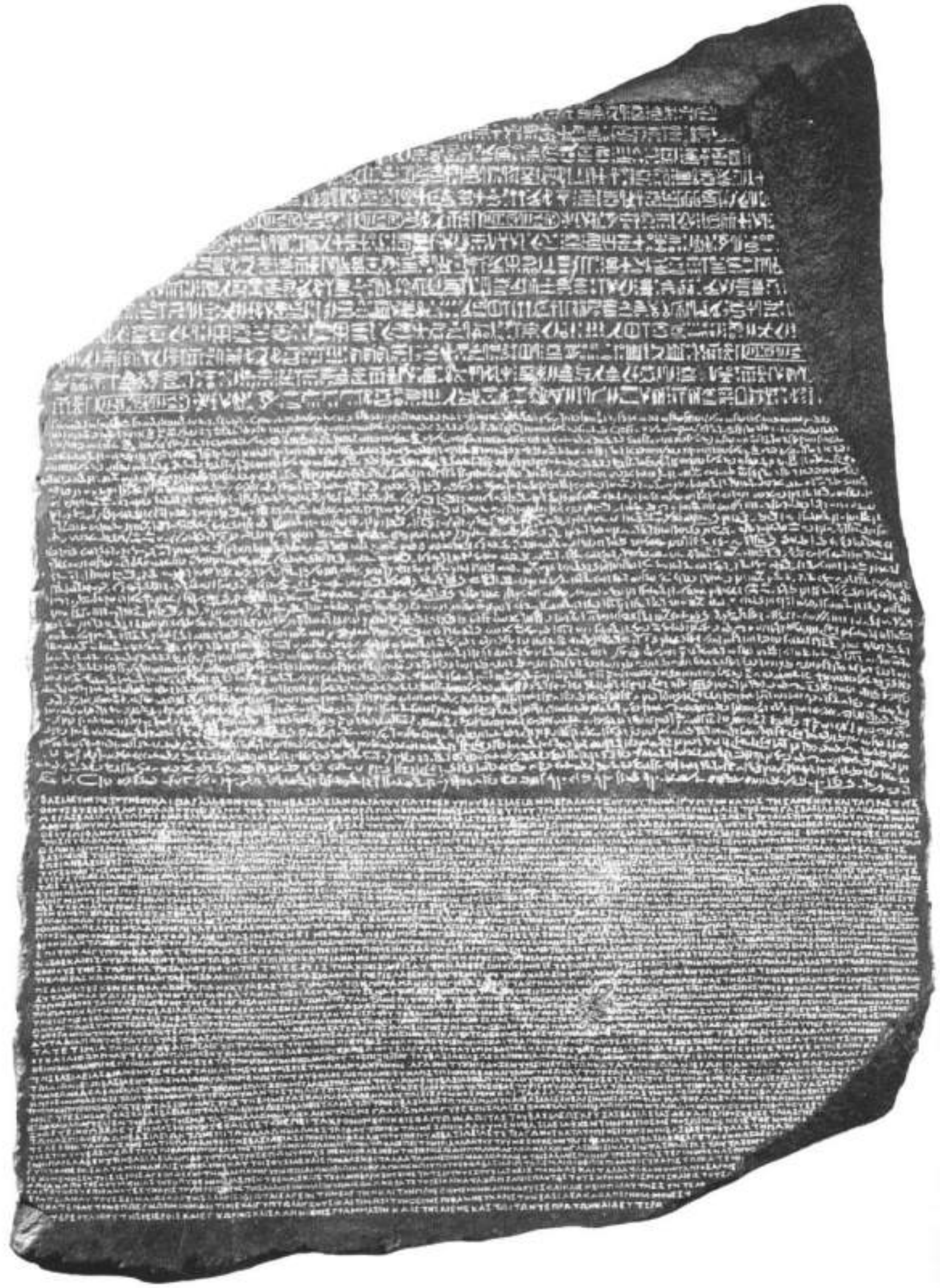
adapted from Adam Lopez's slides

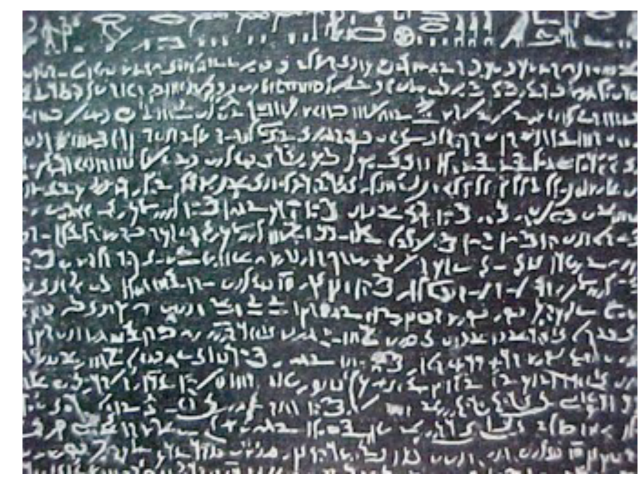
Rosetta Stone

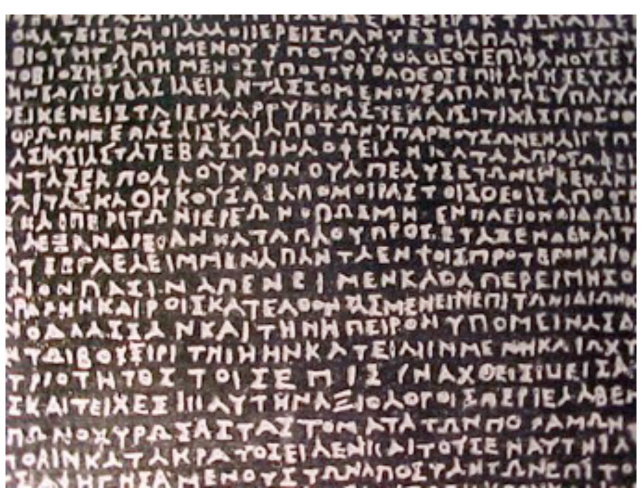


adapted from Adam Lopez's slides

Rosetta Stone







adapted from Adam Lopez's slides

Learning to Translate

Garcia y asociados .

los clients y los asociados son enemigos .

sus asociados no son fuertes .

Garcia y sus asociados no son enemigos .

adapted from Kevin Knight's slides

Learning to Translate

Garcia y asociados .

Garcia and associates .

los clients y los asociados son enemigos .

the clients and the associates are enemies .

sus asociados no son fuertes .

his associates are not strong .

Garcia y sus asociados no son enemigos .

adapted from Kevin Knight's slides

Learning to Translate

Garcia y asociados .

Garcia and associates .

los clients y los asociados son enemigos .

the clients and the associates are enemies .

sus asociados no son fuertes .

his associates are not strong .

Garcia y sus asociados no son enemigos .

adapted from Kevin Knight's slides

Learning to Translate

Garcia y asociados .

Garcia and associates .

los clients y los asociados son enemigos .

the clients and the associates are enemies .

sus asociados no son fuertes .

his associates are not strong .

Garcia y sus asociados no son enemigos .

adapted from Kevin Knight's slides

Learning to Translate

Garcia y asociados .

Garcia and associates .

los clients y los asociados son enemigos .

the clients and the associates are enemies .

sus asociados no son fuertes .

his associates are not strong .

Garcia y sus asociados no son enemigos .

adapted from Kevin Knight's slides

Learning to Translate

Garcia y asociados .

Garcia and associates .

los clients y los asociados son enemigos .

the clients and the associates are enemies .

sus asociados no son fuertes .

his associates are not strong .

Garcia y sus asociados no son enemigos .

adapted from Kevin Knight's slides

Learning to Translate

Garcia y asociados .

Garcia and associates .

los clients y los asociados son enemigos .

the clients and the associates are enemies .

sus asociados no son fuertes .

his associates are not strong .

Garcia y sus asociados no son enemigos .

adapted from Kevin Knight's slides

Learning to Translate

Garcia y asociados .

Garcia and associates .

los clients y los asociados son enemigos .

the clients and the associates are enemies .

sus asociados no son fuertes .

his associates are not strong .

Garcia y sus asociados no son enemigos .

adapted from Kevin Knight's slides

Learning to Translate

Garcia y asociados .

Garcia and associates .

los clients y los asociados son enemigos .

the clients and the associates are enemies .

sus asociados no son fuertes .

his associates are not strong .

Garcia y sus asociados no son enemigos .

adapted from Kevin Knight's slides

Learning to Translate

Garcia y asociados .

Garcia and associates .

los clients y los asociados son enemigos .

the clients and the associates are enemies .

sus asociados no son fuertes .

his associates are not strong .

Garcia y sus asociados no son enemigos .

adapted from Kevin Knight's slides

Learning to Translate

Garcia y asociados .

Garcia and associates .

los clients y los asociados son enemigos .

the clients and the associates are enemies .

sus asociados no son fuertes .

his associates are not strong .

Garcia y sus asociados no son enemigos .

adapted from Kevin Knight's slides

Learning to Translate

Garcia y asociados .

Garcia and associates .

los clients y los asociados son enemigos .

the clients and the associates are enemies .

sus asociados no son fuertes .

his associates are not strong .

Garcia y sus asociados no son enemigos .

adapted from Kevin Knight's slides

Learning to Translate

Garcia y asociados .

Garcia and associates .

los clients y los asociados son enemigos .

the clients and the associates are enemies .

sus asociados no son fuertes .

his associates are not strong .

Garcia y sus asociados no son enemigos .

adapted from Kevin Knight’s slides

Learning to Translate

Garcia y asociados .

Garcia and associates .

los clients y los asociados son enemigos .

the clients and the associates are enemies .

sus asociados no son fuertes .

his associates are not strong .

Garcia y sus asociados no son enemigos .

Spanish	English
Garcia	Garcia
y	and
asociados	associates
.	.
los	the
clients	clients
son	are
enemigos	enemies
sus	his
no	not
fuertes	strong

adapted from Kevin Knight’s slides

Learning to Translate

Garcia y asociados .
Garcia and associates .

los clients y los asociados son enemigos .
the clients and the associates are enemies .

sus asociados no son fuertes .
his associates are not strong .

Garcia y sus asociados no son enemigos .
Garcia and his associates are not enemies .

Spanish	English
Garcia	Garcia
y	and
asociados	associates
.	.
los	the
clients	clients
son	are
enemigos	enemies
sus	his
no	not
fuertes	strong

adapted from Kevin Knight's slides

MT Approaches

Q: How machines learn translation knowledge?

MT Approaches

Q: How machines learn translation knowledge?

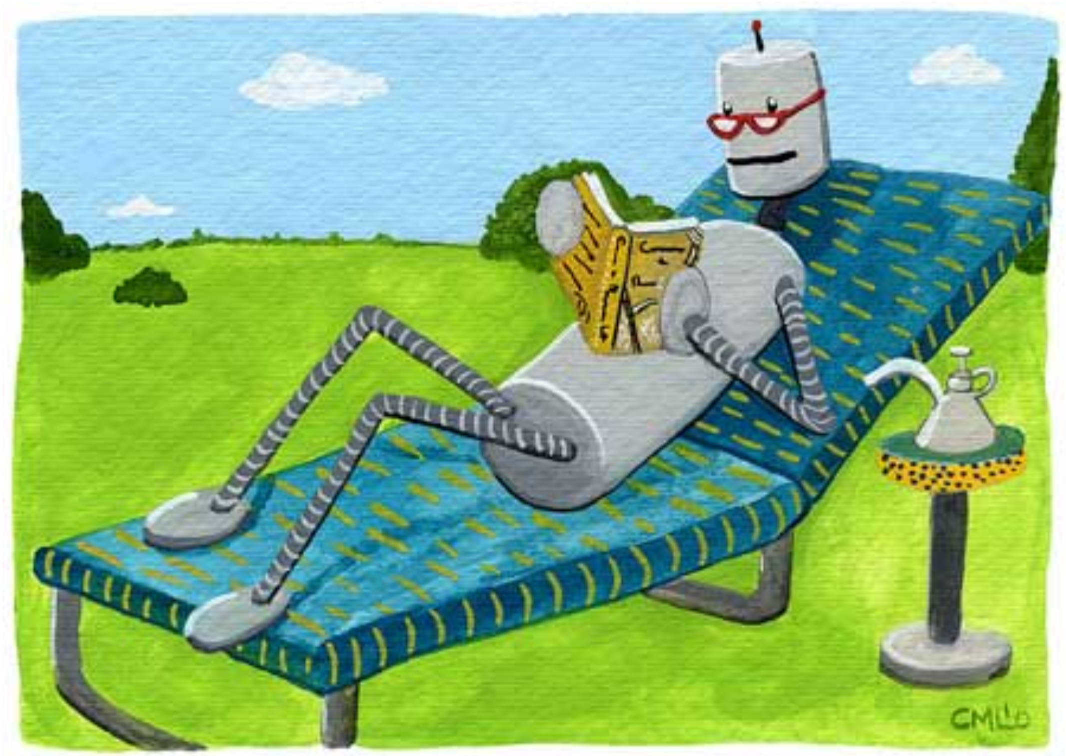
MT Approaches

Q: How machines learn translation knowledge?

rule-based MT

MT Approaches

Q: How machines learn translation knowledge?



rule-based MT

MT Approaches

Q: How machines learn translation knowledge?

rule-based MT

data-driven MT

Data-driven MT

data-driven MT

Data-driven MT

data-driven MT

Example-based MT

(Nagao, 1984)

Data-driven MT

data-driven MT

Example-based MT

(Nagao, 1984)

Statistical MT

(Brown et al., 1993)

Statistical MT

> Statistical machine translation is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora.
>
> -- Wikipedia

Statistical MT

Statistical machine translation is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora.

-- Wikipedia

Modeling

Tell machine how to translate

Learning

Machine learns translation knowledge from data

Decoding

Machine translates text using learned knowledge

Big Data

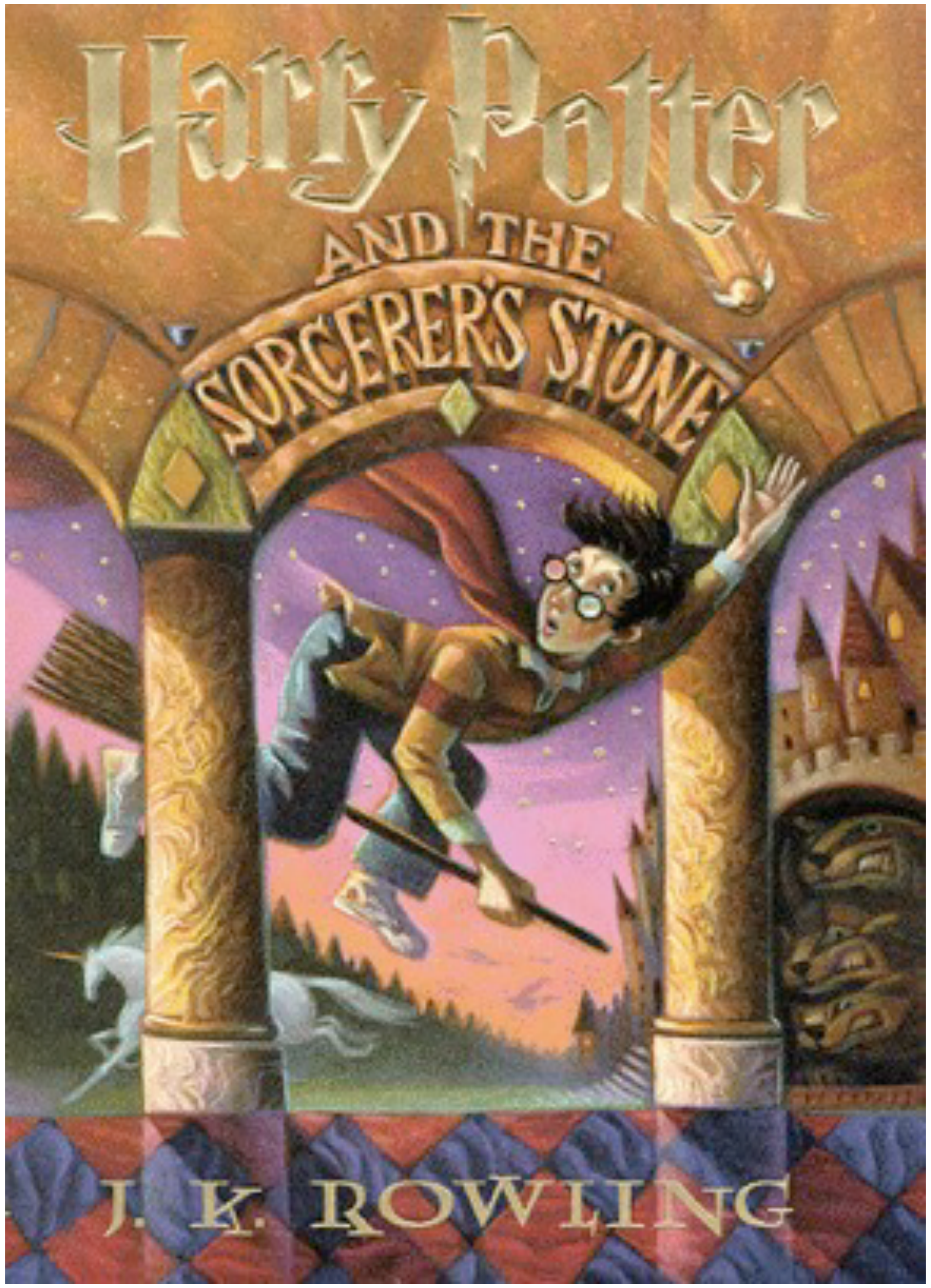
- An explosion of data across numerous languages
  - 60,000 news websites per day
  - 2 million blog posts per day
  - 175 million Tweets per day
  - 293,000 new Facebook status updates per minute
- 2 billion Internet users speaking 6000 languages

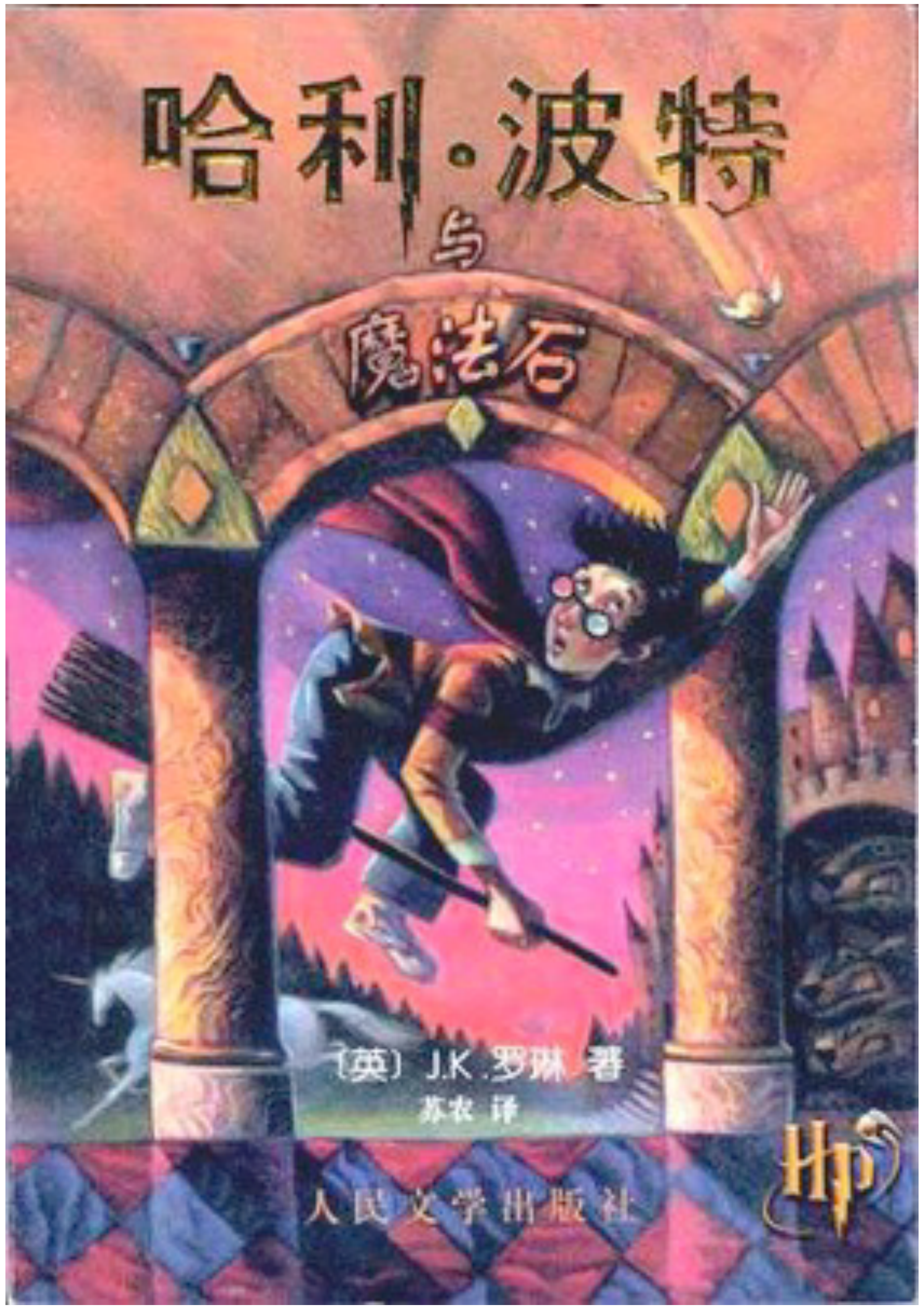
Data for SMT

欢迎来到联合国，您的世界！
搜索联合国网站
搜索
عربي 中文 English Français Русский Español
联 合 国
我联合国人民，团结起来，追求更美好的世界！
和平与安全
发展
人权
人道主义事务
国际法
联合国：索马里饥荒结束 但危机并未过去
相关链接
电台报道 • 相关图片 • 中文视频
关注非洲之角干旱 • 人道主义事务
※ 联合国微博最新报道
最新动态
RSS
2012年2月04日 中国和俄罗斯就安理会关于叙利亚问题决议草案投否决票
2012年2月03日 联合国：索马里饥荒结束 但危机并未过去
2012年2月03日 国际法院：意大利法院要求德国赔偿二战受害者侵犯了德国的国家豁免权
你的联合国
联合国一览
联合国宪章
组织与结构
会员国
加强联合国*
信息中心
常见问题
主要机关
大会
第六十六届会议主席
安全理事会
每月轮值主席
经济及社会理事会
第六十八任主席
托管理事会
国际法院
更多机构 >>
联合国与……
民间社会
秘书长
秘书处
发言人
聚焦
北非局势
气候变化
千年发展目标
中东局势 加沙 | 巴勒斯坦 | 黎巴嫩
联合国与商业界
妇女、和平与安全
日常议题
艾滋病、环境、青年、原子能……
资源与服务
文件
图书馆
地图
出版物
工作
网上书店
采购
实习方案
邮票
数据中心
媒体核证
参观联合国
会议、活动与展览
联大第66届会议的主要会议闭幕

Welcome to the United Nations. It's your world.
Search UN Website
Go
عربي 中文 English Français Русский Español
UNITED NATIONS
We the peoples... A stronger UN for a better world.
Peace and Security
Development
Human Rights
Humanitarian Affairs
International Law
Somalia: UN says famine is over, but warns action is needed to forestall new crisis
Related
S-G's statement • Photos • Radio
OCHA • FAO • WFP
In the News
RSS
04/02/2012 Syria: Ban voices deep regret after Security Council fails to agree on resolution
04/02/2012 Civilian casualty numbers in Afghanistan rise again, UN reports
03/02/2012 UN says Somali famine over, but warns action needed to forestall new crisis
Your United Nations
UN at a Glance
UN Charter
Structure and Organization
Member States
Strengthening the UN
UN Information Centres
Events Calendar
Frequently Asked Questions
Main Bodies
General Assembly
President
Security Council
President
Economic & Social Council
President
Trusteeship Council
International Court of Justice
More info >>
The UN and . . .
Civil Society
Global Compact
Secretary-General
Secretariat
Spokesperson
In Focus
Winds of change: North Africa and the Middle East
Climate Change
Millennium Development Goals
Situation in the Middle East
UN-Business Partnerships
Women, Peace and Security
Global Issues
Africa... Environment... Women
Resources and Services
Documents
Library
Maps
Publications
Employment
Bookshop
Procurement
Internships
Stamps
Databases
Media Accreditation
Visiting UN Headquarters
Conferences, Meetings, Events
UN General Assembly 66th session

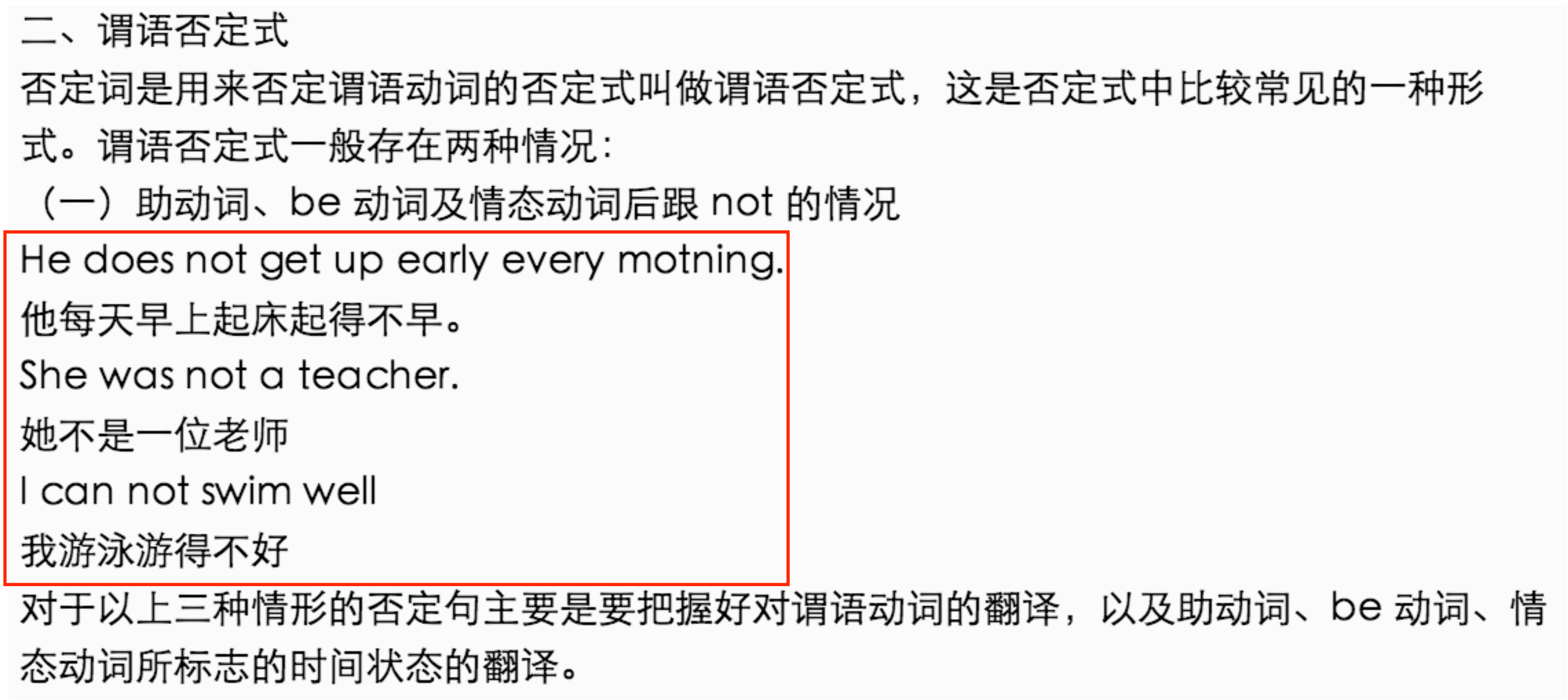
Data for SMT





from Adam Lopez's slides

Data for SMT


二、谓语否定式
否定词是用来否定谓语动词的否定式叫做谓语否定式，这是否定式中比较常见的一种形式。谓语否定式一般存在两种情况：
（一）助动词、be 动词及情态动词后跟 not 的情况
He does not get up early every motning.
他每天早上起床起得不早。
She was not a teacher.
她不是一位老师
I can not swim well
我游泳游得不好
对于以上三种情形的否定句主要是要把握好对谓语动词的翻译，以及助动词、be 动词、情态动词所标志的时间状态的翻译。

Data for SMT

中新网10月5日电 据外电报道，阿富汗官员5日称，驻阿富汗北约部队4日晚对阿富汗东部一地区进行了空袭，造成包括3名未成年人在内至少5名平民死亡。

楠格哈尔省警方发言人马什莱齐瓦尔(Hazrat Hussain Mashreqiwal)说，这5名年龄介于12到20岁的平民，在该省首府贾拉拉巴德郊外萨拉查(Saracha)地区捕鸟时遭到北约空袭致死。

北约发言人称知道此次空袭，但是不愿证实死亡人数。

马士莱齐瓦尔说，昨天(4日)晚11点左右，5名12到20岁的平民在距离贾拉拉巴德市中心约8公里的地区，拿气枪捕鸟时，遭到外国部队空袭死亡。他们的尸体已经被送到中心医院。

楠格哈尔省政府发言人亚布杜拉塞(Ahmad Zia Abdulzai)证实了此次空袭事件。

楠格哈尔省教育部发言人辛瓦利(Mohammad Atif Shinwari)说，3名在空袭中死亡的平民是学童，2人为兄弟。

Data for SMT

中新网10月5日电 据外电报道，阿富汗官员5日称，驻阿富汗北约部队4日晚对阿富汗东部一地区进行了空袭，造成包括3名未成年人在内至少5名平民死亡。

楠格哈尔省警方发言人马什莱齐瓦尔(Hazrat Hussain Mashreqiwal)说，这5名年龄介于12到20岁的平民，在该省首府贾拉拉巴德郊外萨拉查(Saracha)地区捕鸟时遭到北约空袭致死。

北约发言人称知道此次空袭，但是不愿证实死亡人数。

马士莱齐瓦尔说，昨天(4日)晚11点左右，5名12到20岁的平民在距离贾拉拉巴德市中心约8公里的地区，拿气枪捕鸟时，遭到外国部队空袭死亡。他们的尸体已经被送到中心医院。

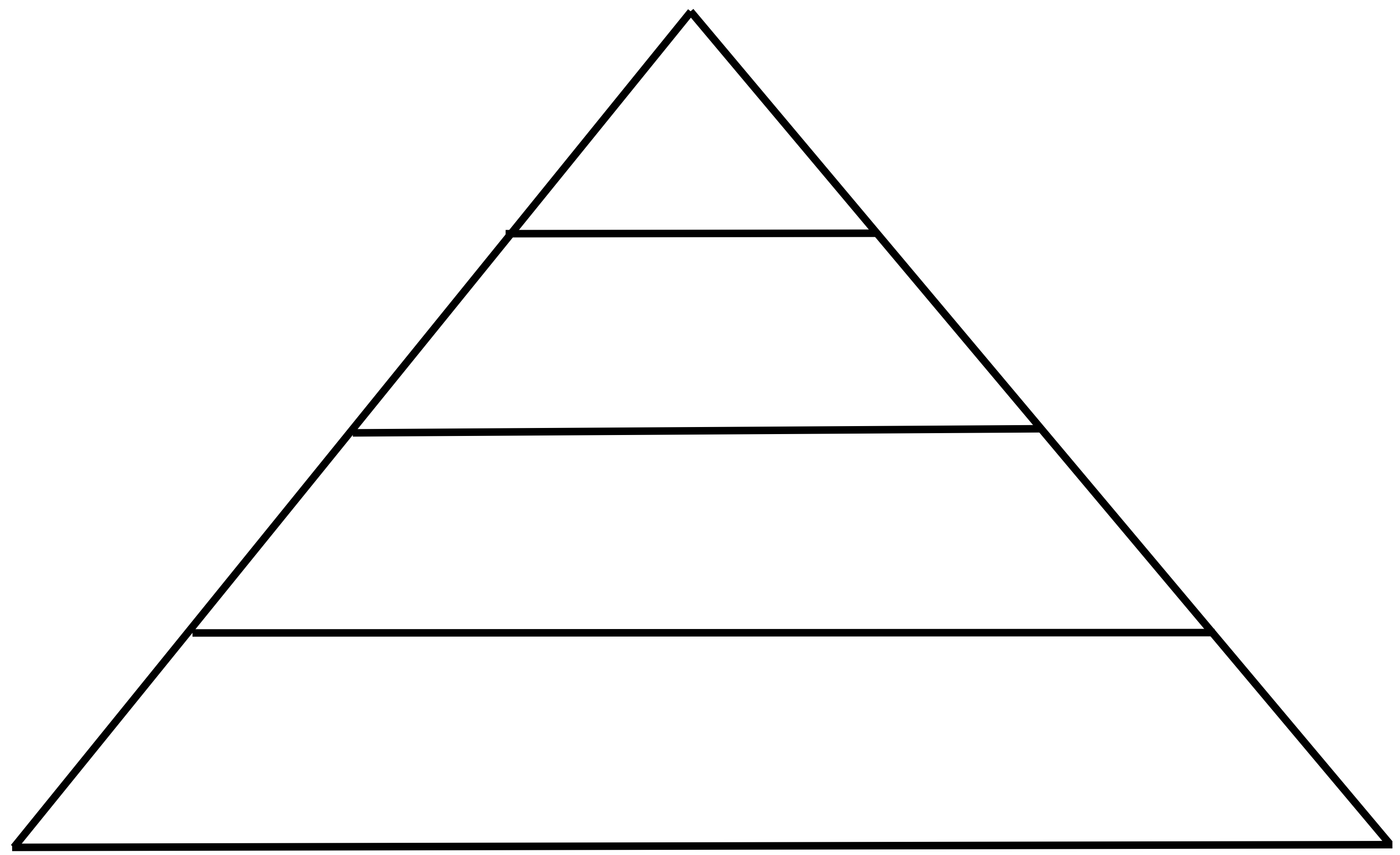
楠格哈尔省政府发言人亚布杜拉塞(Ahmad Zia Abdulzai)证实了此次空袭事件。

楠格哈尔省教育部发言人辛瓦利(Mohammad Atif Shinwari)说，3名在空袭中死亡的平民是学童，2人为兄弟。

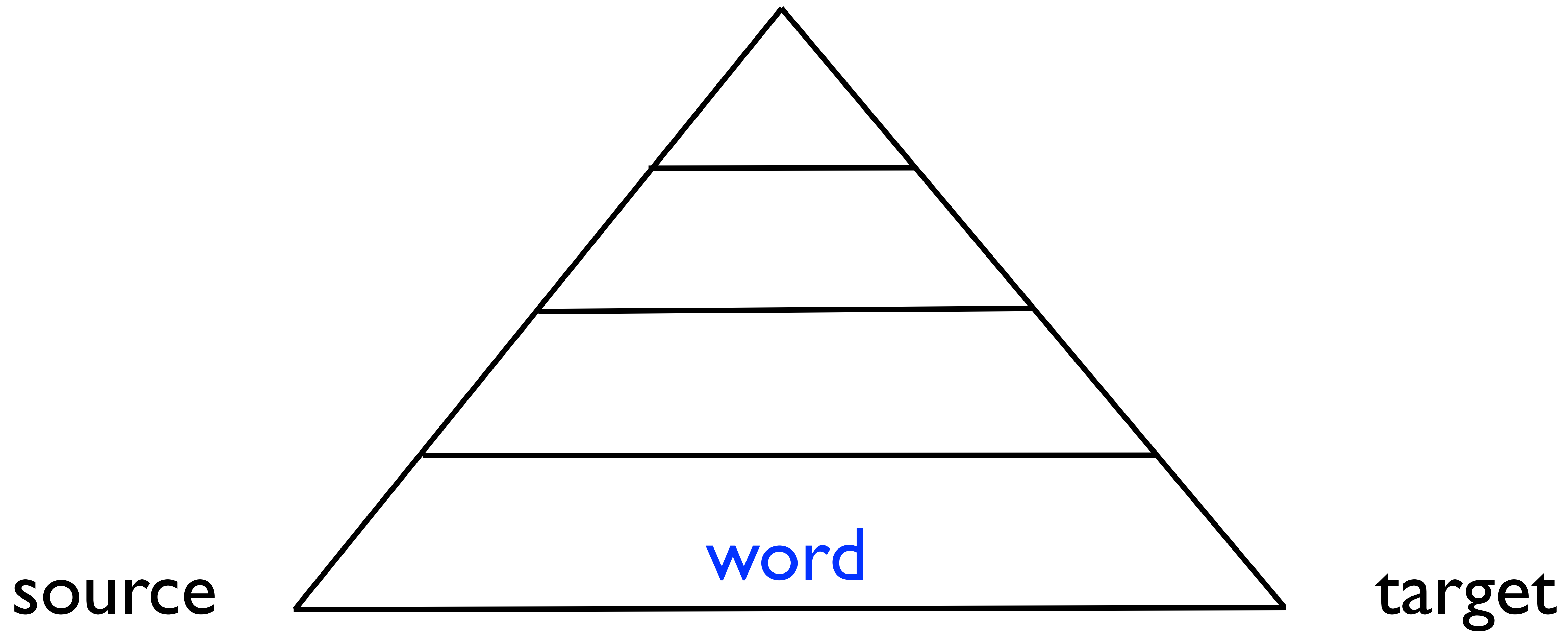
Data for SMT

訂座電話:3427 9228
11:30a.m.-2:00a.m.
彌敦道店
Nathan Road
628-636 彌敦道 638-664
Subway to cross
Nathan Road
請用隧道
橫過彌敦道
Watsons
屈臣氏
萬寧
mannings
SAMSUNG

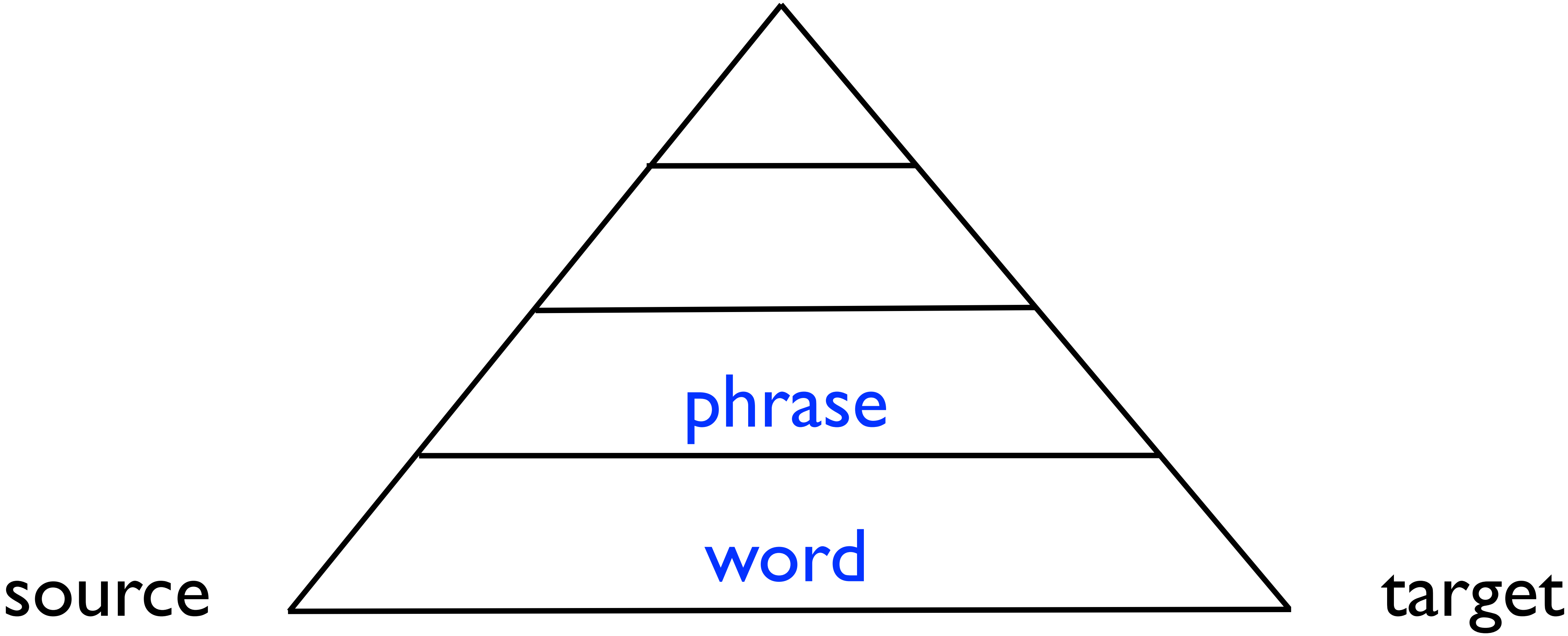
The SMT Pyramid

source  target

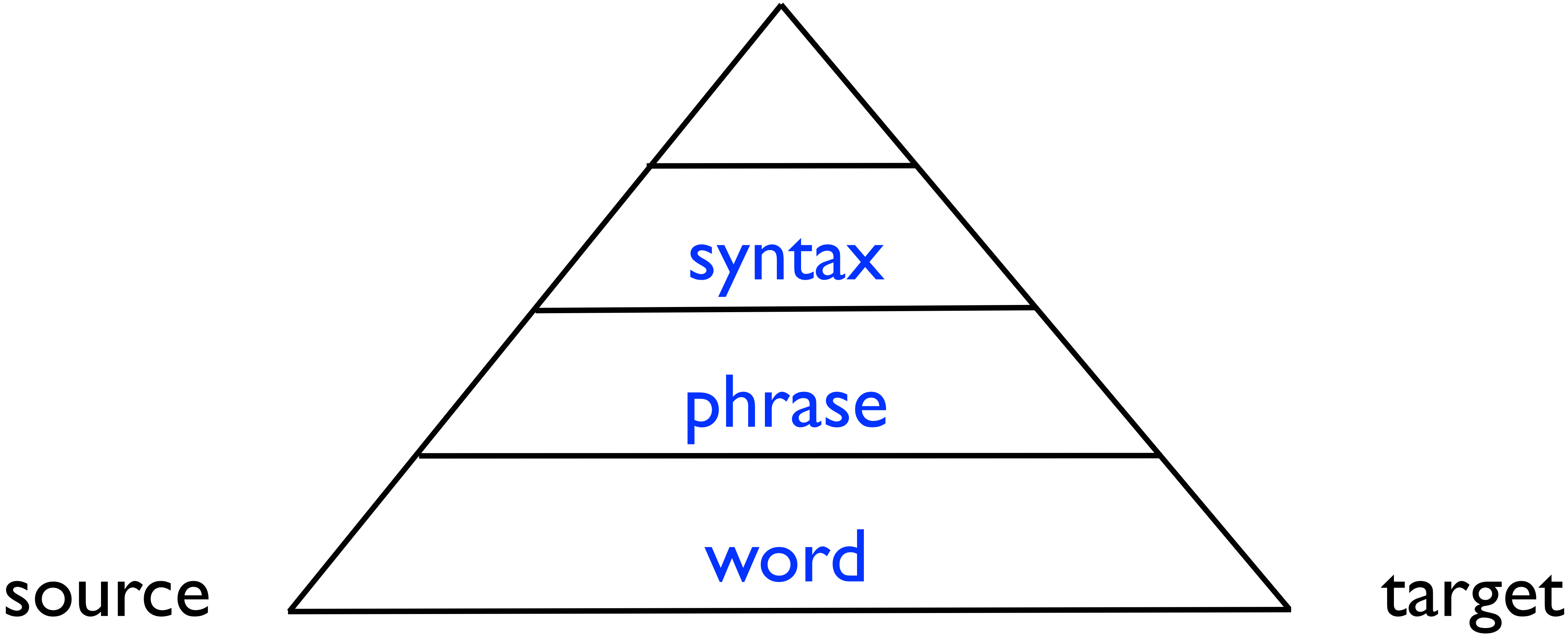
The SMT Pyramid



The SMT Pyramid



The SMT Pyramid



Part 2: Word-based MT

The Origin of SMT

Warren Weaver

When I look at an article in Russian, I say: "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

Weaver (1955)

IBM and Machine Translation

Fred Jelinek

Some of us started to wonder in the mid of 1980s whether our ASR methods could be applied to new fields. Bob Mercer and I ... came up with two: machine translation and stock market modeling.

Jelinek (2009)

The Noisy-Channel Model

The Noisy-Channel Model

$$P(E|F)$$

The Noisy-Channel Model

$$P(E|F) = \frac{P(E) \times P(F|E)}{P(F)}$$

The Noisy-Channel Model

$$P(E|F) = \frac{P(E) \times P(F|E)}{P(F)}$$

$$\hat{E} = \text{argmax}_E \Big\{ P(E|F) \Big\}$$

The Noisy-Channel Model

$$P(E|F) = \frac{P(E) \times P(F|E)}{P(F)}$$

$$\hat{E} = \mathrm{argmax}_E \Big\{ P(E|F) \Big\}$$

$$= \mathrm{argmax}_E \Big\{ P(E) \times P(F|E) \Big\}$$

Language Model

$P(\text{“Bush held a talk with Sharon”})$

Language Model

$$P(\text{“Bush held a talk with Sharon”})$$

$$= P(\text{“Bush”}) \times$$

Language Model

$$P(\text{“Bush held a talk with Sharon”})$$

$$= P(\text{“Bush”})\times$$

$$P(\text{“held”}|\text{“Bush”})\times$$

Language Model

$$
\begin{aligned}
& P(\text{“Bush held a talk with Sharon”}) \\
= & P(\text{“Bush”})\times \\
& P(\text{“held”}|\text{“Bush”})\times \\
& P(\text{“a”}|\text{“Bush held”})\times
\end{aligned}
$$

Language Model

$$P(\text{“Bush held a talk with Sharon”})$$

$$= P(\text{“Bush”})\times$$

$$P(\text{“held”}|\text{“Bush”})\times$$

$$P(\text{“a”}|\text{“Bush held”})\times$$

$$P(\text{“talk”}|\text{“held a”})\times$$

Language Model

$$P(\text{“Bush held a talk with Sharon”})$$

$$= P(\text{“Bush”}) \times$$

$$P(\text{“held”} | \text{“Bush”}) \times$$

$$P(\text{“a”} | \text{“Bush held”}) \times$$

$$P(\text{“talk”} | \text{“held a”}) \times$$

$$P(\text{“with”} | \text{“a talk”}) \times$$

Language Model

$$
\begin{aligned}
& P(\text{“Bush held a talk with Sharon”}) \\
= \; & P(\text{“Bush”}) \times \\
& P(\text{“held”} | \text{“Bush”}) \times \\
& P(\text{“a”} | \text{“Bush held”}) \times \\
& P(\text{“talk”} | \text{“held a”}) \times \\
& P(\text{“with”} | \text{“a talk”}) \times \\
& P(\text{“Sharon”} | \text{“talk with”})
\end{aligned}
$$

Translation Model

Bush held a talk with Sharon

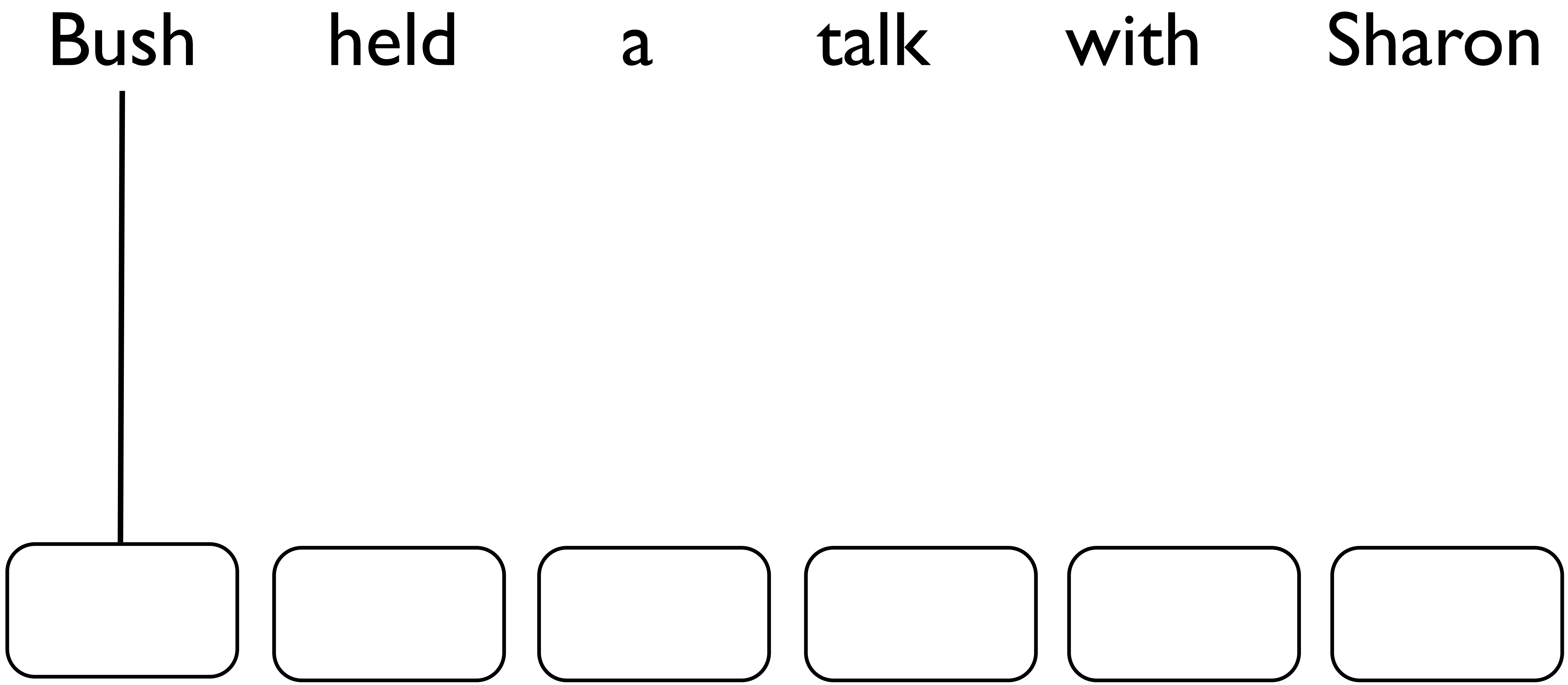
(Brown et al., 1993)

Translation Model

Bush held a talk with Sharon

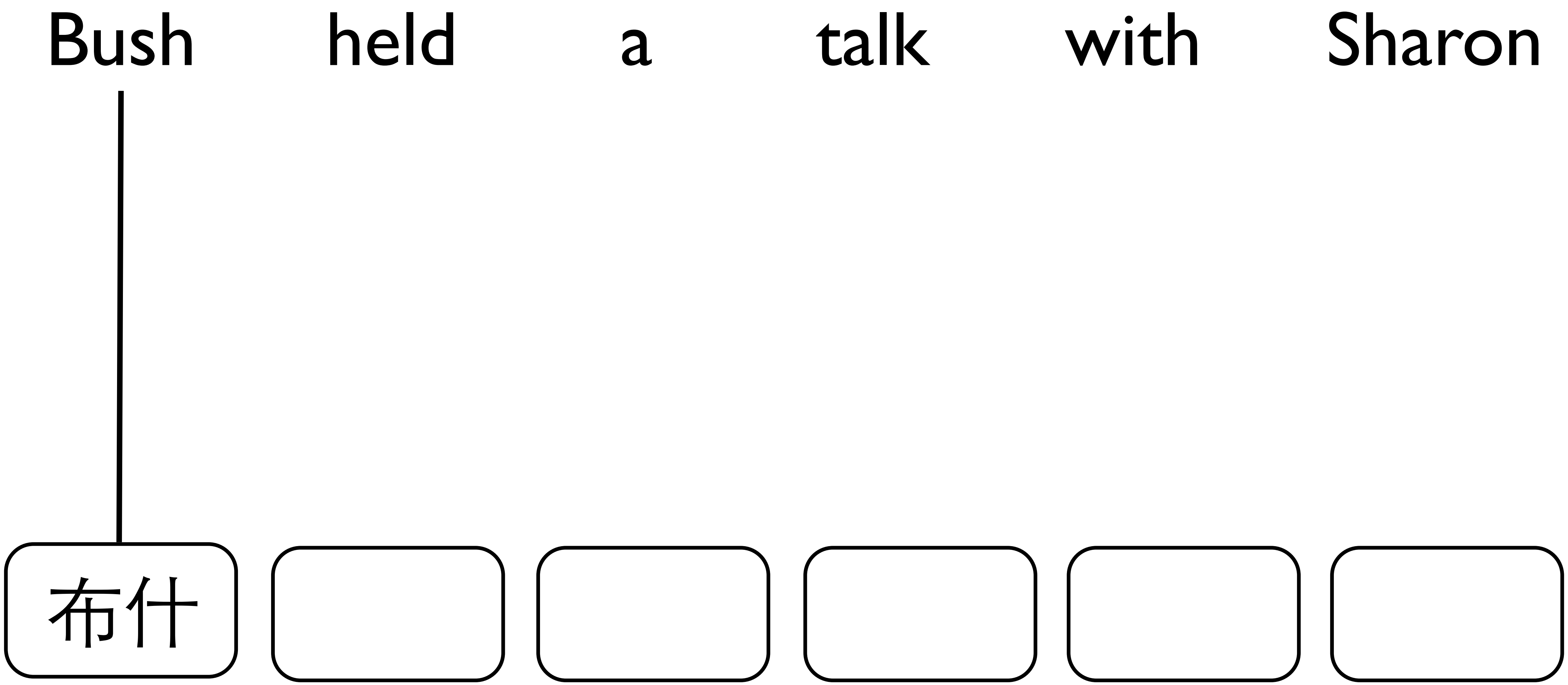
(Brown et al., 1993)

Translation Model



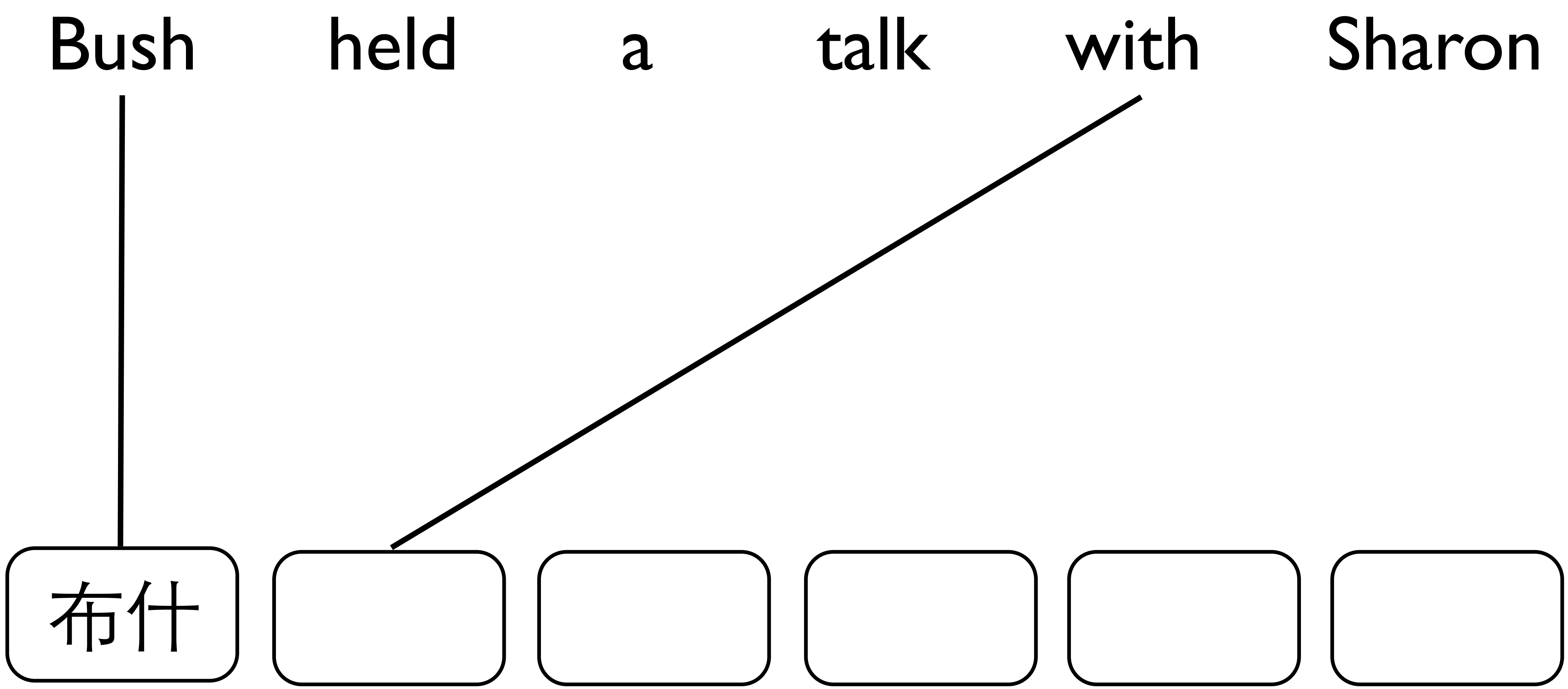
(Brown et al., 1993)

Translation Model



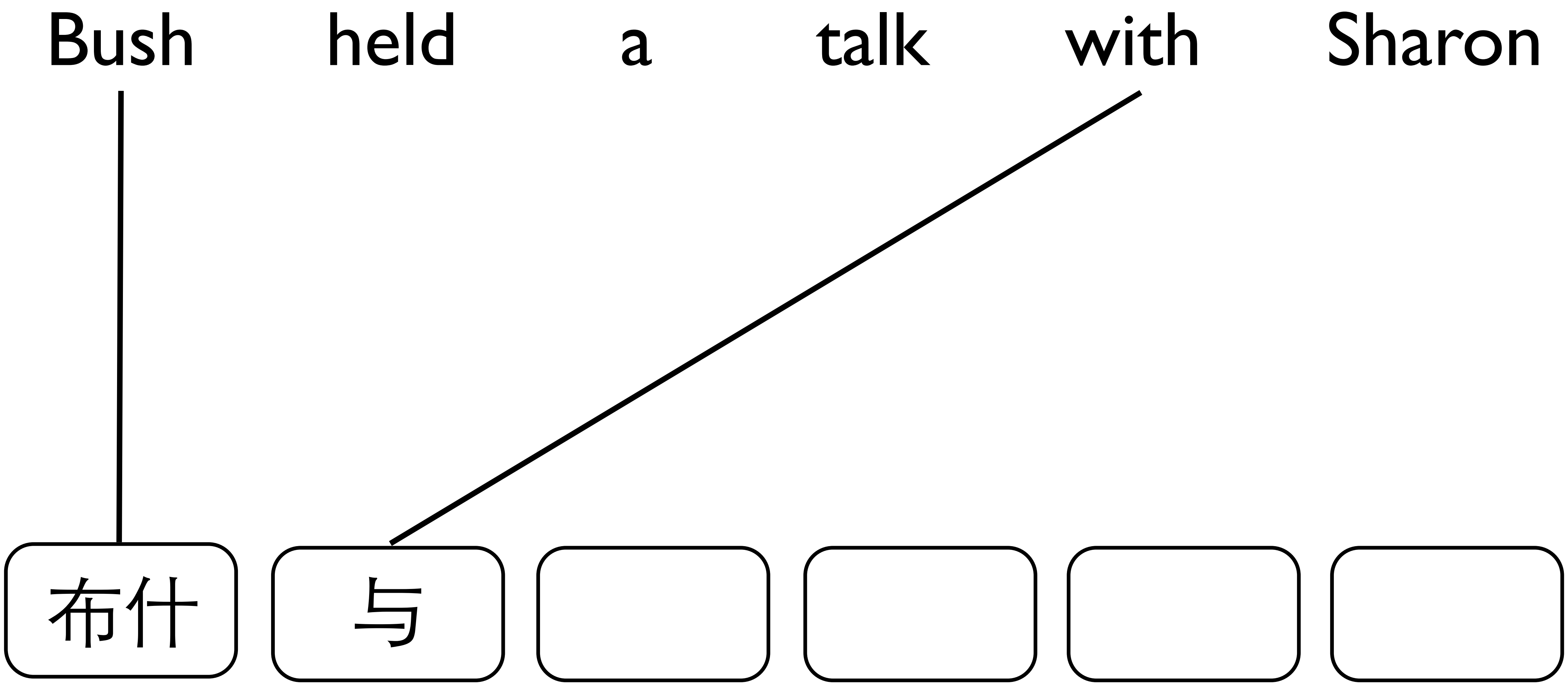
(Brown et al., 1993)

Translation Model



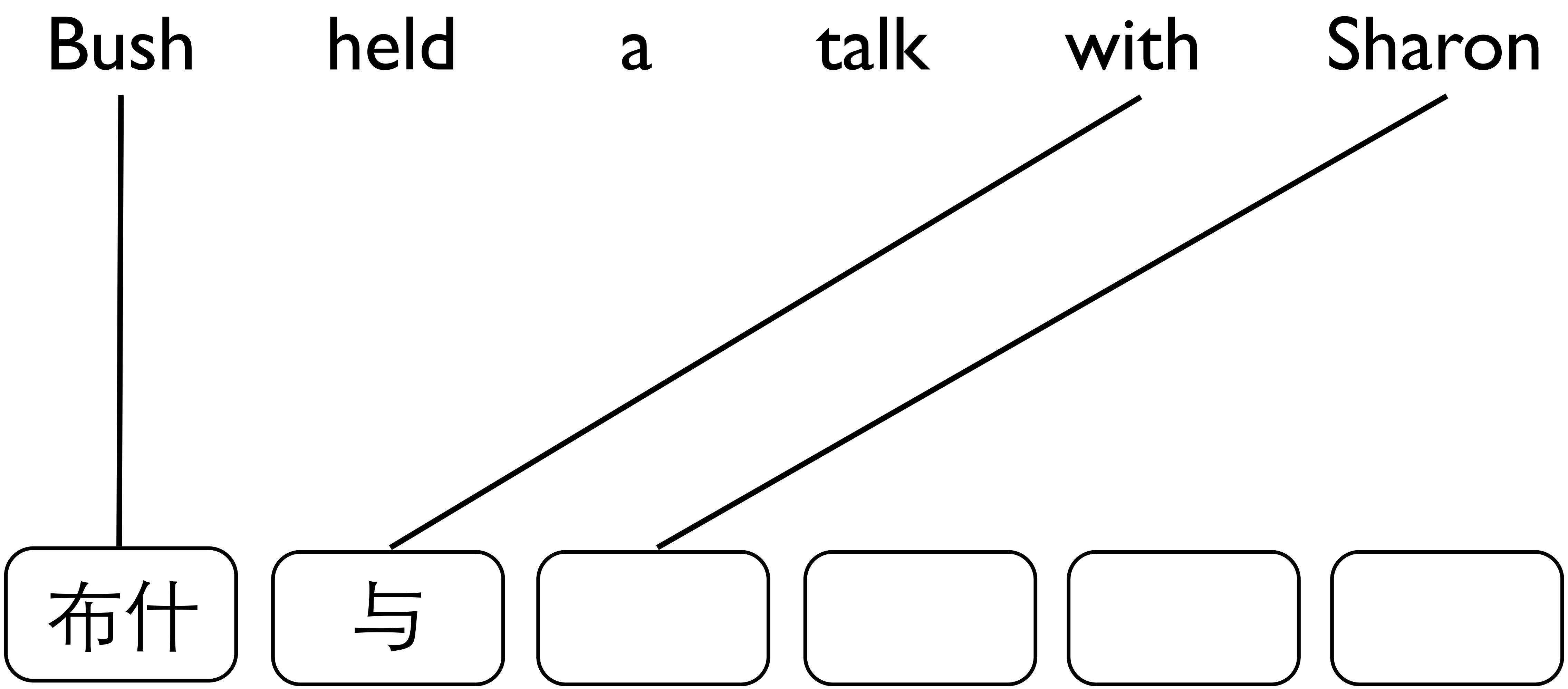
(Brown et al., 1993)

Translation Model



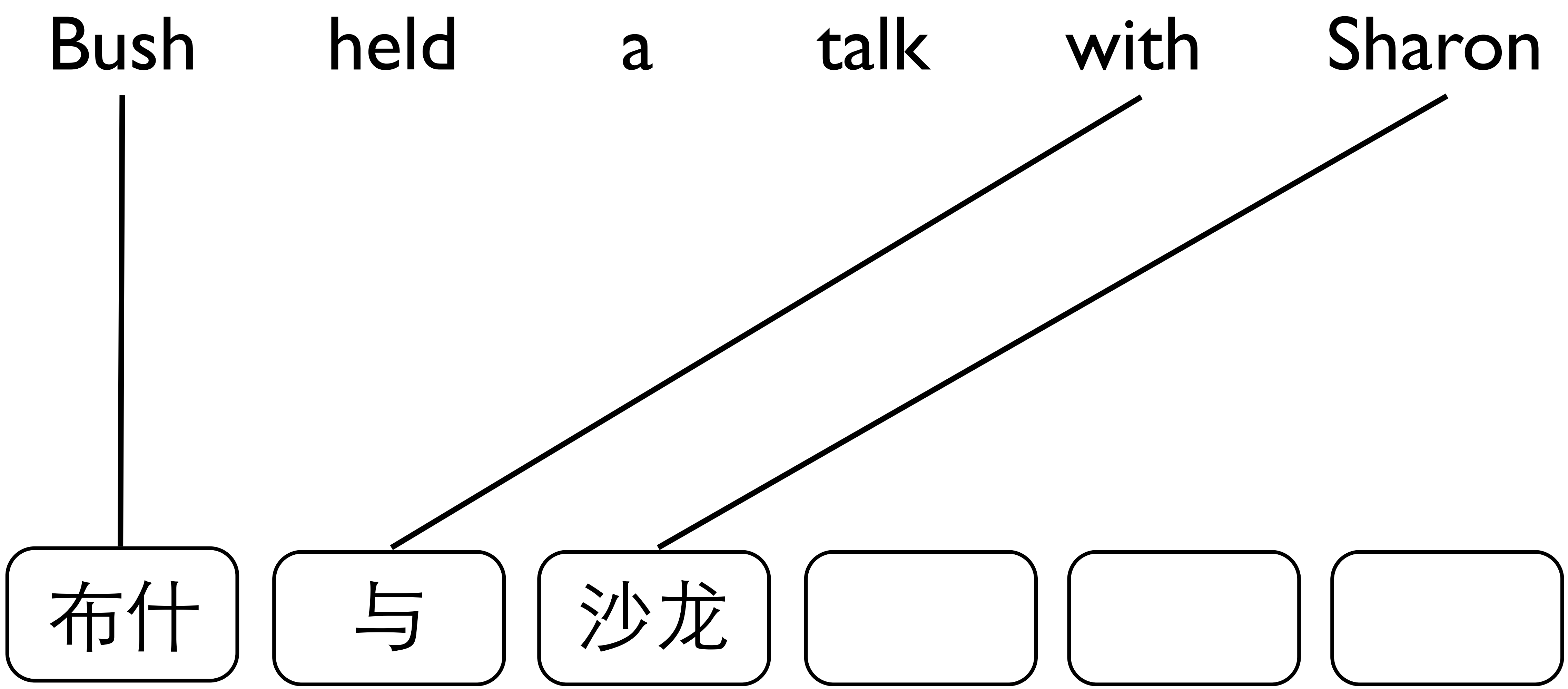
(Brown et al., 1993)

Translation Model



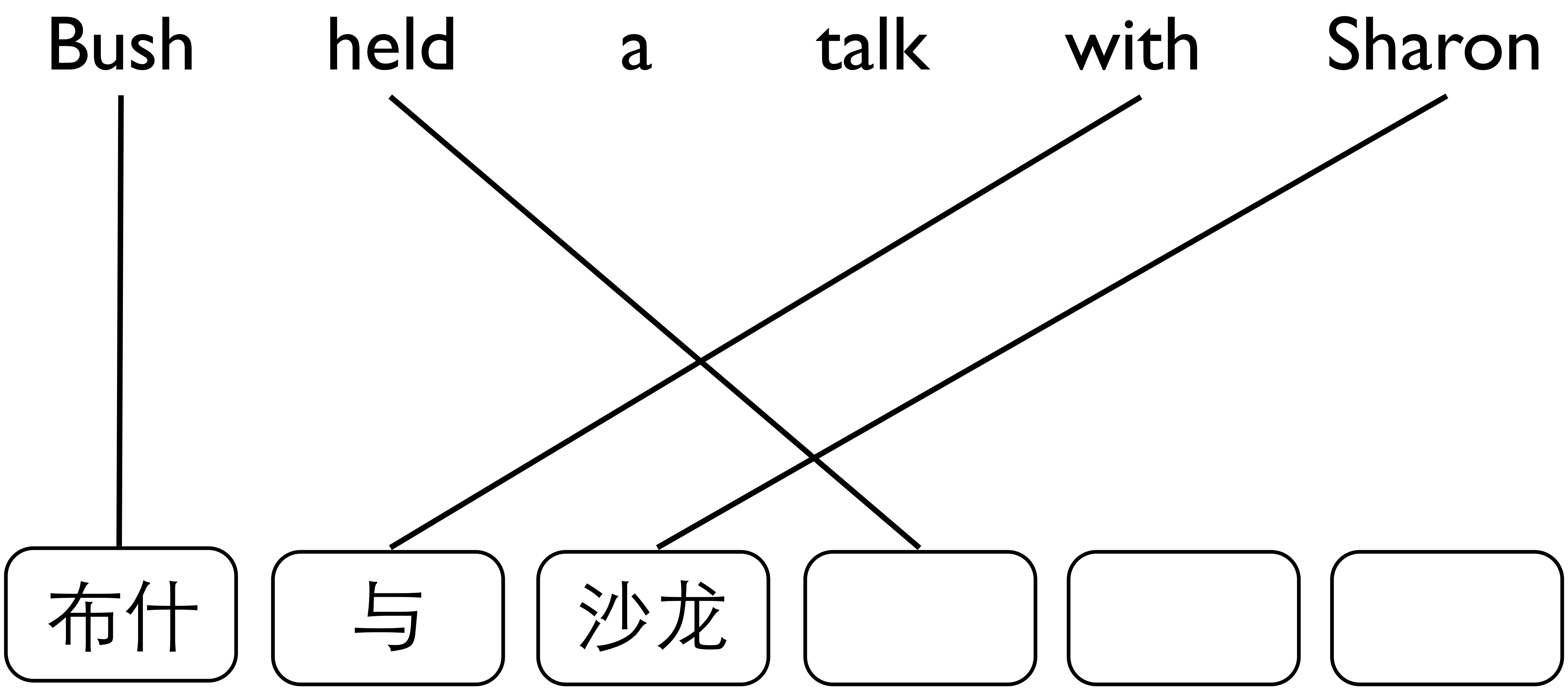
(Brown et al., 1993)

Translation Model



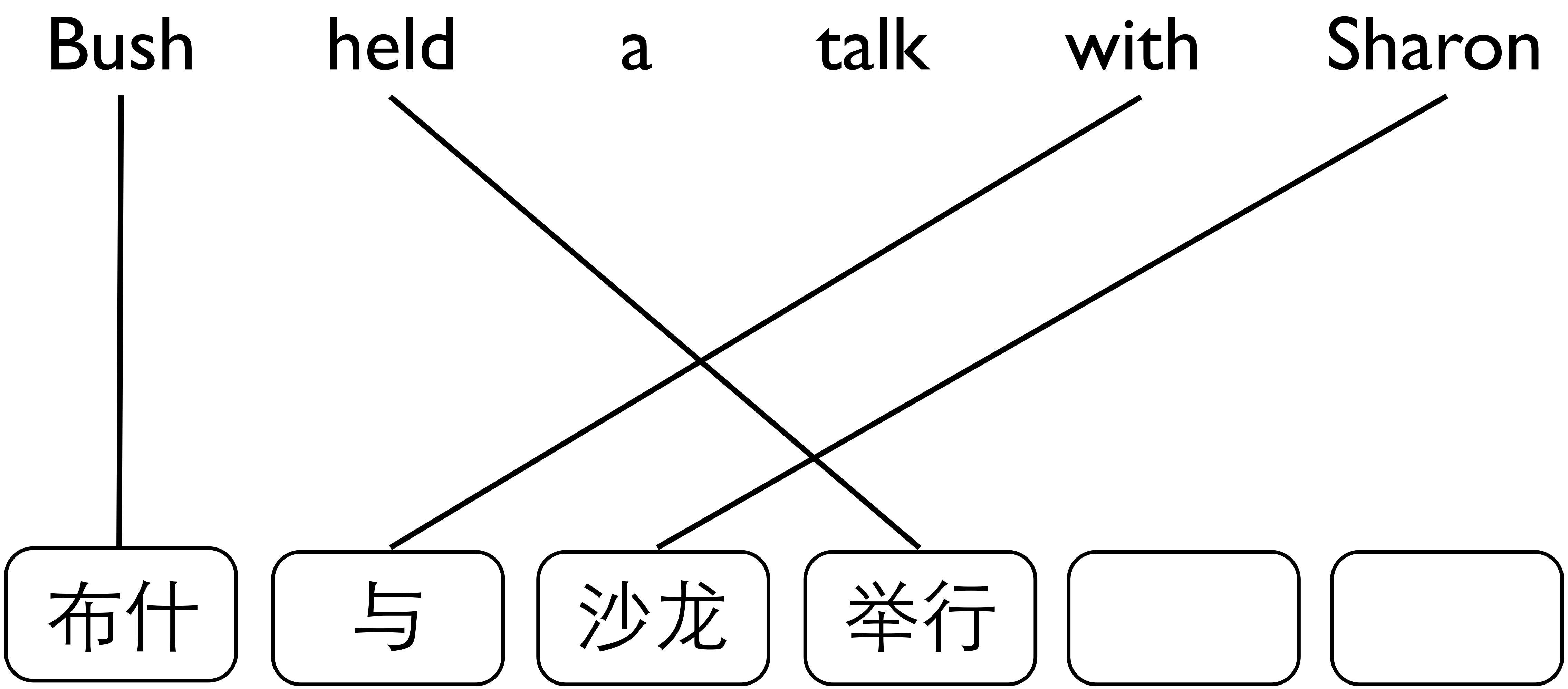
(Brown et al., 1993)

Translation Model



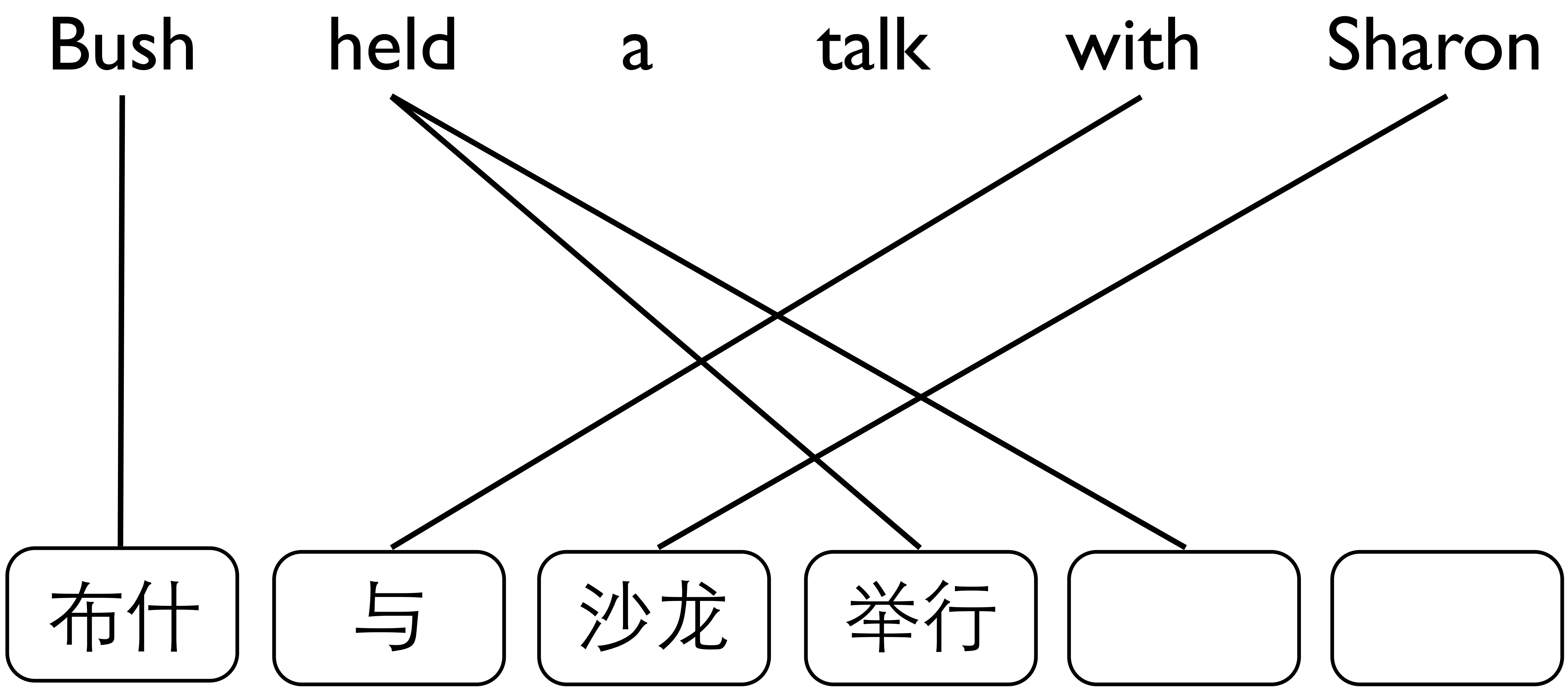
(Brown et al., 1993)

Translation Model



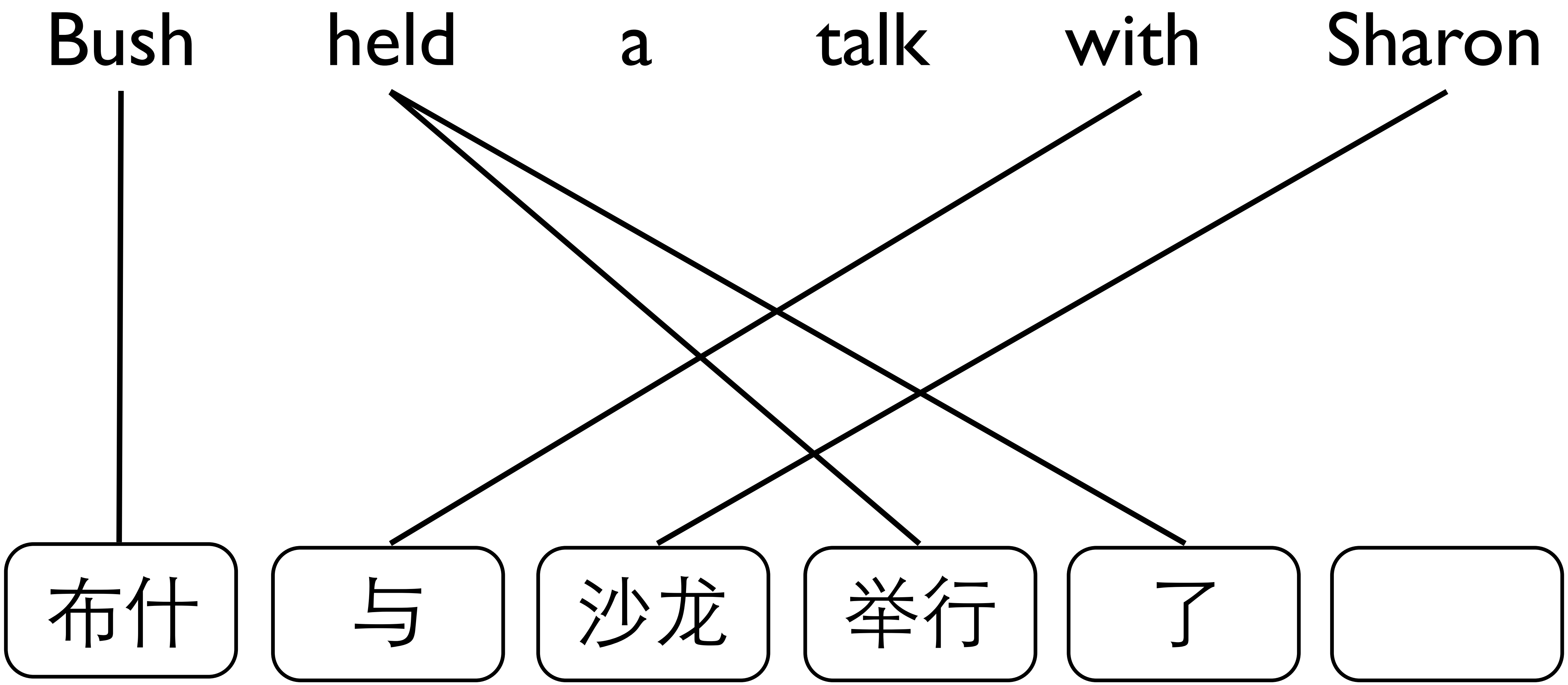
(Brown et al., 1993)

Translation Model



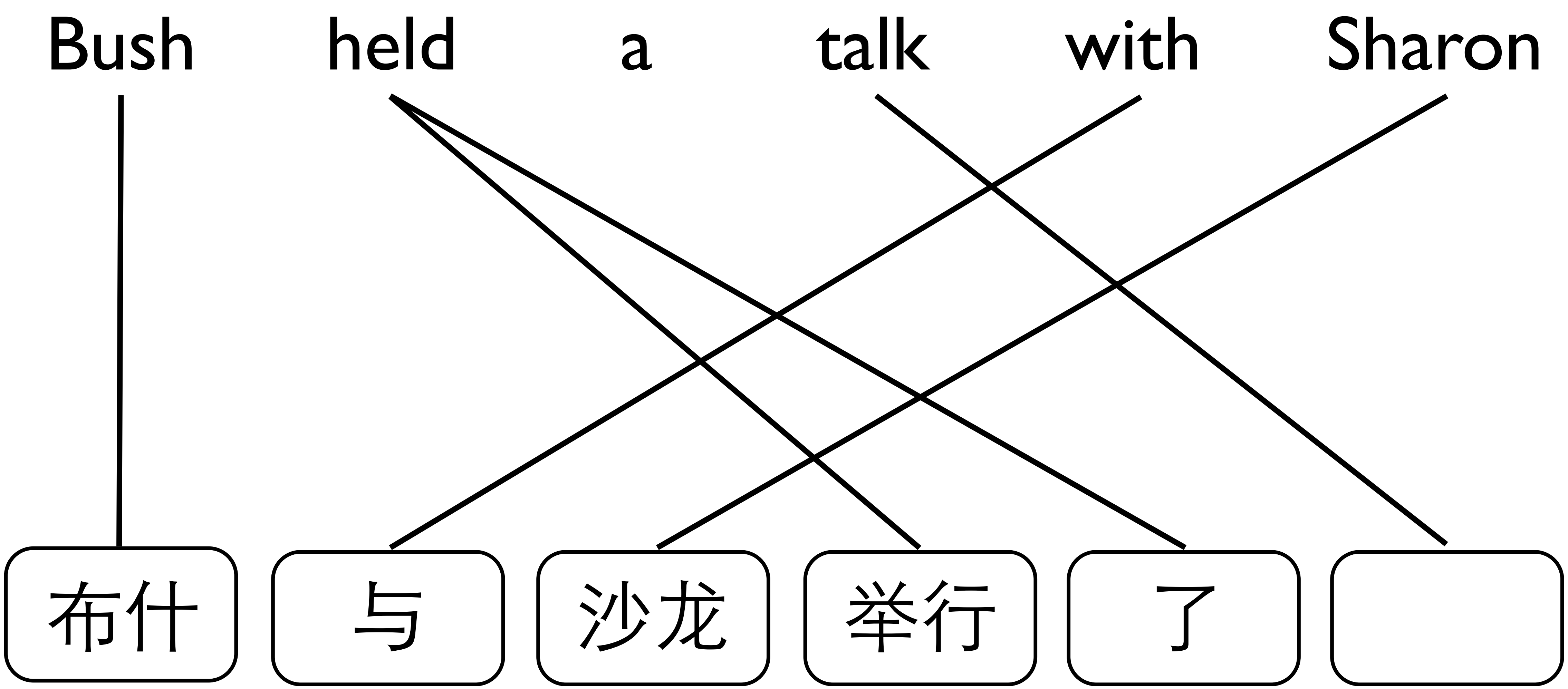
(Brown et al., 1993)

Translation Model



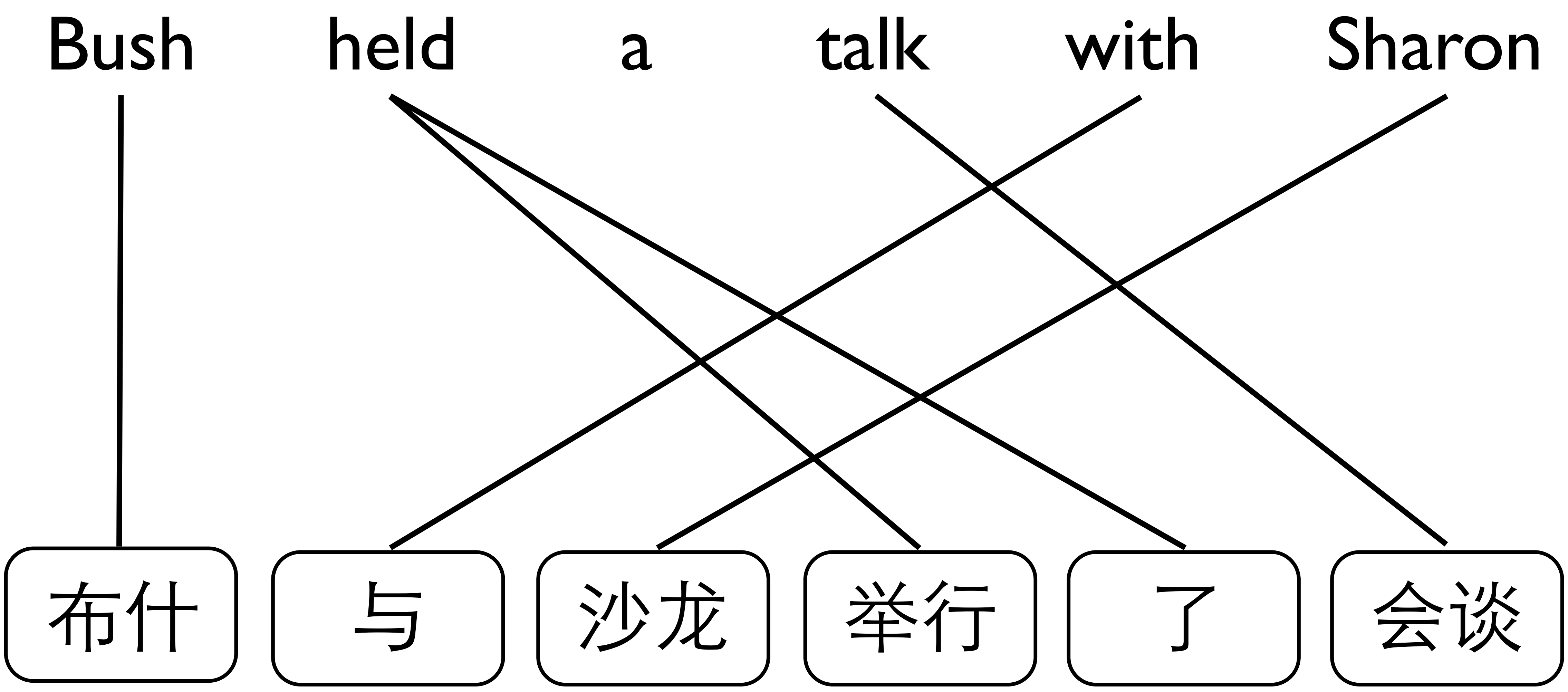
(Brown et al., 1993)

Translation Model



(Brown et al., 1993)

Translation Model



(Brown et al., 1993)

IBM Models 1 & 2

(Brown et al., 1993)

IBM Models 1 & 2

$$\Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}).$$

(Brown et al., 1993)

IBM Models 1 & 2

$$\Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}).$$

$$\Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) \quad = \quad \Pr(m|\mathbf{e}) \prod_{j=1}^{m} \Pr(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) \Pr(f_j|a_1^j, f_1^{j-1}, m, \mathbf{e}).$$

(Brown et al., 1993)

IBM Models 1 & 2

$$\Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}).$$

$$\Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) \quad = \quad \Pr(m|\mathbf{e}) \prod_{j=1}^{m} \Pr(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) \Pr(f_j|a_1^{j}, f_1^{j-1}, m, \mathbf{e}).$$

(Brown et al., 1993)

IBM Models 1 & 2

$$\Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}).$$

$$\Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) \quad = \quad \boxed{\Pr(m|\mathbf{e})} \prod_{j=1}^{m} \Pr(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) \Pr(f_j|a_1^j, f_1^{j-1}, m, \mathbf{e}).$$

length
model

(Brown et al., 1993)

IBM Models 1 & 2

$$\Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}).$$

$$\Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) \quad = \quad \Pr(m|\mathbf{e}) \prod_{j=1}^{m} \Pr(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) \Pr(f_j|a_1^j, f_1^{j-1}, m, \mathbf{e}).$$

length
model

(Brown et al., 1993)

IBM Models 1 & 2

$$\Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}).$$

$$\Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) \quad = \quad \Pr(m|\mathbf{e}) \prod_{j=1}^{m} \Pr(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) \Pr(f_j|a_1^j, f_1^{j-1}, m, \mathbf{e}).$$

length model

alignment model

(Brown et al., 1993)

IBM Models 1 & 2

$$\Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}).$$

$$\Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \Pr(m|\mathbf{e}) \prod_{j=1}^{m} \Pr(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) \Pr(f_j|a_1^j, f_1^{j-1}, m, \mathbf{e}).$$

length model

alignment model

(Brown et al., 1993)

IBM Models 1 & 2

$$\Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} \Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}).$$

$$\Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \Pr(m|\mathbf{e}) \prod_{j=1}^{m} \Pr(a_j|a_1^{j-1}, f_1^{j-1}, m, \mathbf{e}) \Pr(f_j|a_1^j, f_1^{j-1}, m, \mathbf{e}).$$

length model

alignment model

translation model

(Brown et al., 1993)

Translation Model

Bush held a talk with Sharon

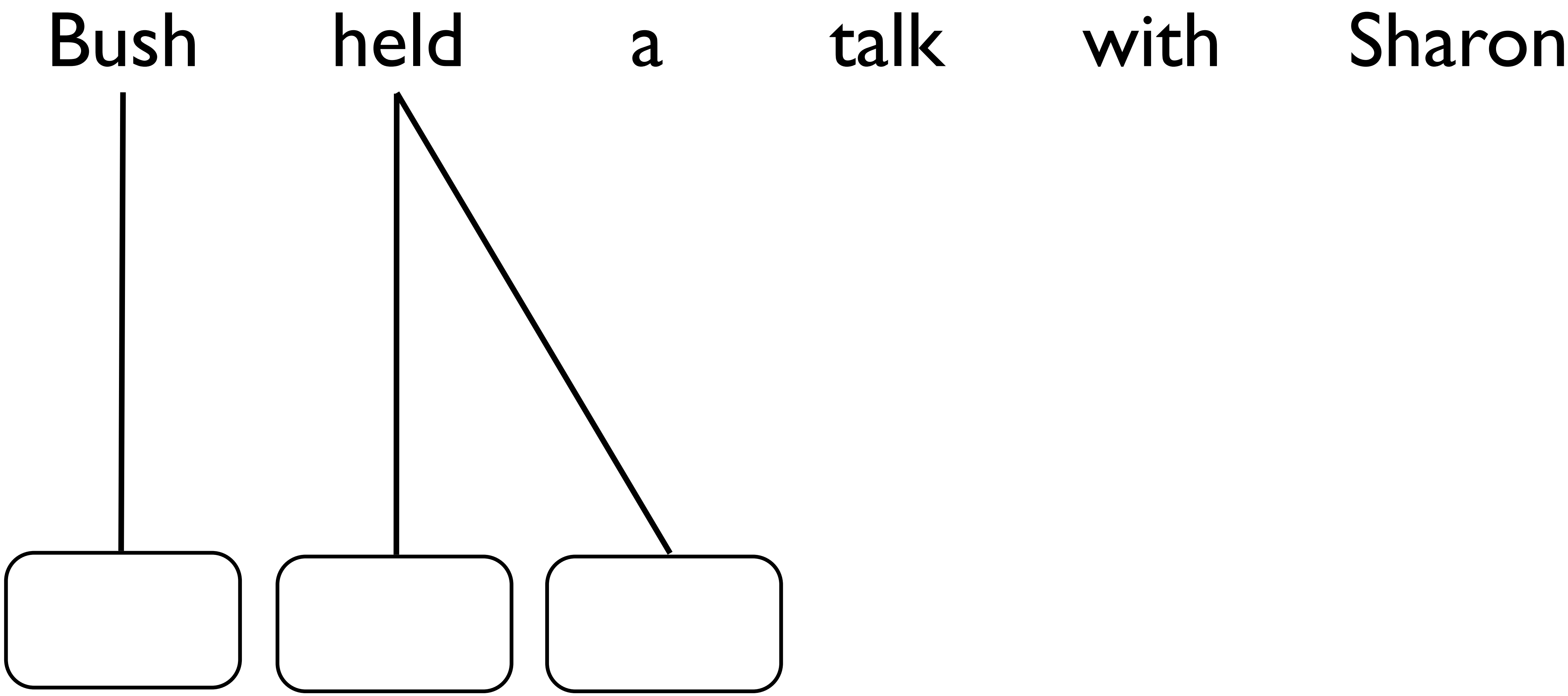
(Brown et al., 1993)

Translation Model

Bush held a talk with Sharon

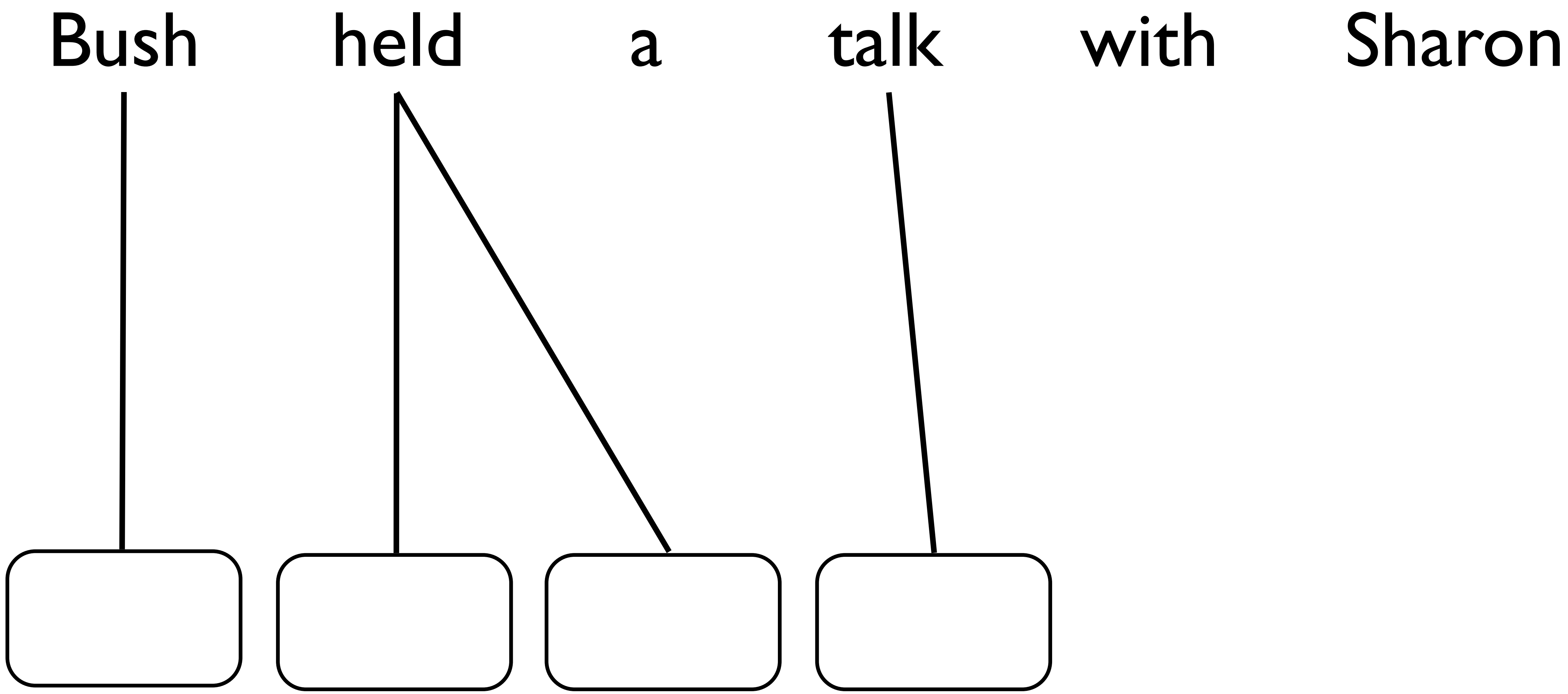
(Brown et al., 1993)

Translation Model



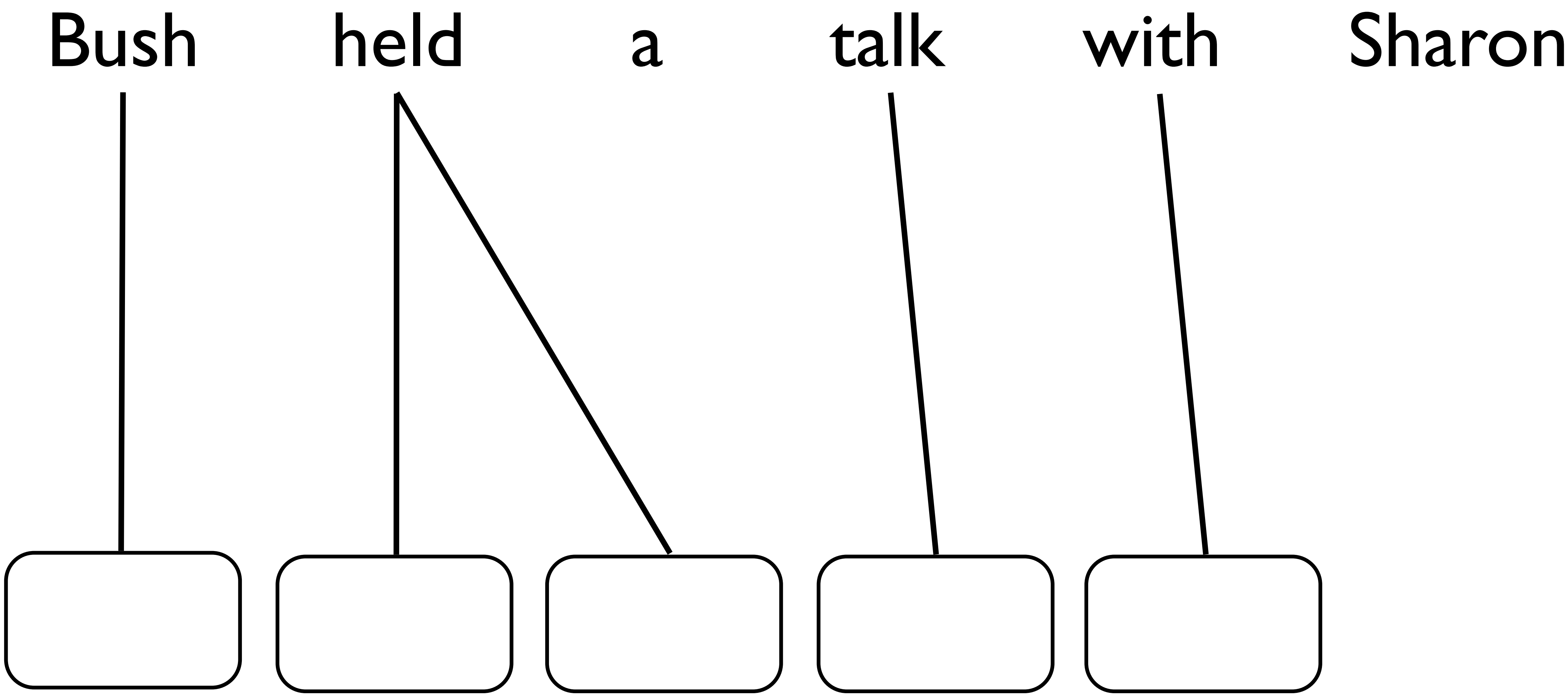
(Brown et al., 1993)

Translation Model



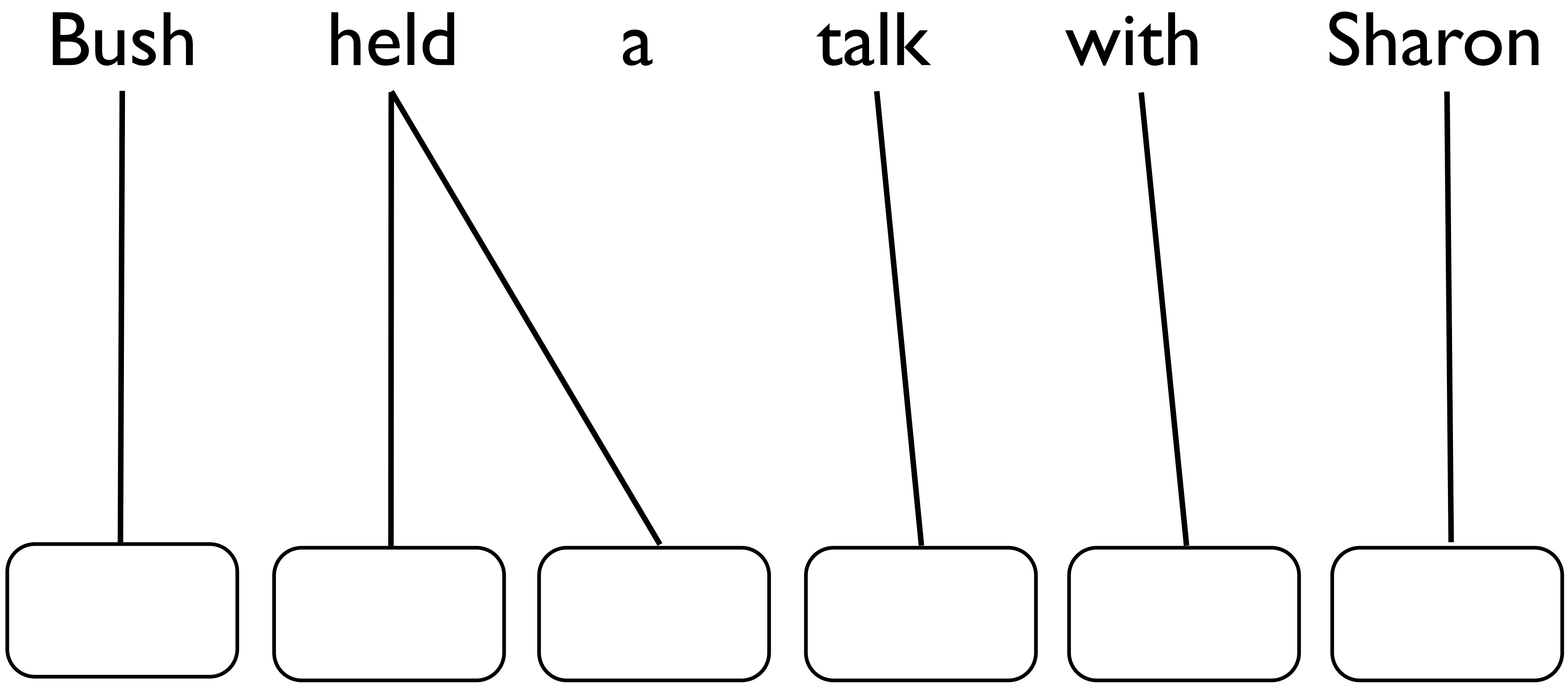
(Brown et al., 1993)

Translation Model



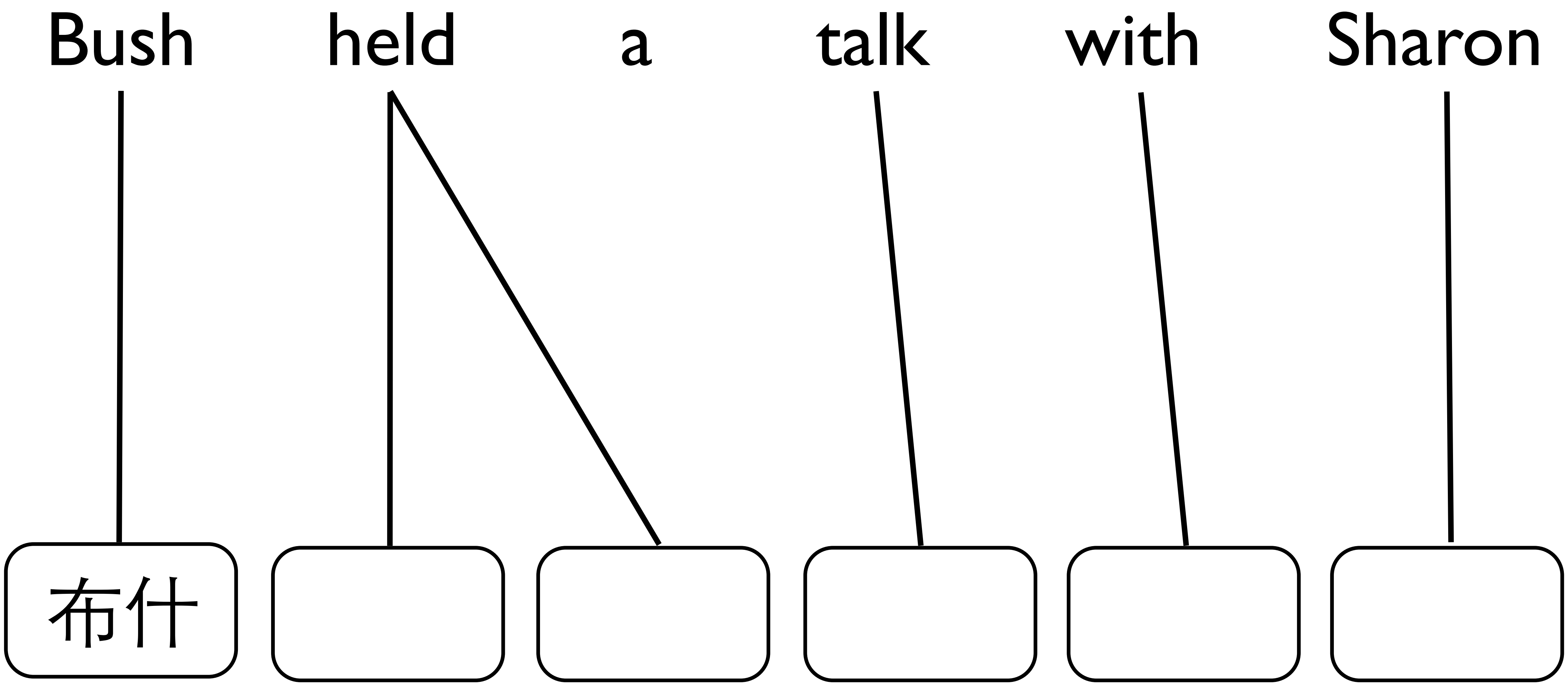
(Brown et al., 1993)

Translation Model



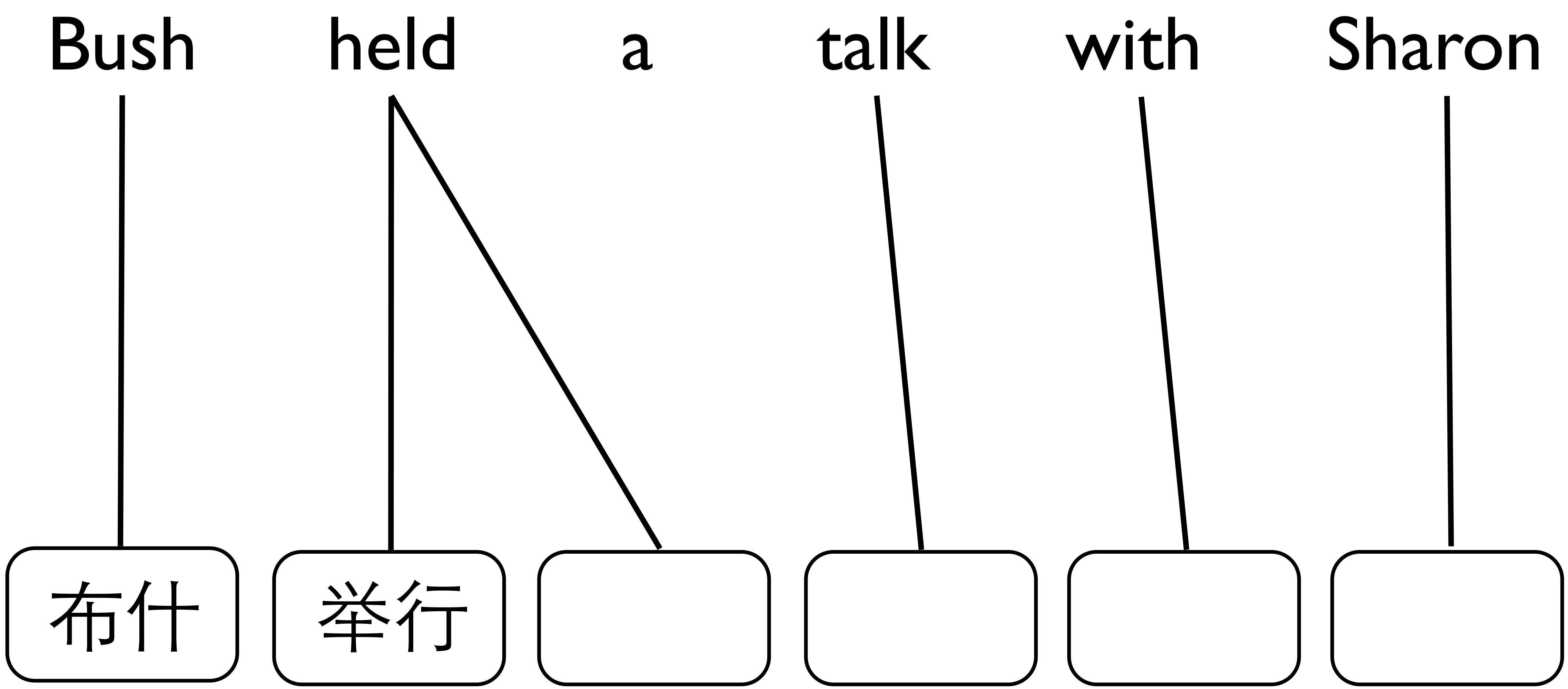
(Brown et al., 1993)

Translation Model



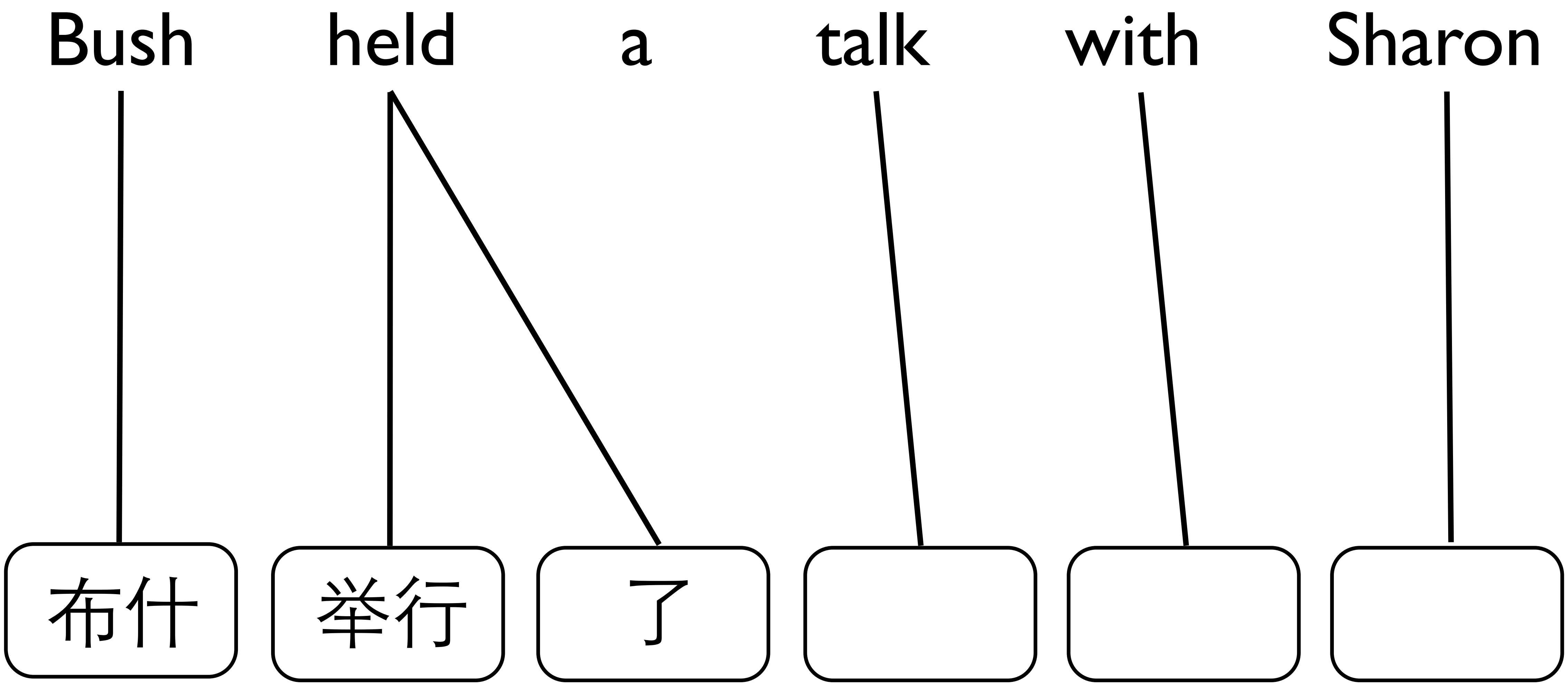
(Brown et al., 1993)

Translation Model



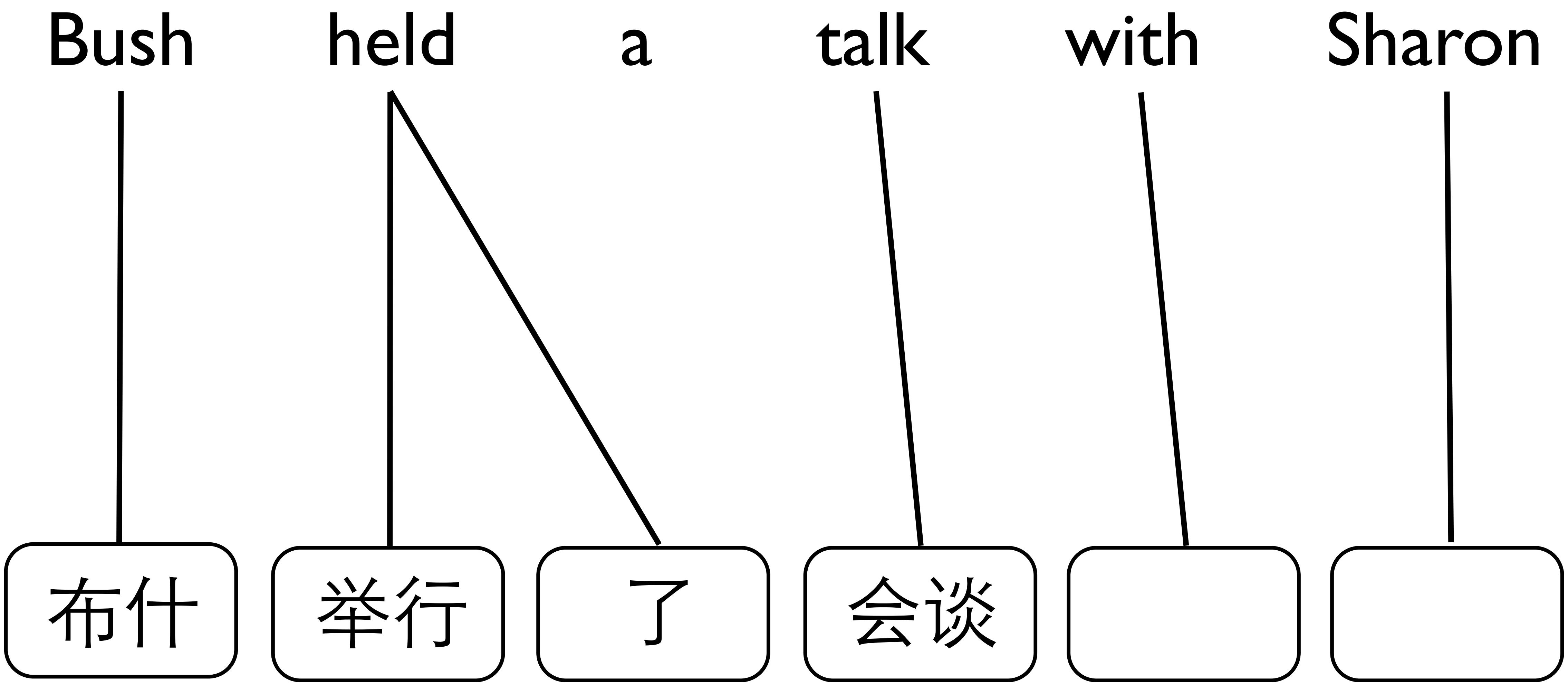
(Brown et al., 1993)

Translation Model



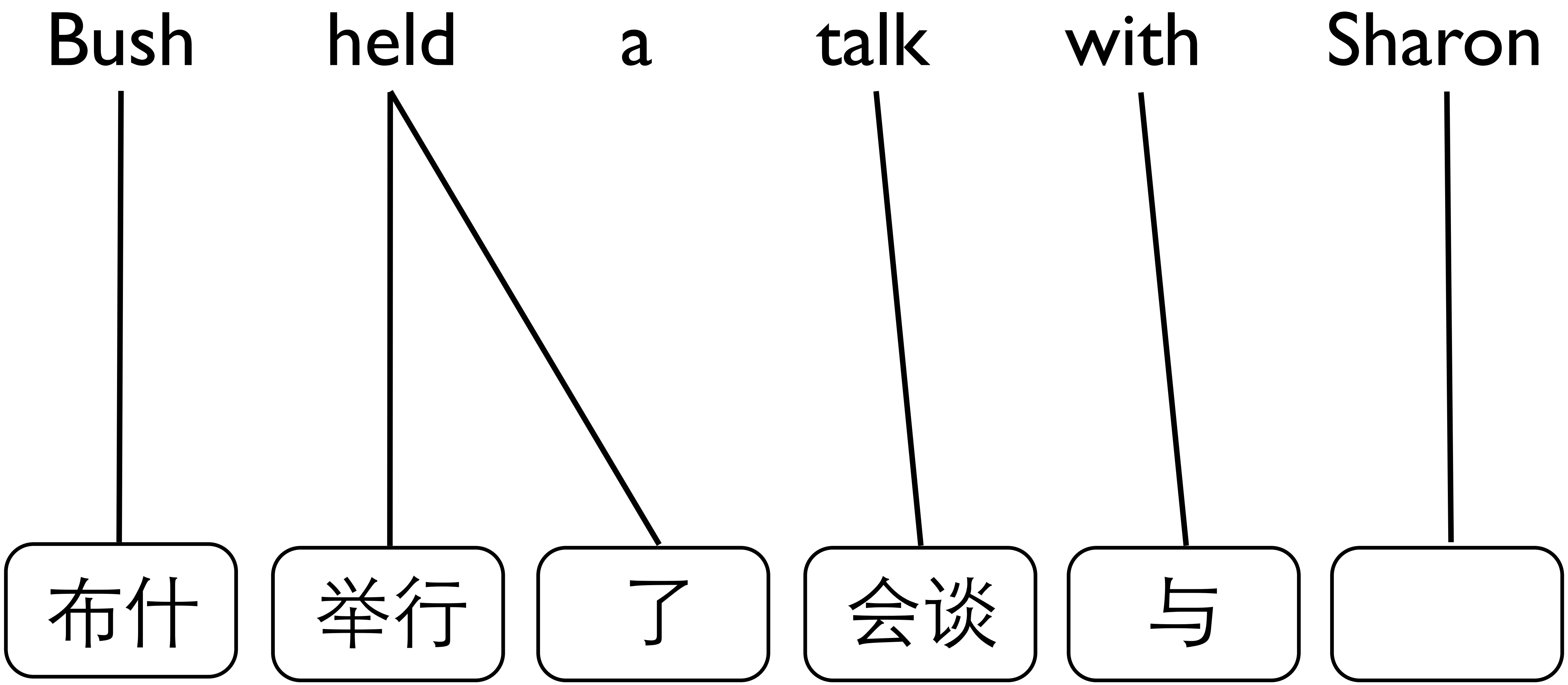
(Brown et al., 1993)

Translation Model



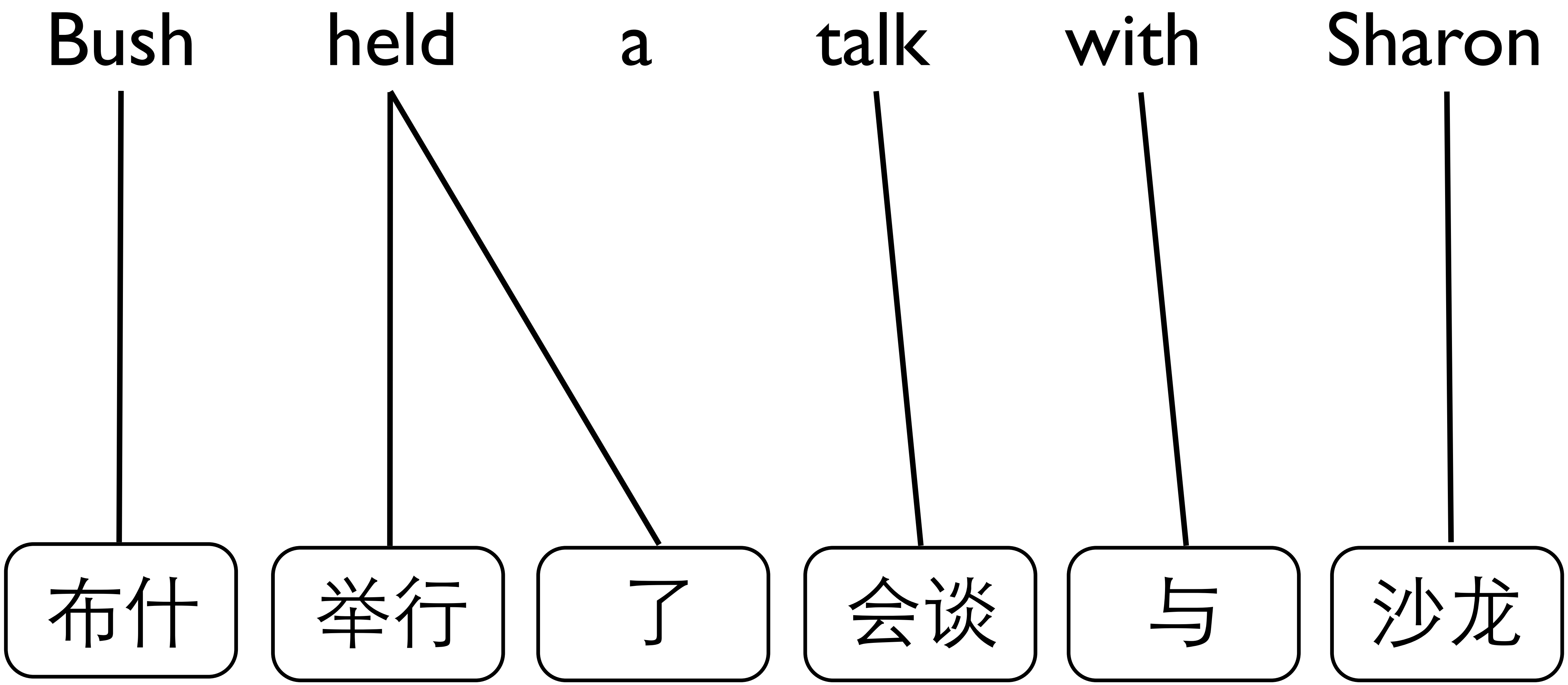
(Brown et al., 1993)

Translation Model



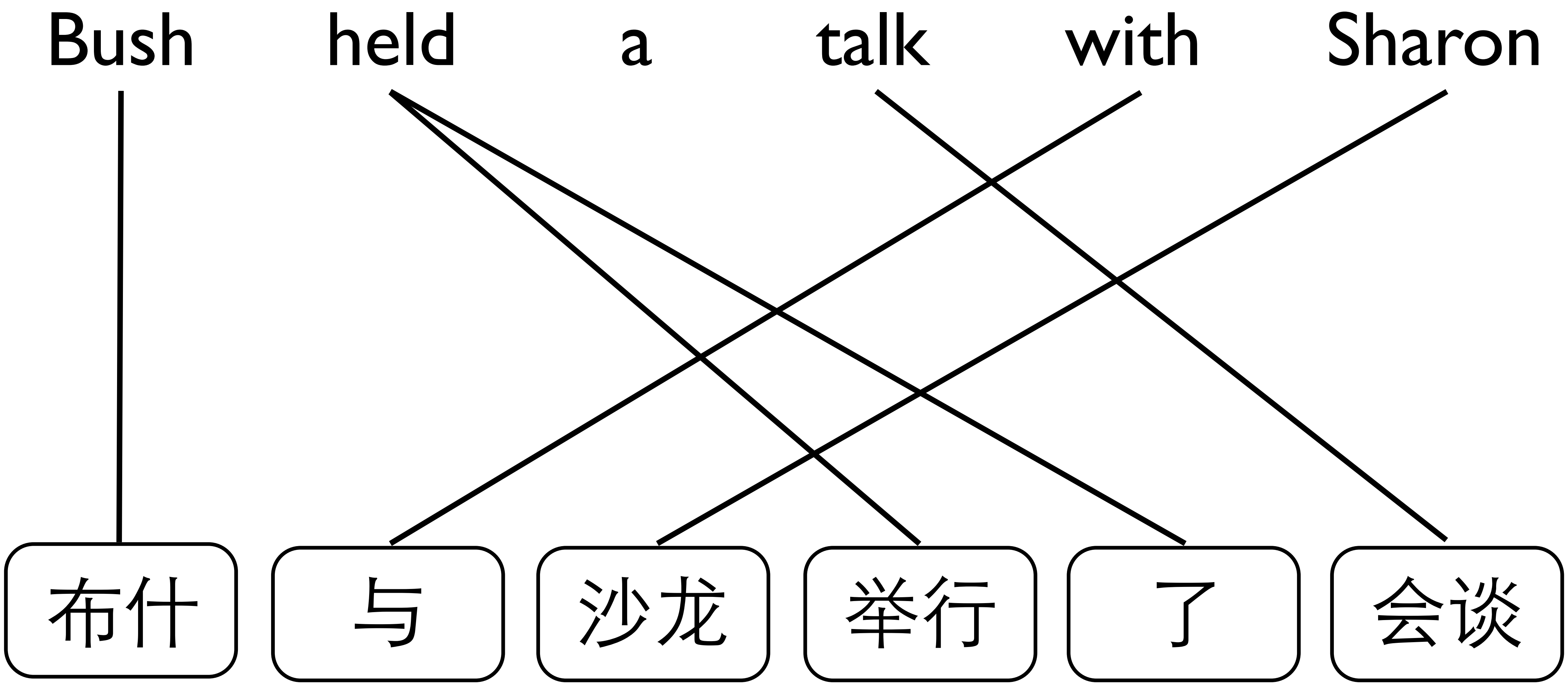
(Brown et al., 1993)

Translation Model



(Brown et al., 1993)

Translation Model



(Brown et al., 1993)

IBM Models 3-5

(Brown et al., 1993)

IBM Models 3-5

$$\Pr(\mathbf{f}, \mathbf{a} | \mathbf{e}) = \sum_{(\tau, \pi) \in \langle \mathbf{f}, \mathbf{a} \rangle} \Pr(\tau, \pi | \mathbf{e}).$$

$$\begin{aligned}
\Pr(\tau, \pi | \mathbf{e}) \quad = \quad & \prod_{i=1}^{l} \Pr(\phi_i | \phi_1^{i-1}, \mathbf{e}) \Pr(\phi_0 | \phi_1^l, \mathbf{e}) \times \\
& \prod_{i=0}^{l} \prod_{k=1}^{\phi_i} \Pr(\tau_{ik} | \tau_{i1}^{k-1}, \tau_0^{i-1}, \phi_0^l, \mathbf{e}) \times \\
& \prod_{i=1}^{l} \prod_{k=1}^{\phi_i} \Pr(\pi_{ik} | \pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^l, \phi_0^l, \mathbf{e}) \times \\
& \prod_{k=1}^{\phi_0} \Pr(\pi_{0k} | \pi_{01}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, \mathbf{e}).
\end{aligned}$$

(Brown et al., 1993)

IBM Models 3-5

$$\Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \sum_{(\tau,\pi)\in\langle\mathbf{f},\mathbf{a}\rangle} \Pr(\tau, \pi|\mathbf{e}).$$

$$\begin{aligned}\Pr(\tau, \pi|\mathbf{e}) = &\prod_{i=1}^{l} \Pr(\phi_i|\phi_1^{i-1}, \mathbf{e}) \Pr(\phi_0|\phi_1^l, \mathbf{e}) \times \\ &\prod_{i=0}^{l}\prod_{k=1}^{\phi_i} \Pr(\tau_{ik}|\tau_{i1}^{k-1}, \tau_0^{i-1}, \phi_0^l, \mathbf{e}) \times \\ &\prod_{i=1}^{l}\prod_{k=1}^{\phi_i} \Pr(\pi_{ik}|\pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^l, \phi_0^l, \mathbf{e}) \times \\ &\prod_{k=1}^{\phi_0} \Pr(\pi_{0k}|\pi_{01}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, \mathbf{e}).\end{aligned}$$

(Brown et al., 1993)

IBM Models 3-5

$$\Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \sum_{(\tau,\pi)\in\langle\mathbf{f},\mathbf{a}\rangle} \Pr(\tau, \pi|\mathbf{e}).$$

$$\begin{aligned}\Pr(\tau, \pi|\mathbf{e}) = & \prod_{i=1}^{l} \boxed{\Pr(\phi_i|\phi_1^{i-1}, \mathbf{e}) \Pr(\phi_0|\phi_1^l, \mathbf{e})} \times \\ & \prod_{i=0}^{l}\prod_{k=1}^{\phi_i} \Pr(\tau_{ik}|\tau_{i1}^{k-1}, \tau_0^{i-1}, \phi_0^l, \mathbf{e}) \times \\ & \prod_{i=1}^{l}\prod_{k=1}^{\phi_i} \Pr(\pi_{ik}|\pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^l, \phi_0^l, \mathbf{e}) \times \\ & \prod_{k=1}^{\phi_0} \Pr(\pi_{0k}|\pi_{01}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, \mathbf{e}).\end{aligned}$$

fertility model

(Brown et al., 1993)

IBM Models 3-5

$$\Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \sum_{(\tau,\pi)\in\langle\mathbf{f},\mathbf{a}\rangle} \Pr(\tau, \pi|\mathbf{e}).$$

$$\begin{aligned}\Pr(\tau, \pi|\mathbf{e}) \quad = \quad & \prod_{i=1}^{l} \Pr(\phi_i|\phi_1^{i-1}, \mathbf{e}) \Pr(\phi_0|\phi_1^{l}, \mathbf{e}) \times \\ & \prod_{i=0}^{l}\prod_{k=1}^{\phi_i} \Pr(\tau_{ik}|\tau_{i1}^{k-1}, \tau_0^{i-1}, \phi_0^l, \mathbf{e}) \times \\ & \prod_{i=1}^{l}\prod_{k=1}^{\phi_i} \Pr(\pi_{ik}|\pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^l, \phi_0^l, \mathbf{e}) \times \\ & \prod_{k=1}^{\phi_0} \Pr(\pi_{0k}|\pi_{01}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, \mathbf{e}).\end{aligned}$$

fertility model

(Brown et al., 1993)

IBM Models 3-5

$$\Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \sum_{(\tau,\pi)\in\langle\mathbf{f},\mathbf{a}\rangle} \Pr(\tau, \pi|\mathbf{e}).$$

$$\begin{aligned}\Pr(\tau, \pi|\mathbf{e}) \quad = \quad & \prod_{i=1}^{l} \Pr(\phi_i|\phi_1^{i-1}, \mathbf{e}) \Pr(\phi_0|\phi_1^l, \mathbf{e}) \times \\ & \prod_{i=0}^{l}\prod_{k=1}^{\phi_i} \Pr(\tau_{ik}|\tau_{i1}^{k-1}, \tau_0^{i-1}, \phi_0^l, \mathbf{e}) \times \\ & \prod_{i=1}^{l}\prod_{k=1}^{\phi_i} \Pr(\pi_{ik}|\pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^l, \phi_0^l, \mathbf{e}) \times \\ & \prod_{k=1}^{\phi_0} \Pr(\pi_{0k}|\pi_{01}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, \mathbf{e}).\end{aligned}$$

fertility model

translation model

(Brown et al., 1993)

IBM Models 3-5

$$\Pr(\mathbf{f}, \mathbf{a} | \mathbf{e}) = \sum_{(\tau,\pi) \in \langle \mathbf{f}, \mathbf{a} \rangle} \Pr(\tau, \pi | \mathbf{e}).$$

$$\begin{aligned}
\Pr(\tau, \pi | \mathbf{e}) \quad = \quad & \prod_{i=1}^{l} \Pr(\phi_i | \phi_1^{i-1}, \mathbf{e}) \Pr(\phi_0 | \phi_1^l, \mathbf{e}) \times \\
& \prod_{i=0}^{l} \prod_{k=1}^{\phi_i} \Pr(\tau_{ik} | \tau_{i1}^{k-1}, \tau_0^{i-1}, \phi_0^l, \mathbf{e}) \times \\
& \prod_{i=1}^{l} \prod_{k=1}^{\phi_i} \Pr(\pi_{ik} | \pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^l, \phi_0^l, \mathbf{e}) \times \\
& \prod_{k=1}^{\phi_0} \Pr(\pi_{0k} | \pi_{01}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, \mathbf{e}).
\end{aligned}$$

fertility model

translation model

(Brown et al., 1993)

IBM Models 3-5

$$\Pr(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \sum_{(\tau,\pi)\in\langle\mathbf{f},\mathbf{a}\rangle} \Pr(\tau, \pi|\mathbf{e}).$$

$$\begin{aligned}\Pr(\tau, \pi|\mathbf{e}) \quad = \quad & \prod_{i=1}^{l} \Pr(\phi_i|\phi_1^{i-1}, \mathbf{e}) \Pr(\phi_0|\phi_1^{l}, \mathbf{e}) \times \\ & \prod_{i=0}^{l}\prod_{k=1}^{\phi_i} \Pr(\tau_{ik}|\tau_{i1}^{k-1}, \tau_0^{i-1}, \phi_0^l, \mathbf{e}) \times \\ & \prod_{i=1}^{l}\prod_{k=1}^{\phi_i} \Pr(\pi_{ik}|\pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^l, \phi_0^l, \mathbf{e}) \times \\ & \prod_{k=1}^{\phi_0} \Pr(\pi_{0k}|\pi_{01}^{k-1}, \pi_1^l, \tau_0^l, \phi_0^l, \mathbf{e}).\end{aligned}$$

fertility model

translation model

distortion model

(Brown et al., 1993)

Learning from Data

Q: how to learn model parameters from data?

Garcia y asociados .

Garcia and associates .

los clients y los asociados son enemigos .

the clients and the associates are enemies .

sus asociados no son fuertes .

his associates are not strong .

Spanish	English
Garcia	Garcia
y	and
asociados	associates
.	.
los	the
clients	clients
son	are
enemigos	enemies
sus	his
no	not
fuertes	strong

Maximum Likelihood Estimation

(Brown et al., 1993)

Maximum Likelihood Estimation

input: $(\mathbf{f}^{(1)}, \mathbf{e}^{(1)}) \ldots (\mathbf{f}^{(S)}, \mathbf{e}^{(S)})$

(Brown et al., 1993)

Maximum Likelihood Estimation

alignment is unobserved

input: $(\mathbf{f}^{(1)}, \mathbf{e}^{(1)}) \ldots (\mathbf{f}^{(S)}, \mathbf{e}^{(S)})$

(Brown et al., 1993)

Maximum Likelihood Estimation

alignment is unobserved

input: $(\mathbf{f}^{(1)}, \mathbf{e}^{(1)}) \ldots (\mathbf{f}^{(S)}, \mathbf{e}^{(S)})$

output: θ

(Brown et al., 1993)

Maximum Likelihood Estimation

alignment is unobserved

input: $(\mathbf{f}^{(1)}, \mathbf{e}^{(1)}) \ldots (\mathbf{f}^{(S)}, \mathbf{e}^{(S)})$

output: θ

$$\hat{\theta} = \mathrm{argmax}_{\theta} \left\{ \prod_{s=1}^{S} P_{\theta}(\mathbf{f}^{(s)} | \mathbf{e}^{(s)}) \right\}$$

(Brown et al., 1993)

Maximum Likelihood Estimation

alignment is unobserved

input: $(\mathbf{f}^{(1)}, \mathbf{e}^{(1)}) \dots (\mathbf{f}^{(S)}, \mathbf{e}^{(S)})$

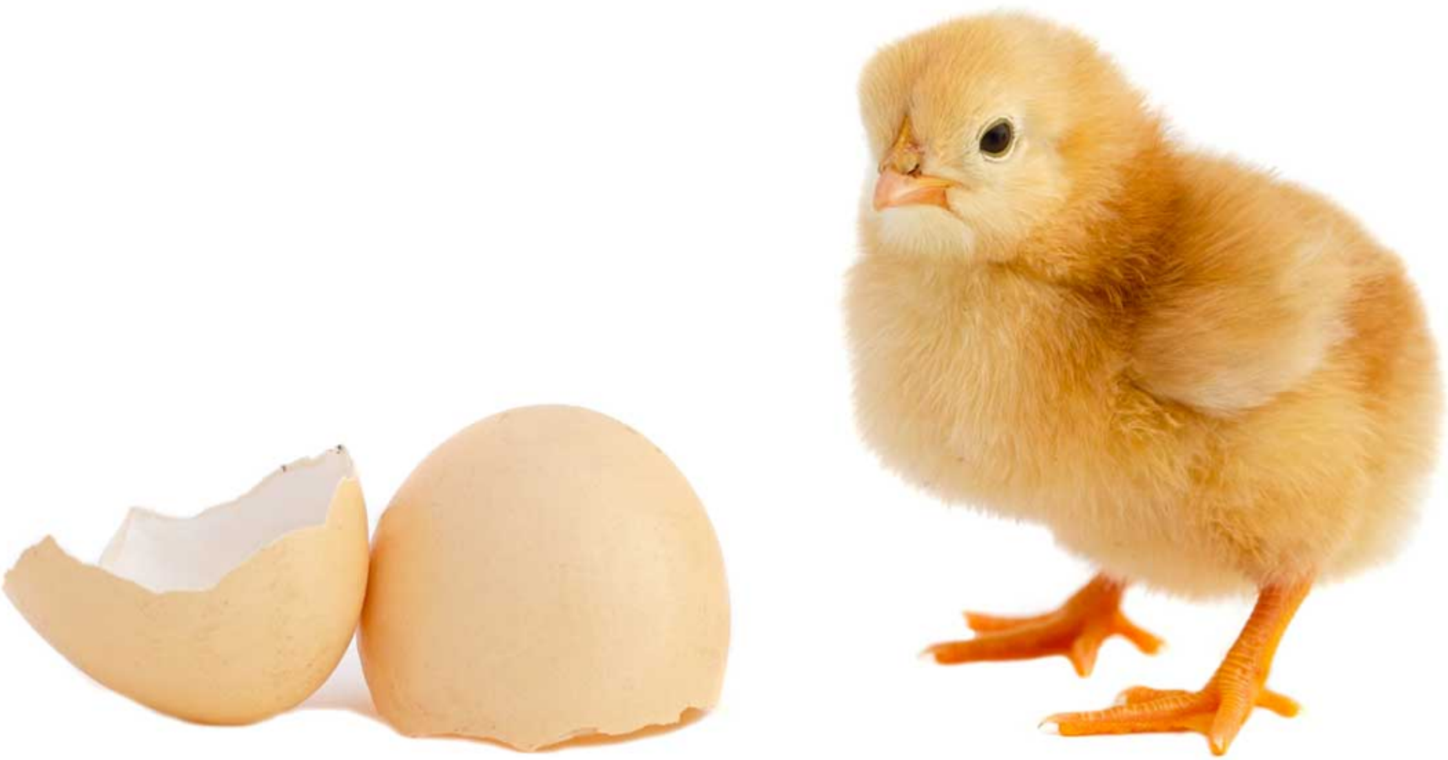
output: θ

$$\hat{\theta} = \mathrm{argmax}_{\theta} \left\{ \prod_{s=1}^{S} P_{\theta}(\mathbf{f}^{(s)} | \mathbf{e}^{(s)}) \right\}$$

The EM algorithm is often used for estimating parameters from unlabeled data

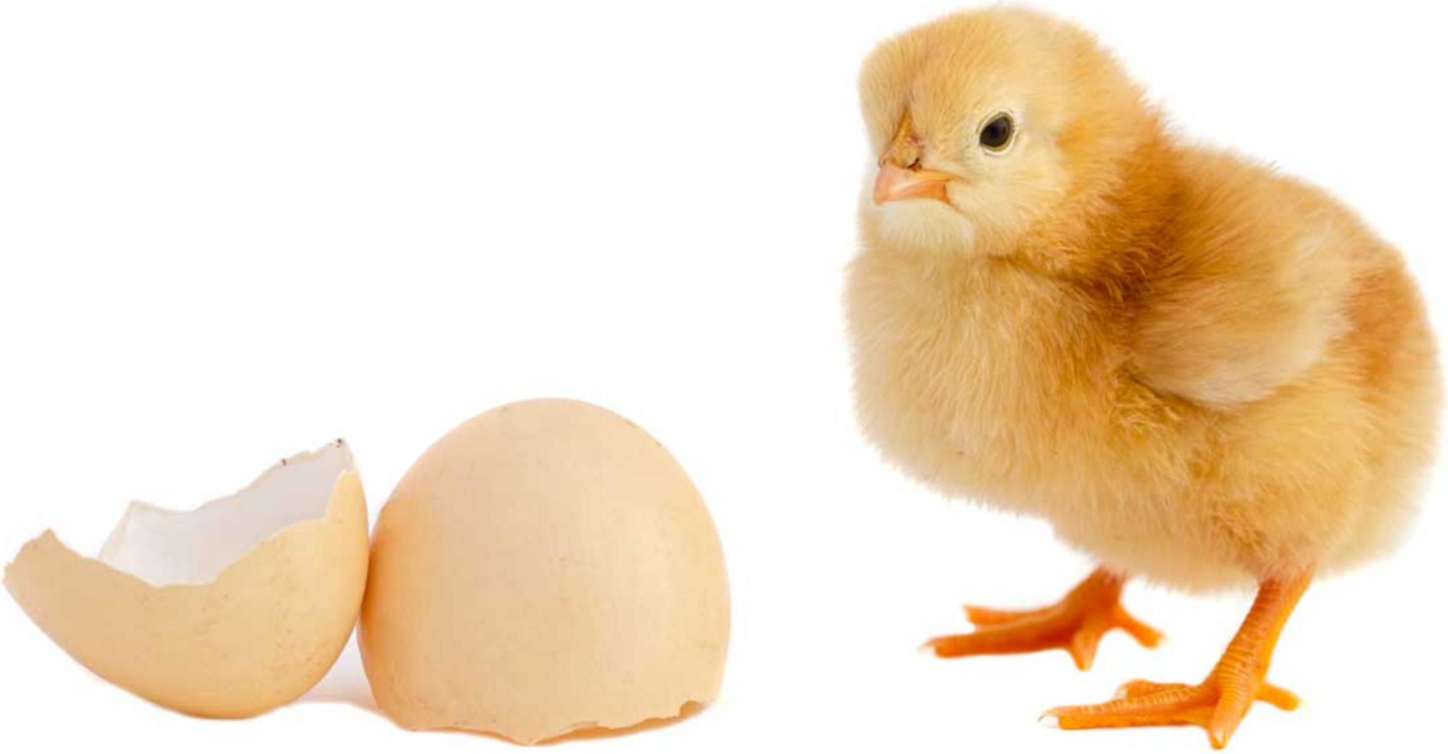
(Brown et al., 1993)

Learning from Unlabeled Data



Learning from Unlabeled Data

labels



Learning from Unlabeled Data

labels

parameters

Example

与 沙龙

with Sharon

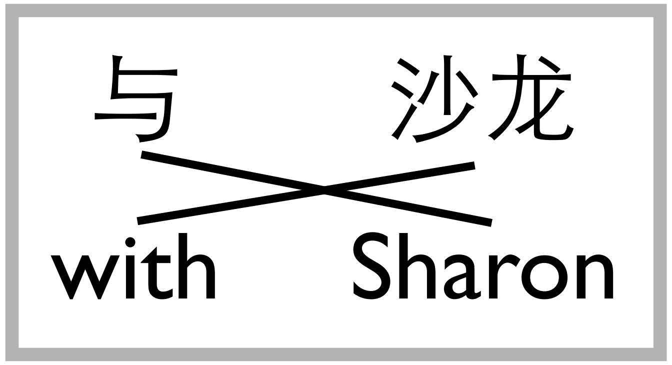
与

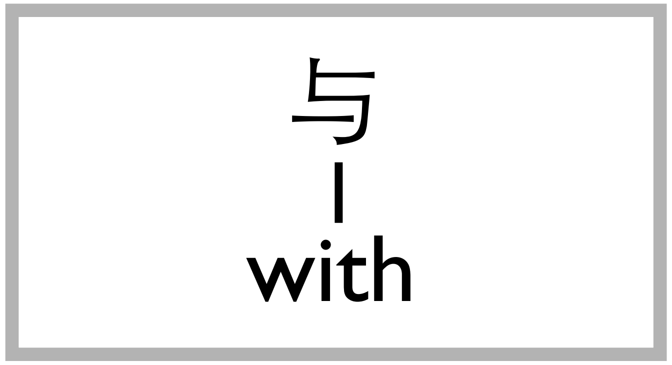
with

f	e	tc	t
与	with	N/A	0.5
	Sharon	N/A	0.5
沙龙	with	N/A	0.5
	Sharon	N/A	0.5

(Brown et al., 1993)

Example



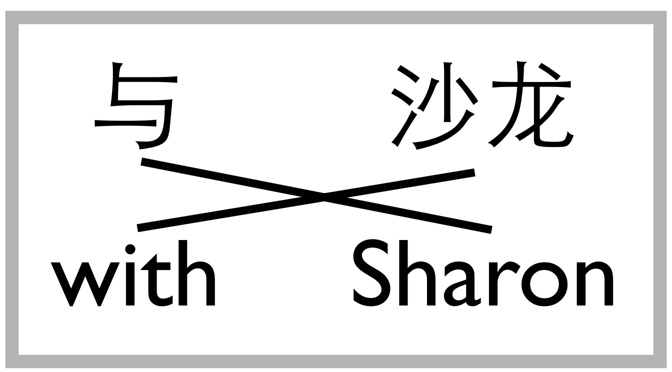


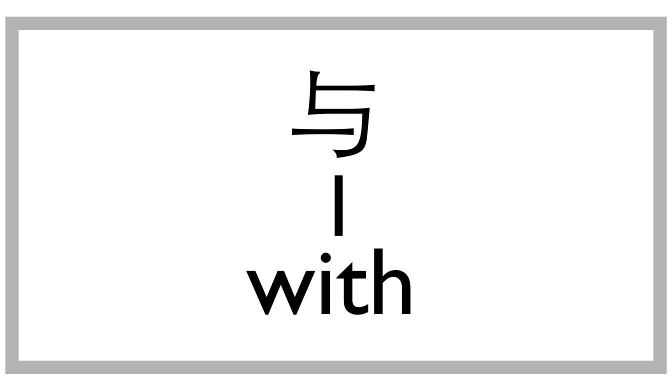
f	e	tc	t
与	with	N/A	0.5
	Sharon	N/A	0.5
沙龙	with	N/A	0.5
	Sharon	N/A	0.5

(Brown et al., 1993)

Example

$$P(\mathbf{f}, \mathbf{a} | \mathbf{e})$$



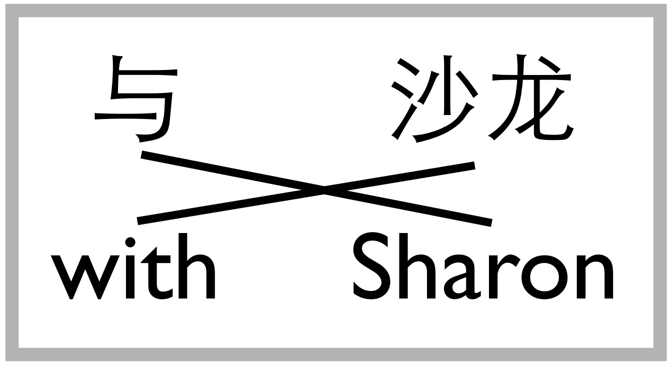


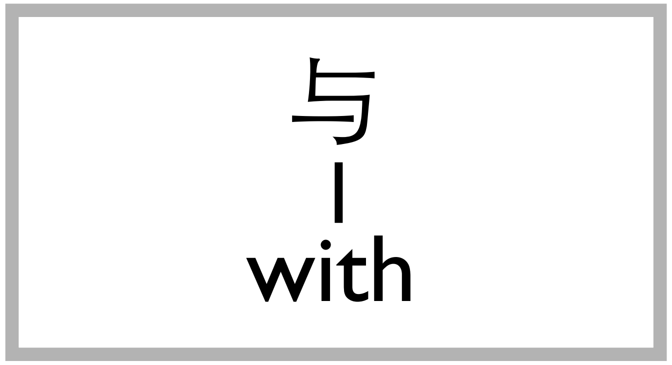
f	e	tc	t
与	with	N/A	0.5
	Sharon	N/A	0.5
沙龙	with	N/A	0.5
	Sharon	N/A	0.5

(Brown et al., 1993)

Example

$P(\mathbf{f}, \mathbf{a}|\mathbf{e})$





$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \prod_{j=1}^{|\mathbf{f}|} t(\mathbf{f}_j|\mathbf{e}_{\mathbf{a}_j})$$

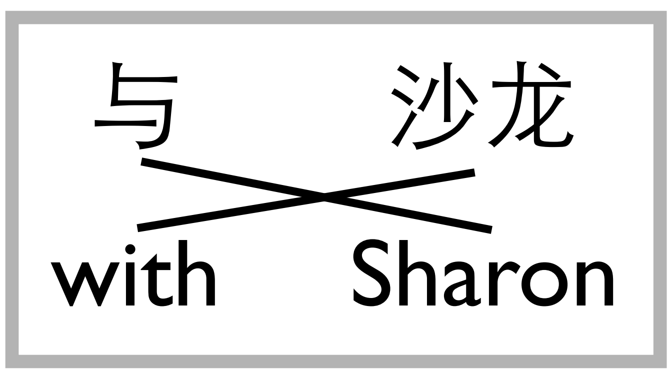
f	e	tc	t
与	with	N/A	0.5
	Sharon	N/A	0.5
沙龙	with	N/A	0.5
	Sharon	N/A	0.5

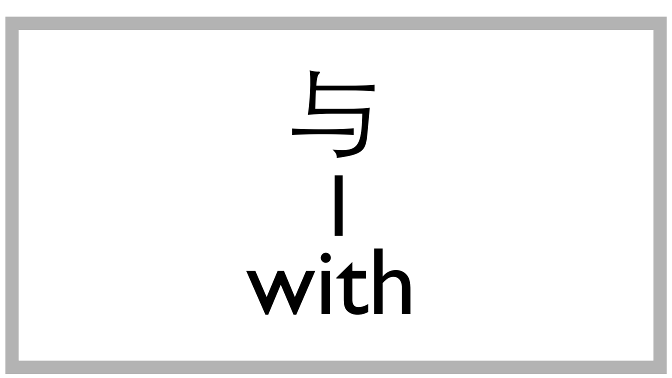
(Brown et al., 1993)

Example

$$P(\mathbf{f}, \mathbf{a}|\mathbf{e})$$

0.25





$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \prod_{j=1}^{|\mathbf{f}|} t(\mathbf{f}_j|\mathbf{e}_{\mathbf{a}_j})$$

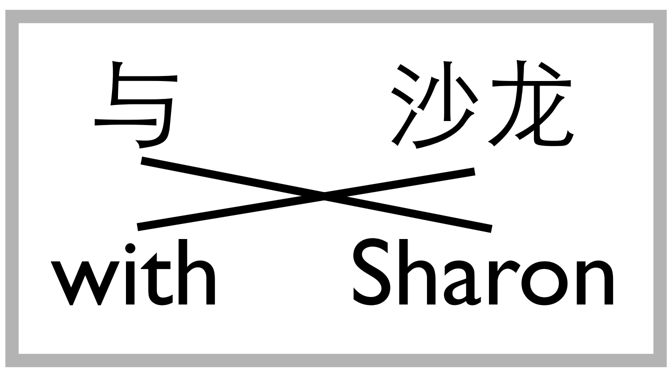
f	e	tc	t
与	with	N/A	0.5
	Sharon	N/A	0.5
沙龙	with	N/A	0.5
	Sharon	N/A	0.5

(Brown et al., 1993)

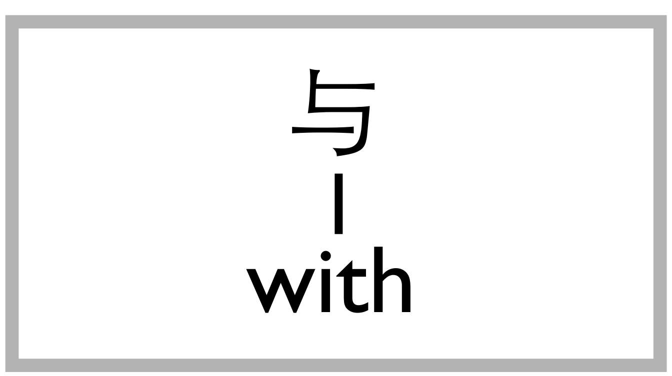
Example

$P(\mathbf{f}, \mathbf{a}|\mathbf{e})$

0.25



0.25



$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \prod_{j=1}^{|\mathbf{f}|} t(\mathbf{f}_j|\mathbf{e}_{\mathbf{a}_j})$$

f	e	tc	t
与	with	N/A	0.5
	Sharon	N/A	0.5
沙龙	with	N/A	0.5
	Sharon	N/A	0.5

(Brown et al., 1993)

Example

$P(\mathbf{f}, \mathbf{a}|\mathbf{e})$

与 — with, 沙龙 — Sharon

0.25

与 — Sharon, 沙龙 — with

0.25

与 — with

0.5

$$P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \prod_{j=1}^{|\mathbf{f}|} t(\mathbf{f}_j|\mathbf{e}_{\mathbf{a}_j})$$

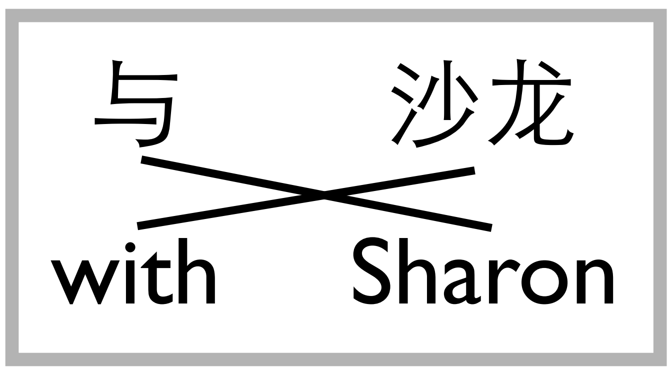
f	e	tc	t
与	with	N/A	0.5
	Sharon	N/A	0.5
沙龙	with	N/A	0.5
	Sharon	N/A	0.5

(Brown et al., 1993)

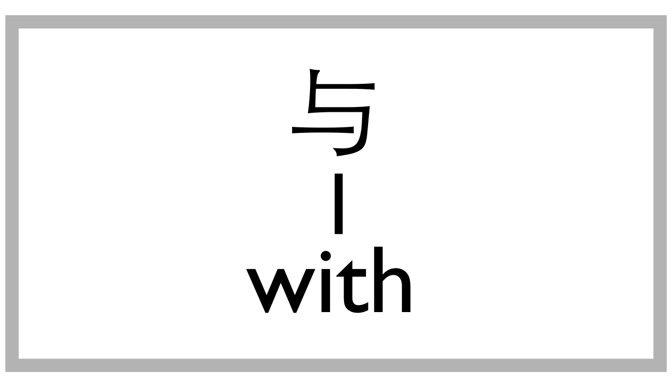
Example

$$P(\mathbf{f}, \mathbf{a} | \mathbf{e})$$

0.25



0.25



0.5

f	e	tc	t
与	with	N/A	0.5
	Sharon	N/A	0.5
沙龙	with	N/A	0.5
	Sharon	N/A	0.5

(Brown et al., 1993)

Example

$P(\mathbf{f}, \mathbf{a}|\mathbf{e})$ $\quad$ $P(\mathbf{a}|\mathbf{f}, \mathbf{e})$

Alignment	$P(\mathbf{f}, \mathbf{a}\vert\mathbf{e})$	$P(\mathbf{a}\vert\mathbf{f}, \mathbf{e})$
与 — with, 沙龙 — Sharon	0.25	
与 — Sharon, 沙龙 — with	0.25	
与 — with	0.5	

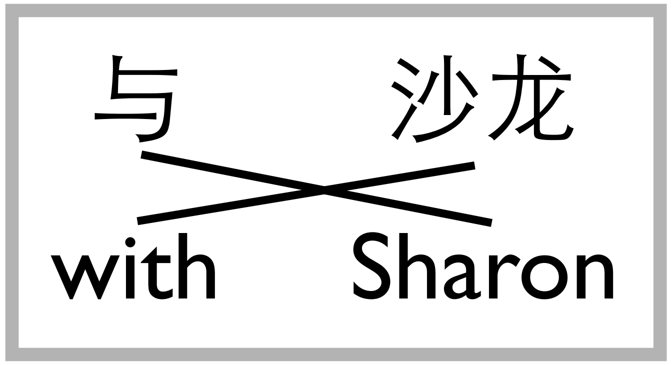
f	e	tc	t
与	with	N/A	0.5
	Sharon	N/A	0.5
沙龙	with	N/A	0.5
	Sharon	N/A	0.5

(Brown et al., 1993)

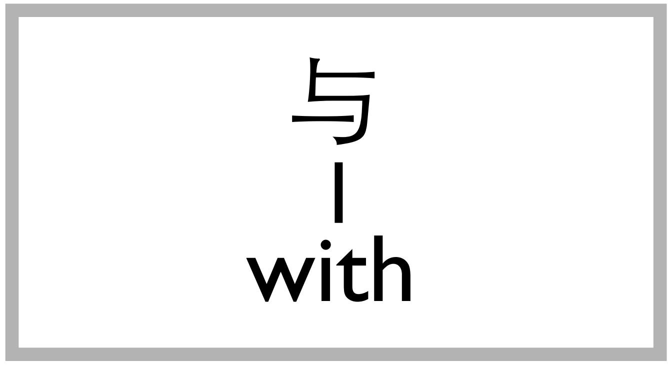
Example

$P(\mathbf{f}, \mathbf{a}|\mathbf{e})$ $P(\mathbf{a}|\mathbf{f}, \mathbf{e})$

0.25



0.25



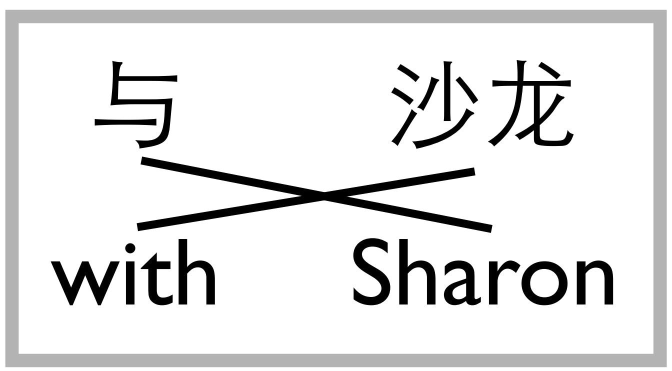
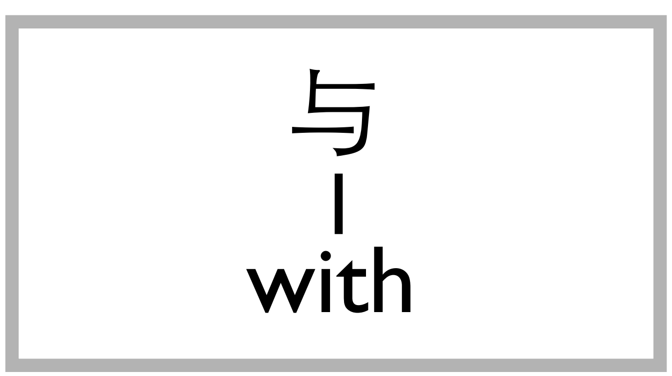
0.5

$$P(\mathbf{a}|\mathbf{f}, \mathbf{e}) = \frac{P(\mathbf{f}, \mathbf{a}|\mathbf{e})}{\sum_{\mathbf{a}'} P(\mathbf{f}, \mathbf{a}'|\mathbf{e})}$$

f	e	tc	t
与	with	N/A	0.5
	Sharon	N/A	0.5
沙龙	with	N/A	0.5
	Sharon	N/A	0.5

(Brown et al., 1993)

Example

	$P(\mathbf{f}, \mathbf{a}\vert\mathbf{e})$	$P(\mathbf{a}\vert\mathbf{f}, \mathbf{e})$
	0.25	0.5
	0.25	
	0.5	

$$P(\mathbf{a}|\mathbf{f}, \mathbf{e}) = \frac{P(\mathbf{f}, \mathbf{a}|\mathbf{e})}{\sum_{\mathbf{a}'} P(\mathbf{f}, \mathbf{a}'|\mathbf{e})}$$

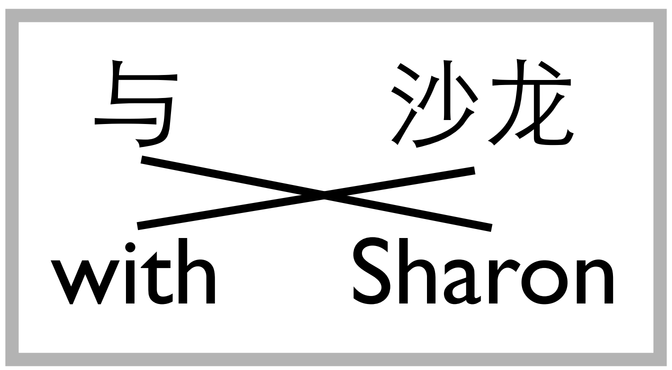
f	e	tc	t
与	with	N/A	0.5
	Sharon	N/A	0.5
沙龙	with	N/A	0.5
	Sharon	N/A	0.5

(Brown et al., 1993)

Example

$P(\mathbf{f}, \mathbf{a}|\mathbf{e})$ $P(\mathbf{a}|\mathbf{f}, \mathbf{e})$

0.25 0.5



0.25 0.5

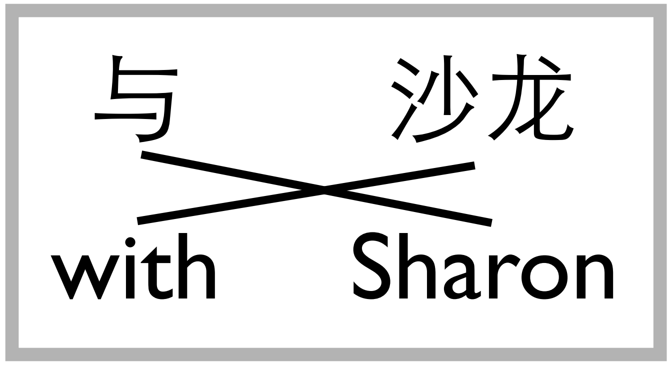
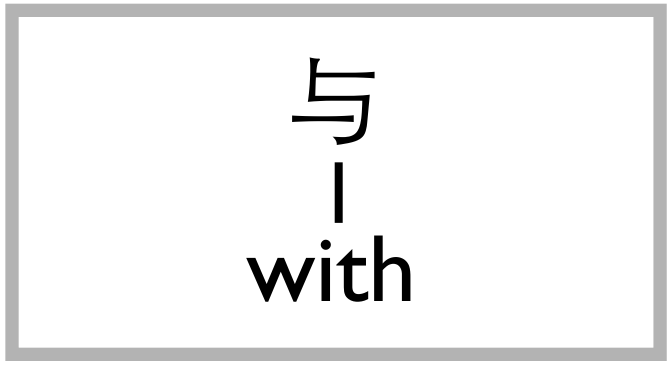
0.5

$$P(\mathbf{a}|\mathbf{f}, \mathbf{e}) = \frac{P(\mathbf{f}, \mathbf{a}|\mathbf{e})}{\sum_{\mathbf{a}'} P(\mathbf{f}, \mathbf{a}'|\mathbf{e})}$$

f	e	tc	t
与	with	N/A	0.5
	Sharon	N/A	0.5
沙龙	with	N/A	0.5
	Sharon	N/A	0.5

(Brown et al., 1993)

Example

	$P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$	$P(\mathbf{a} \mid \mathbf{f}, \mathbf{e})$
	0.25	0.5
	0.25	0.5
	0.5	1.0

$$P(\mathbf{a} \mid \mathbf{f}, \mathbf{e}) = \frac{P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})}{\sum_{\mathbf{a}'} P(\mathbf{f}, \mathbf{a}' \mid \mathbf{e})}$$

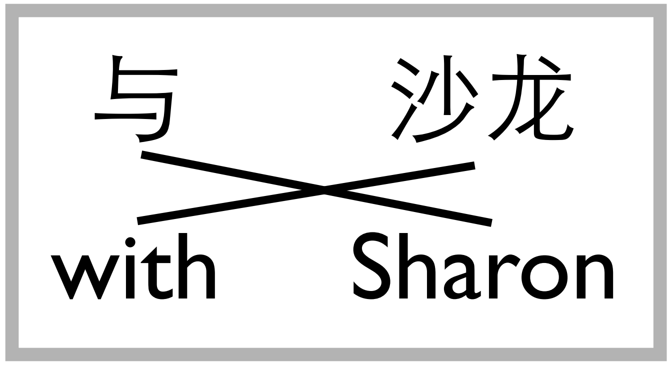
f	e	tc	t
与	with	N/A	0.5
	Sharon	N/A	0.5
沙龙	with	N/A	0.5
	Sharon	N/A	0.5

(Brown et al., 1993)

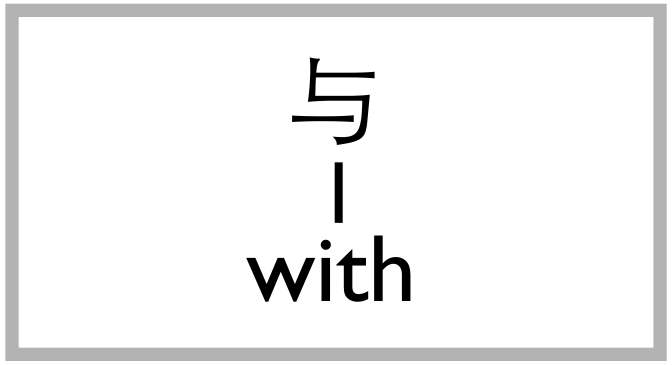
Example

$P(\mathbf{f}, \mathbf{a}|\mathbf{e})$ $P(\mathbf{a}|\mathbf{f}, \mathbf{e})$

0.25 0.5



0.25 0.5

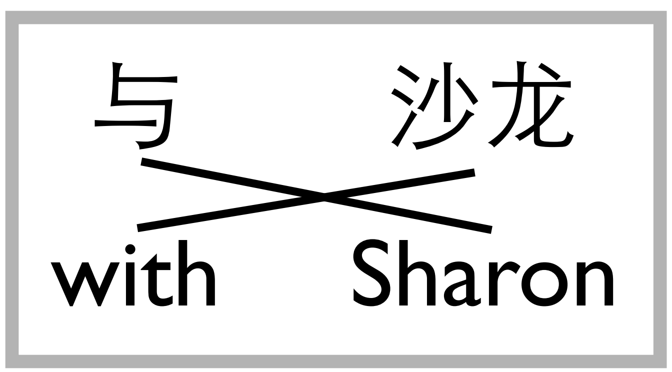
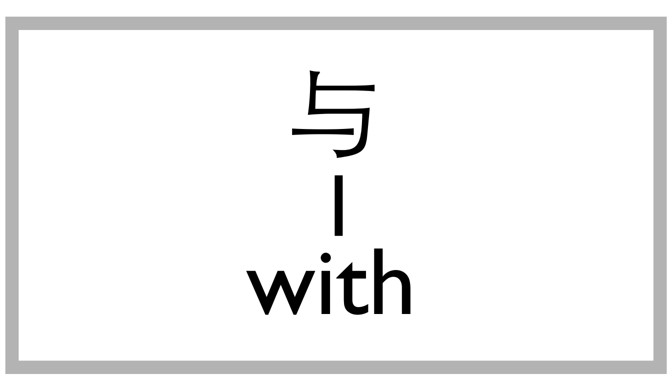


0.5 1.0

f	e	tc	t
与	with	N/A	0.5
	Sharon	N/A	0.5
沙龙	with	N/A	0.5
	Sharon	N/A	0.5

(Brown et al., 1993)

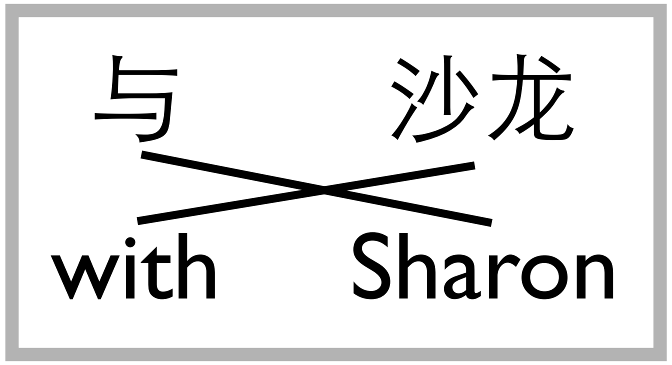
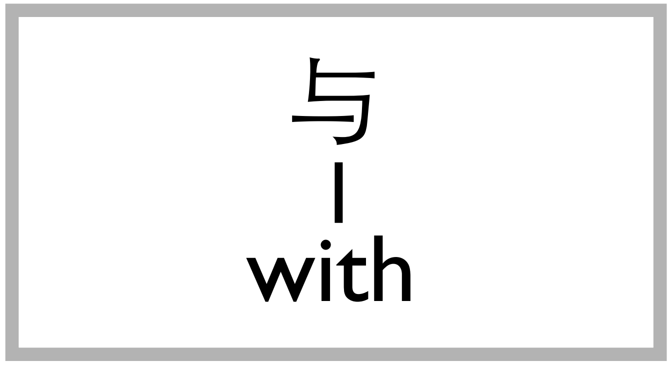
Example

	$P(\mathbf{f},\mathbf{a}\|\mathbf{e})$	$P(\mathbf{a}\|\mathbf{f},\mathbf{e})$
	0.25	0.5
	0.25	0.5
	0.5	1.0

f	e	tc	t
与	with	N/A	0.5
	Sharon	N/A	0.5
沙龙	with	N/A	0.5
	Sharon	N/A	0.5

$$tc(f|e) = \sum_{s=1}^{S} \sum_{\mathbf{a}} P(\mathbf{a}|\mathbf{f}^{(s)}, \mathbf{e}^{(s)}) \sum_{j=1}^{|\mathbf{f}^{(s)}|} \delta(\mathbf{f}_j^{(s)}, f)\delta(\mathbf{e}_{\mathbf{a}_j}^{(s)}, e)$$

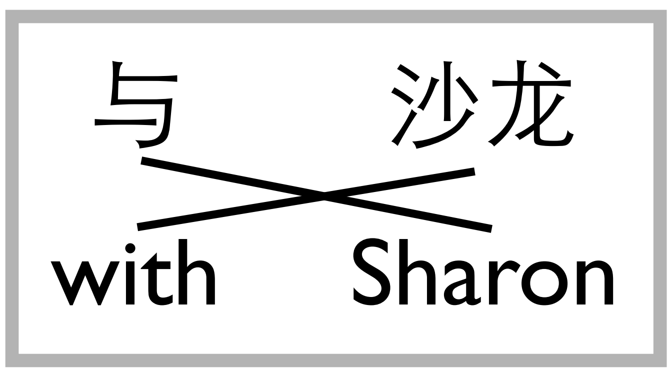
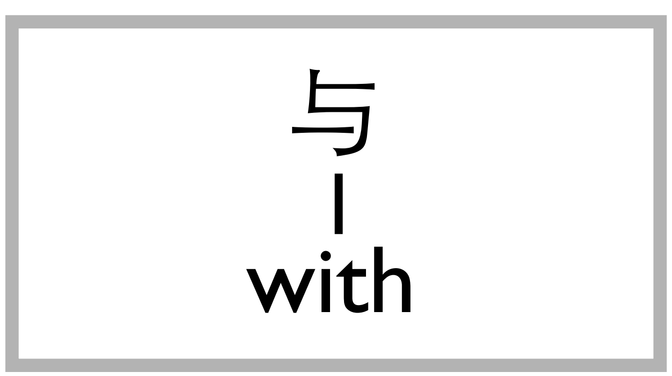
Example

	$P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$	$P(\mathbf{a} \mid \mathbf{f}, \mathbf{e})$
	0.25	0.5
	0.25	0.5
	0.5	1.0

f	e	tc	t
与	with	1.5	0.5
	Sharon	0.5	0.5
沙龙	with	0.5	0.5
	Sharon	0.5	0.5

$$tc(f|e) = \sum_{s=1}^{S} \sum_{\mathbf{a}} P(\mathbf{a}|\mathbf{f}^{(s)}, \mathbf{e}^{(s)}) \sum_{j=1}^{|\mathbf{f}^{(s)}|} \delta(\mathbf{f}_j^{(s)}, f)\delta(\mathbf{e}_{\mathbf{a}_j}^{(s)}, e)$$

Example

	$P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$	$P(\mathbf{a} \mid \mathbf{f}, \mathbf{e})$
	0.25	0.5
	0.25	0.5
	0.5	1.0

f	e	tc	t
与	with	1.5	0.5
	Sharon	0.5	0.5
沙龙	with	0.5	0.5
	Sharon	0.5	0.5

Example

Alignment	$P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$	$P(\mathbf{a} \mid \mathbf{f}, \mathbf{e})$
与 沙龙 / with Sharon	0.25	0.5
与 沙龙 / with Sharon	0.25	0.5
与 / with	0.5	1.0

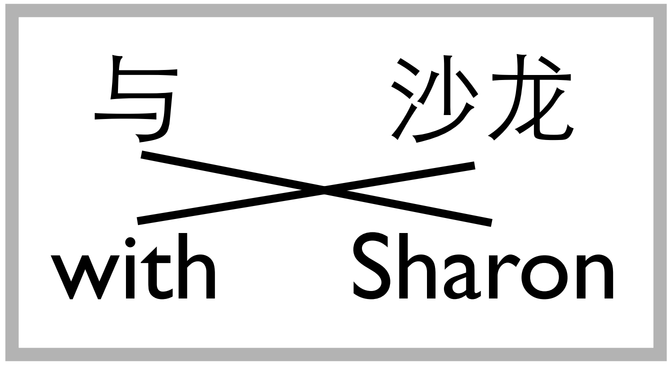
$$t(f|e) = \frac{tc(f|e)}{\sum_{f'} tc(f'|e)}$$

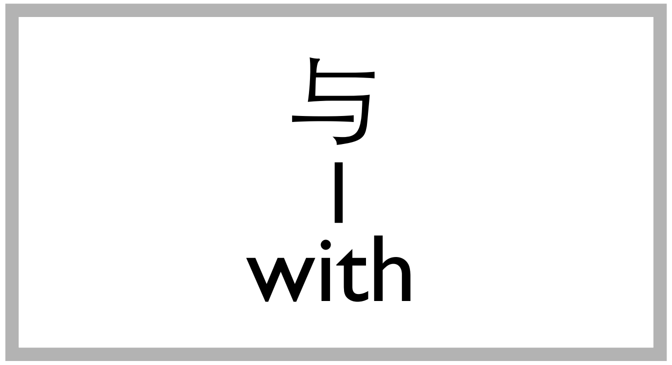
f	e	tc	t
与	with	1.5	0.5
	Sharon	0.5	0.5
沙龙	with	0.5	0.5
	Sharon	0.5	0.5

(Brown et al., 1993)

Example

Alignment	$P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$	$P(\mathbf{a} \mid \mathbf{f}, \mathbf{e})$
(1)	0.25	0.5
(2)	0.25	0.5
(3)	0.5	1.0



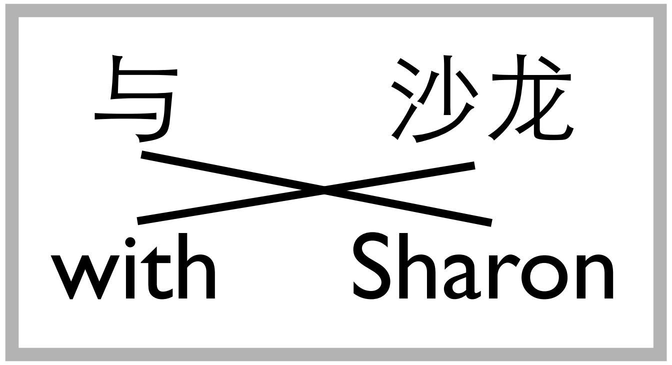
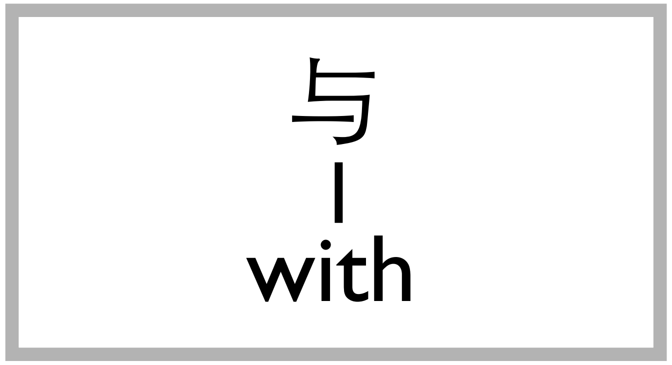


$$t(f|e) = \frac{tc(f|e)}{\sum_{f'} tc(f'|e)}$$

f	e	tc	t
与	with	1.5	0.75
	Sharon	0.5	0.25
沙龙	with	0.5	0.5
	Sharon	0.5	0.5

(Brown et al., 1993)

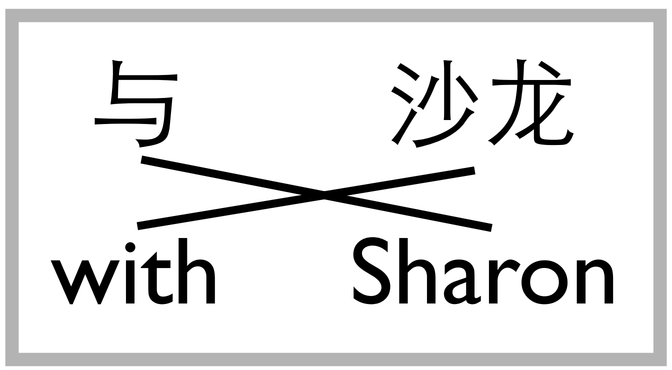
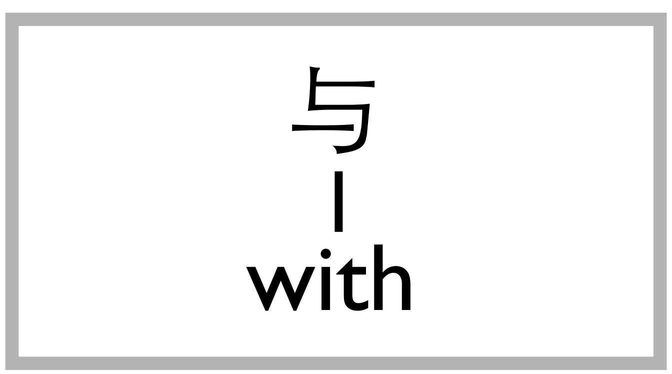
Example

	$P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$	$P(\mathbf{a} \mid \mathbf{f}, \mathbf{e})$
	0.25	0.5
	0.25	0.5
	0.5	1.0

f	e	tc	t
与	with	1.5	0.75
	Sharon	0.5	0.25
沙龙	with	0.5	0.5
	Sharon	0.5	0.5

(Brown et al., 1993)

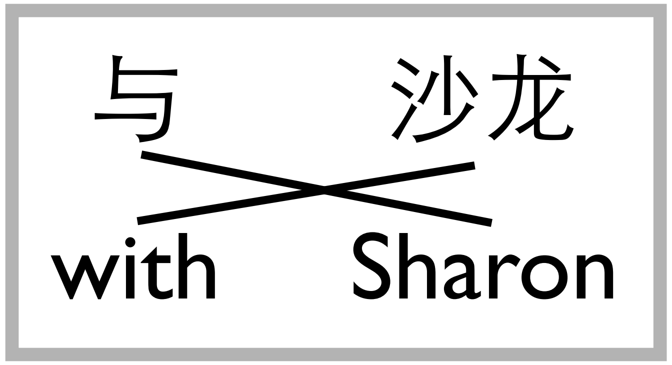
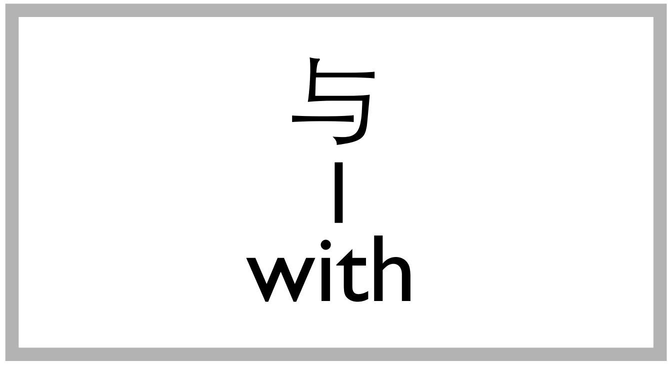
Example

	$P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$	$P(\mathbf{a} \mid \mathbf{f}, \mathbf{e})$
	0.375	0.5
	0.125	0.5
	0.75	1.0

f	e	tc	t
与	with	1.5	0.75
	Sharon	0.5	0.25
沙龙	with	0.5	0.5
	Sharon	0.5	0.5

(Brown et al., 1993)

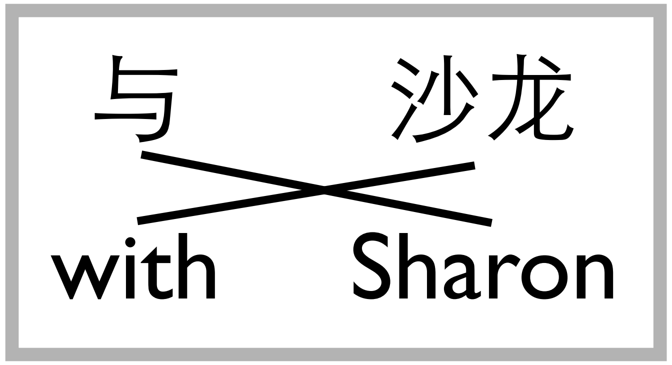
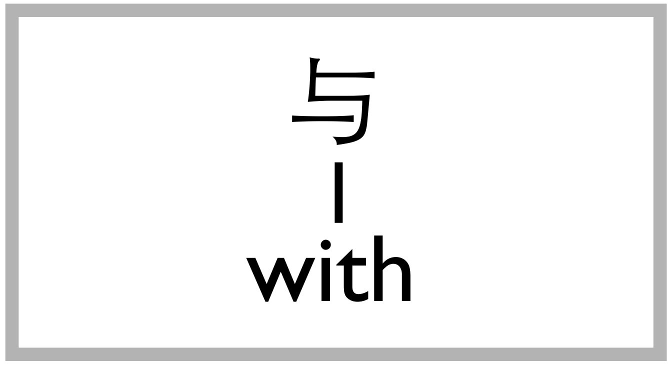
Example

	$P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$	$P(\mathbf{a} \mid \mathbf{f}, \mathbf{e})$
	0.375	0.75
	0.125	0.25
	0.75	1.0

f	e	tc	t
与	with	1.5	0.75
	Sharon	0.5	0.25
沙龙	with	0.5	0.5
	Sharon	0.5	0.5

(Brown et al., 1993)

Example

	$P(\mathbf{f},\mathbf{a}\mid\mathbf{e})$	$P(\mathbf{a}\mid\mathbf{f},\mathbf{e})$
	0.375	0.75
	0.125	0.25
	0.75	1.0

f	e	tc	t
与	with	1.75	0.75
	Sharon	0.25	0.25
沙龙	with	0.25	0.5
	Sharon	0.75	0.5

(Brown et al., 1993)

Example

Alignment	$P(\mathbf{f}, \mathbf{a} \mid \mathbf{e})$	$P(\mathbf{a} \mid \mathbf{f}, \mathbf{e})$
与 – with, 沙龙 – Sharon	0.375	0.75
与 – Sharon, 沙龙 – with	0.125	0.25
与 – with	0.75	1.0

f	e	tc	t
与	with	1.75	0.875
	Sharon	0.25	0.125
沙龙	with	0.25	0.25
	Sharon	0.75	0.75

(Brown et al., 1993)

Problems with EM

- Initialization is important as EM is prone to get stuck in local optima
- Solution: use the output of simpler models as the input of training more complex models

(Brown et al., 1993)

Problems with EM

- Initialization is important as EM is prone to get stuck in local optima
- Solution: use the output of simpler models as the input of training more complex models

model 1

(Brown et al., 1993)

Problems with EM

- Initialization is important as EM is prone to get stuck in local optima
- Solution: use the output of simpler models as the input of training more complex models

model 1

5

(Brown et al., 1993)

Problems with EM

- Initialization is important as EM is prone to get stuck in local optima
- Solution: use the output of simpler models as the input of training more complex models

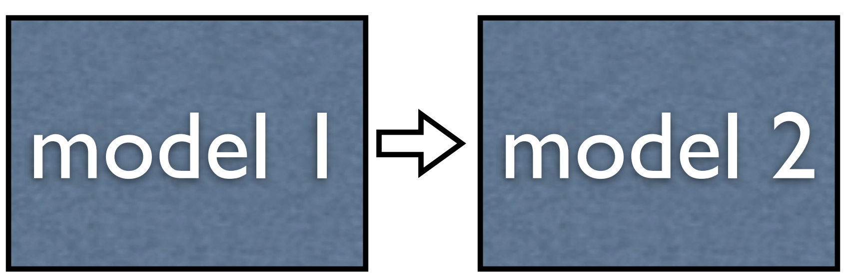
model 1 ⇨ model 2

5

(Brown et al., 1993)

Problems with EM

- Initialization is important as EM is prone to get stuck in local optima
- Solution: use the output of simpler models as the input of training more complex models



5 5

(Brown et al., 1993)

Problems with EM

- Initialization is important as EM is prone to get stuck in local optima
- Solution: use the output of simpler models as the input of training more complex models

model 1 ⇨ model 2 ⇨ model 3

5 5

(Brown et al., 1993)

Problems with EM

- Initialization is important as EM is prone to get stuck in local optima
- Solution: use the output of simpler models as the input of training more complex models

model 1 ⇨ model 2 ⇨ model 3

5 5 3

(Brown et al., 1993)

Problems with EM

- Initialization is important as EM is prone to get stuck in local optima
- Solution: use the output of simpler models as the input of training more complex models

model 1 ⇨ model 2 ⇨ model 3 ⇨ model 4

5 5 3

(Brown et al., 1993)

Problems with EM

- Initialization is important as EM is prone to get stuck in local optima
- Solution: use the output of simpler models as the input of training more complex models

model 1	⇨	model 2	⇨	model 3	⇨	model 4
5		5		3		3

(Brown et al., 1993)

Problems with EM

- Initialization is important as EM is prone to get stuck in local optima
- Solution: use the output of simpler models as the input of training more complex models

model 1 ⇨ model 2 ⇨ model 3 ⇨ model 4 ⇨ model 5

5 5 3 3

(Brown et al., 1993)

Problems with EM

- Initialization is important as EM is prone to get stuck in local optima
- Solution: use the output of simpler models as the input of training more complex models

model 1 ⇨	model 2 ⇨	model 3 ⇨	model 4 ⇨	model 5
5	5	3	3	3

(Brown et al., 1993)

JHU Workshop

- Kevin Knight led a team to develop open-source toolkits for IBM Models in the 1999 JHU Workshop
- Franz Och wrote GIZA++, the trainer of IBM models

Kevin Knight

Franz Och

Problems with Word-based MT

Problems with Word-based MT

Problems with Word-based MT

Problems with Word-based MT

hard to include context

Part 3: Phrase-based MT

Phrase-based Model

Bush held a talk with Sharon

(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Model

Bush	held a talk	with Sharon

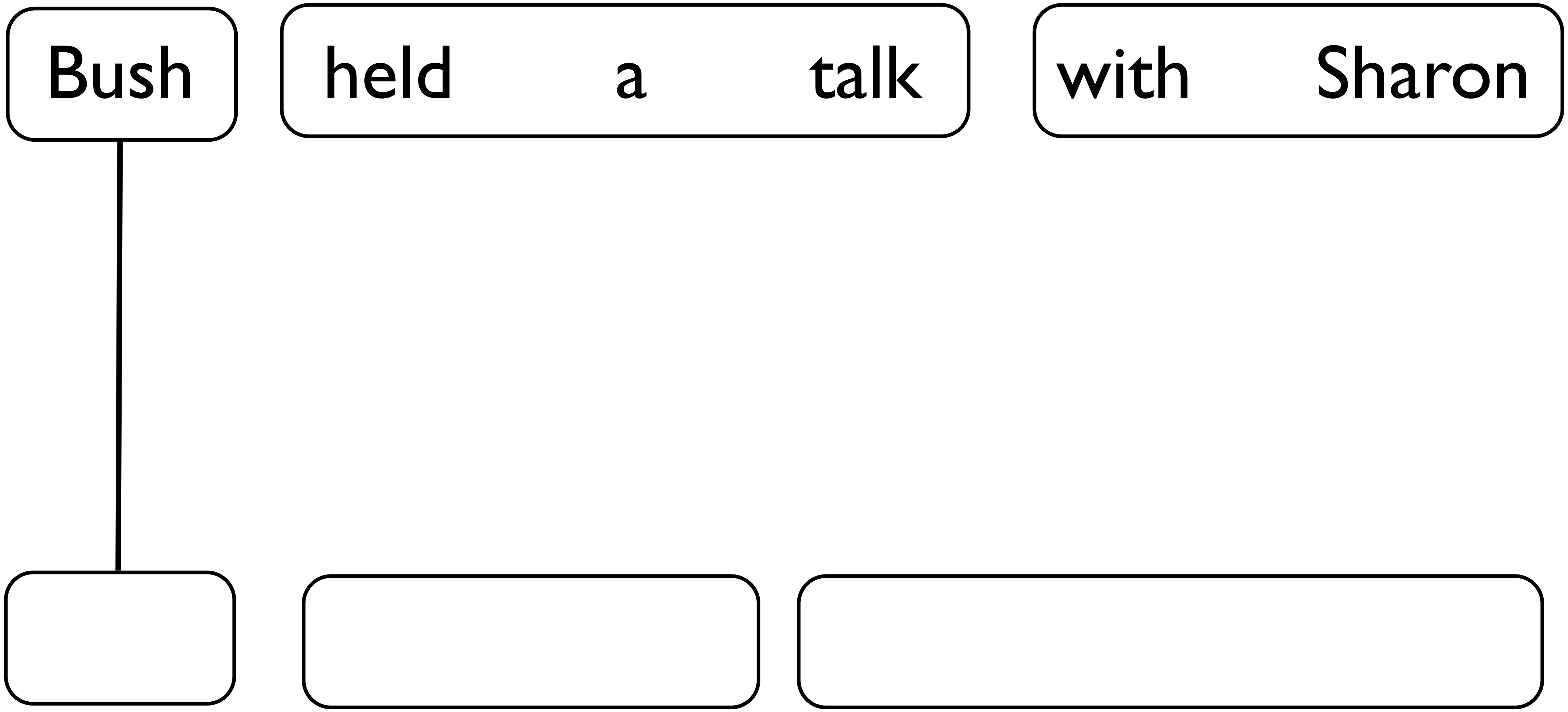
(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Model

Bush | held a talk | with Sharon

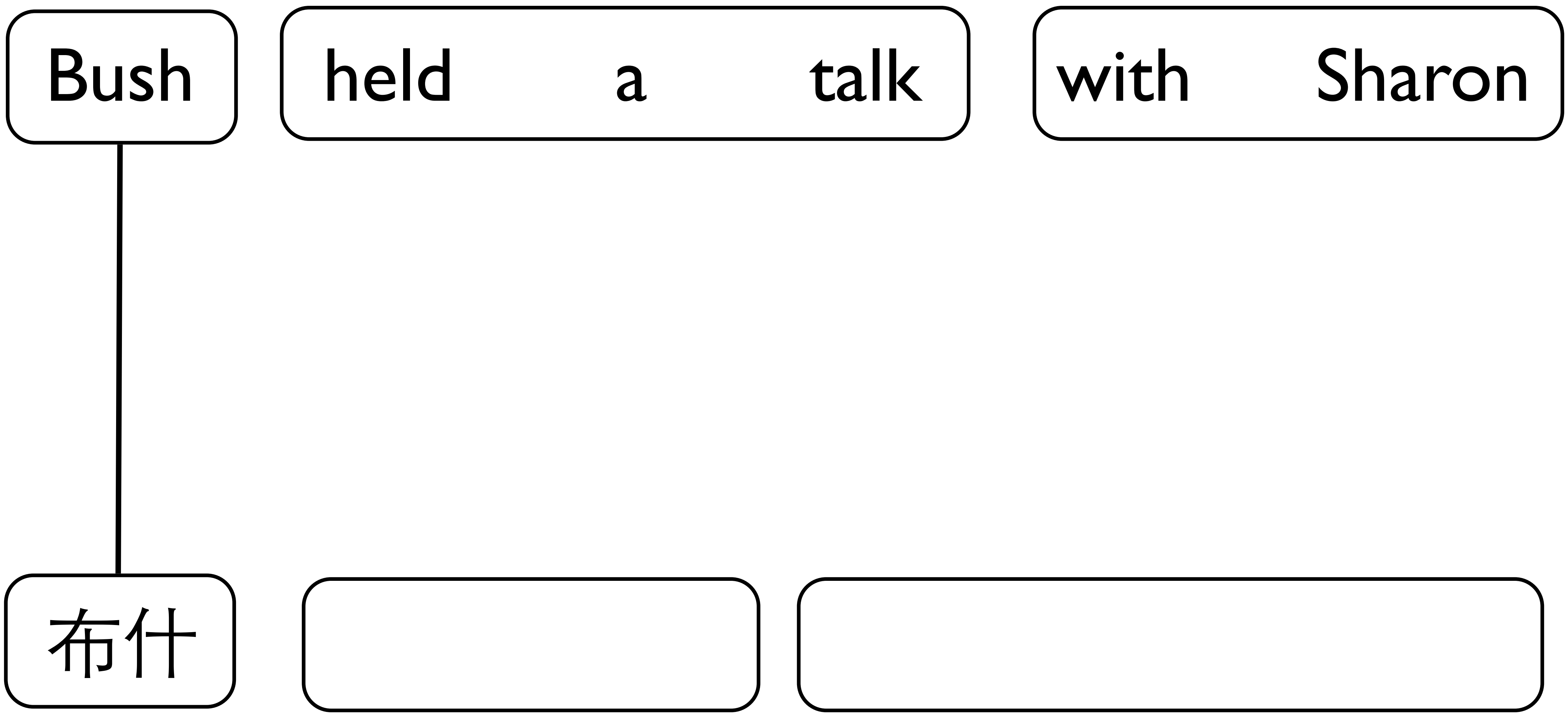
(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Model



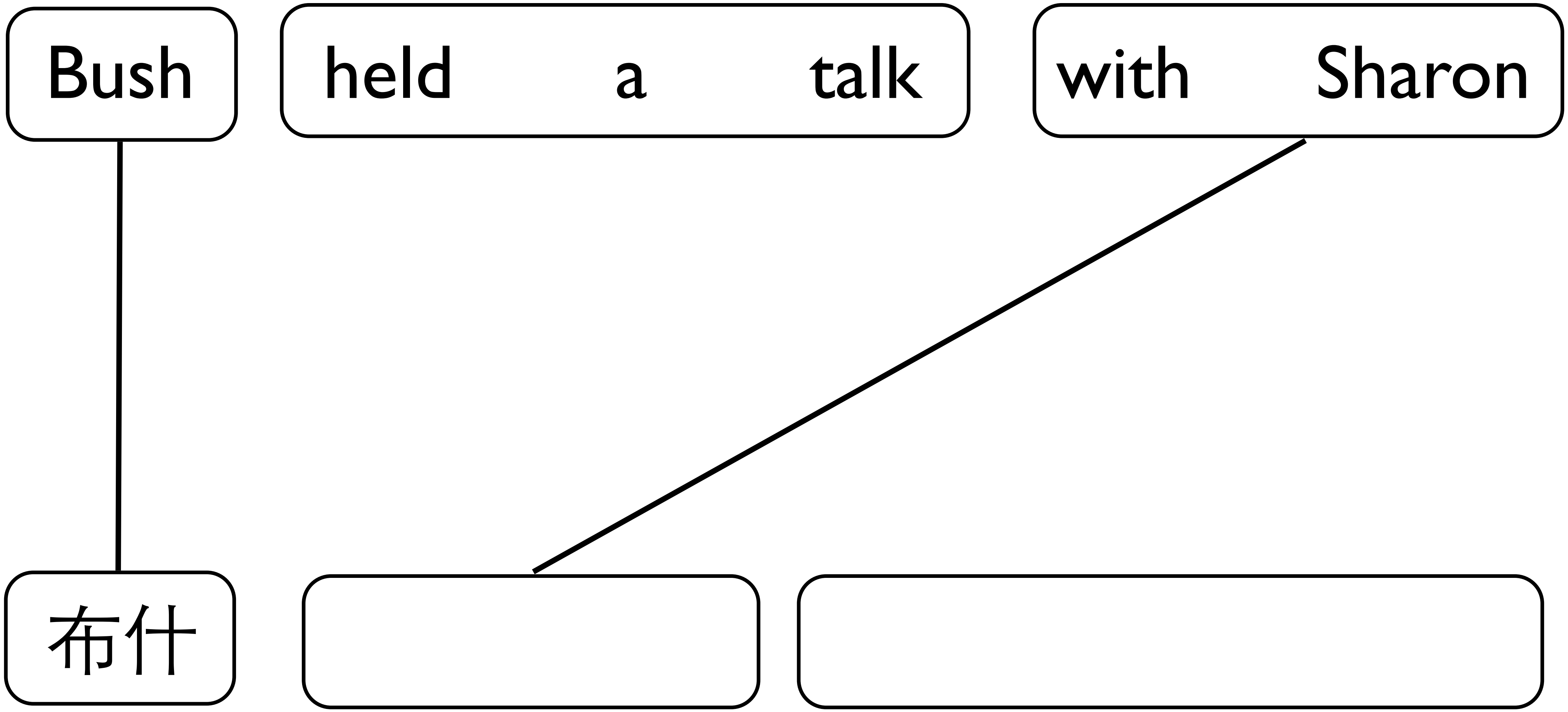
(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Model



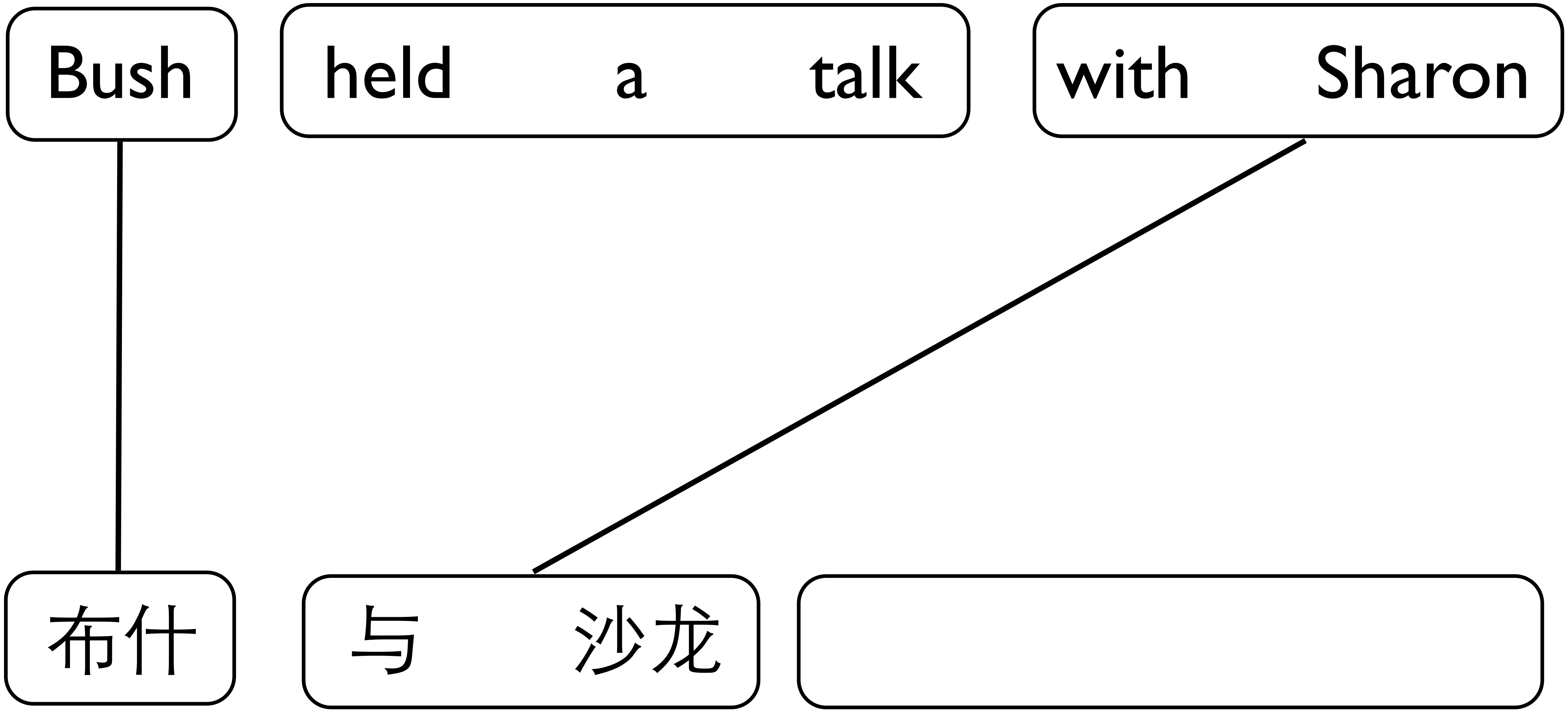
(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Model



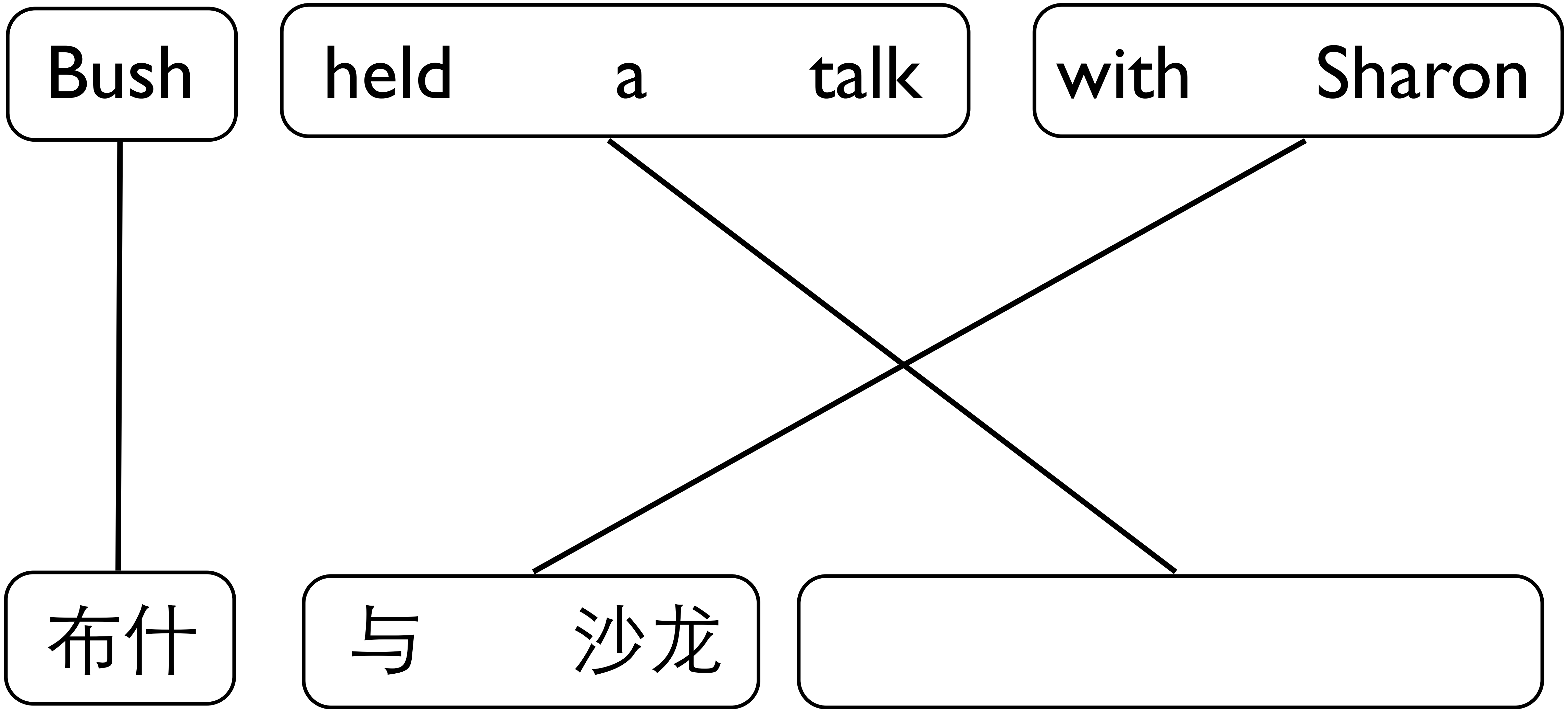
(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Model



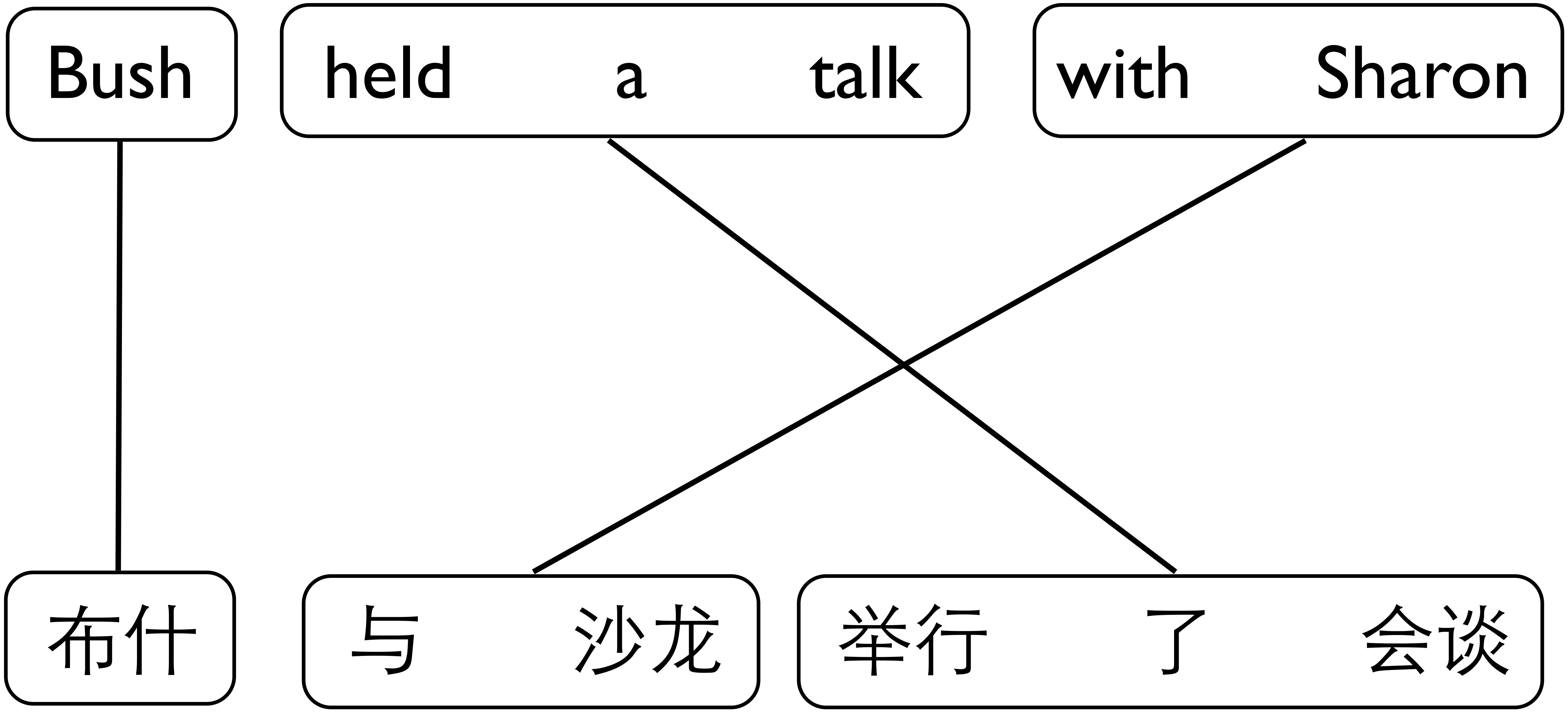
(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Model



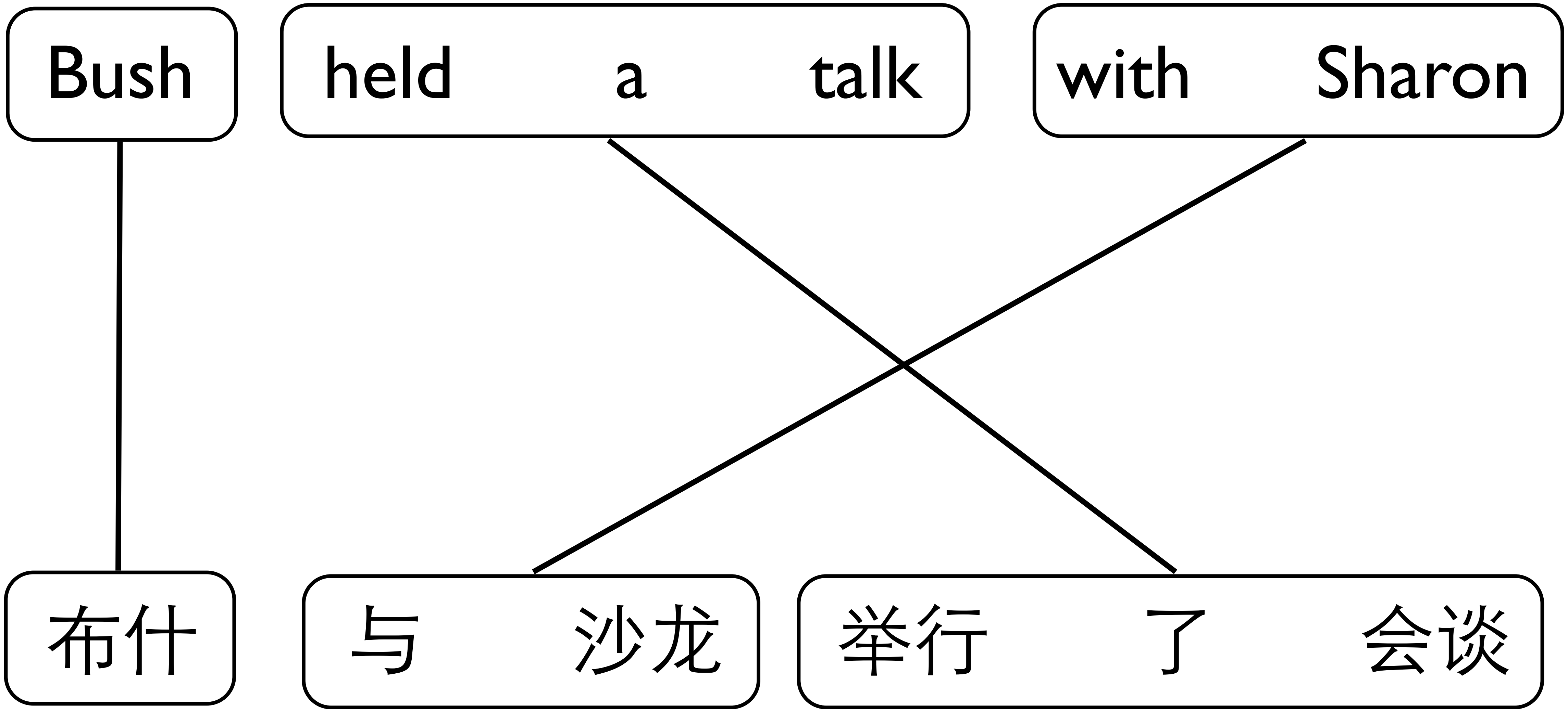
(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Model



(Koehn et al., 2003; Och and Ney, 2004)

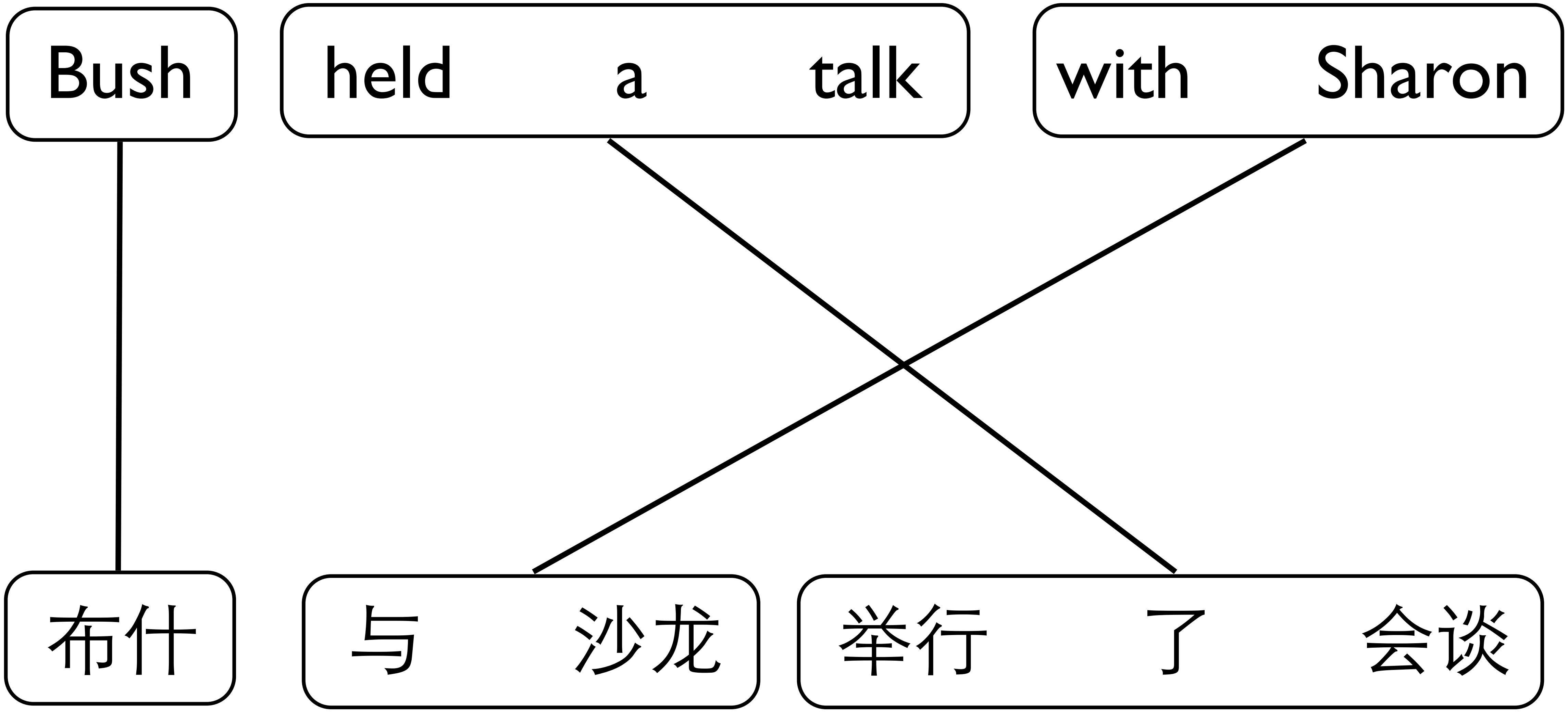
Phrase-based Model



segmentation

(Koehn et al., 2003; Och and Ney, 2004)

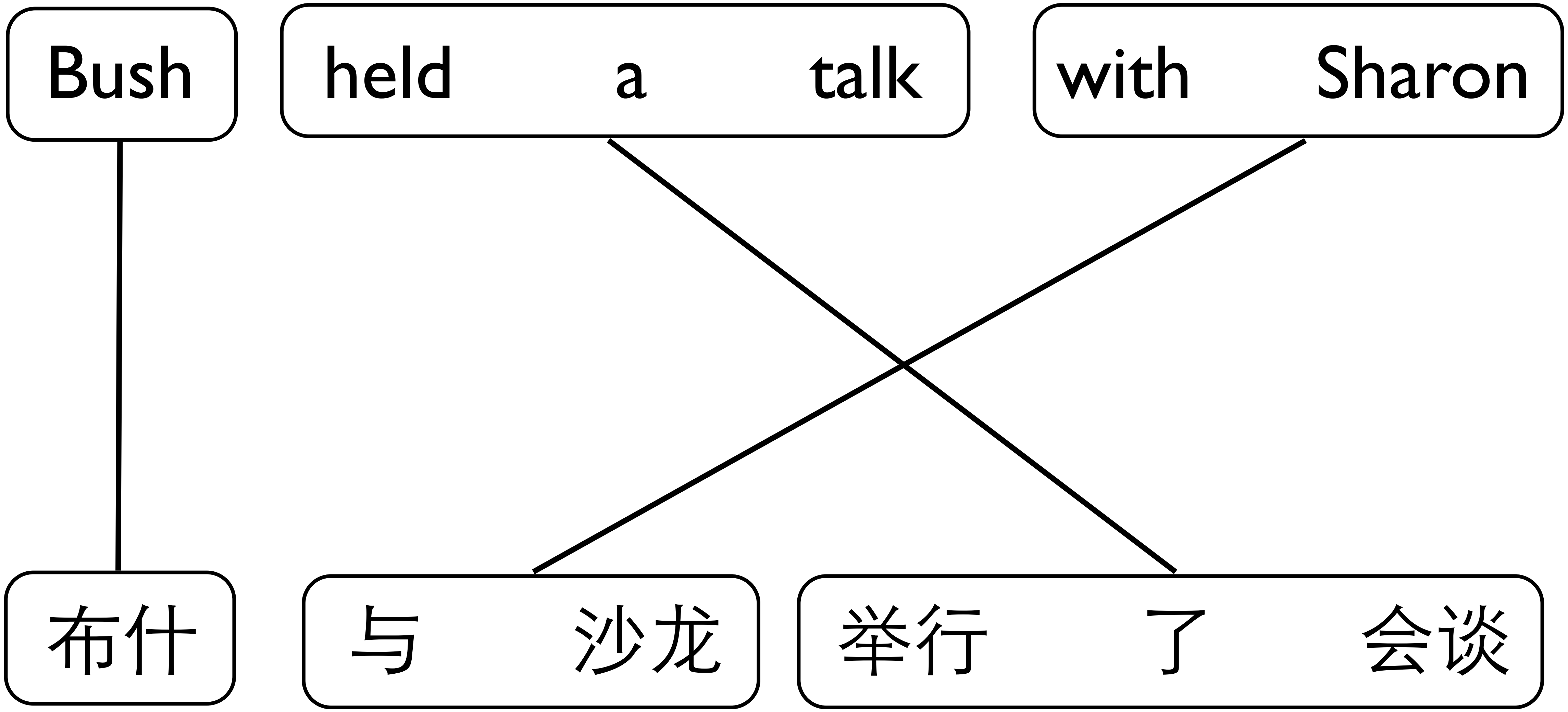
Phrase-based Model



segmentation reordering

(Koehn et al., 2003; Och and Ney, 2004)

Phrase-based Model



segmentation reordering translation

(Koehn et al., 2003; Och and Ney, 2004)

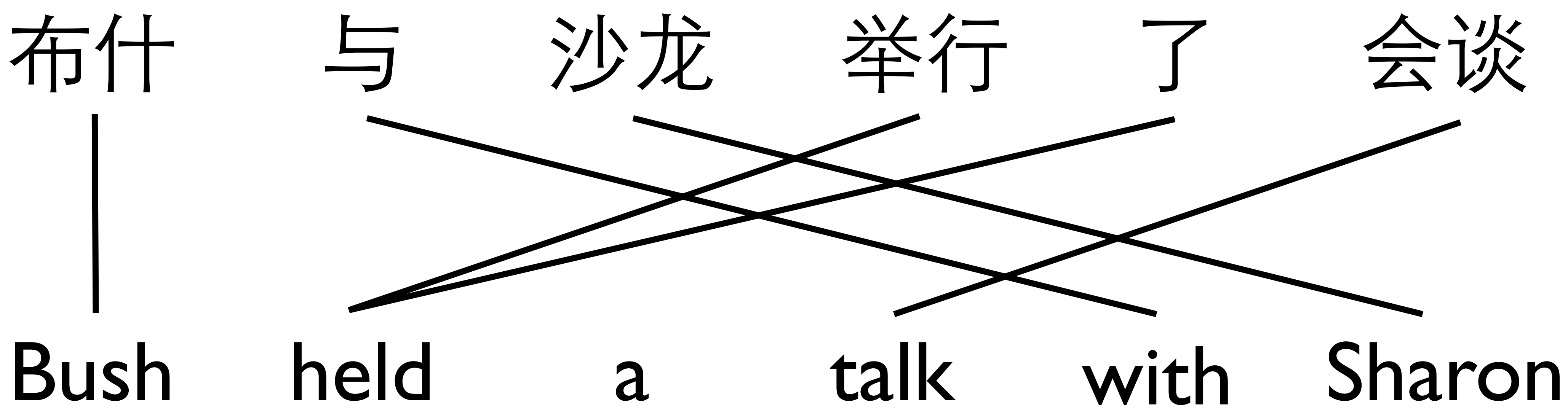
Phrase Extraction

布什　　与　　沙龙　　举行　　了　　会谈

Bush　　held　　a　　talk　　with　　Sharon

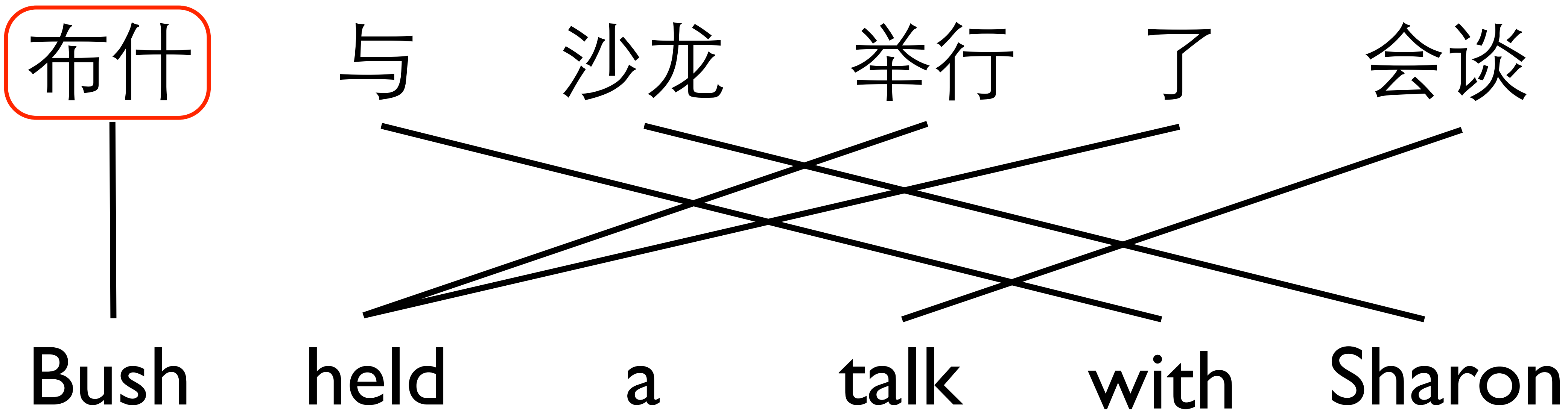
(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction



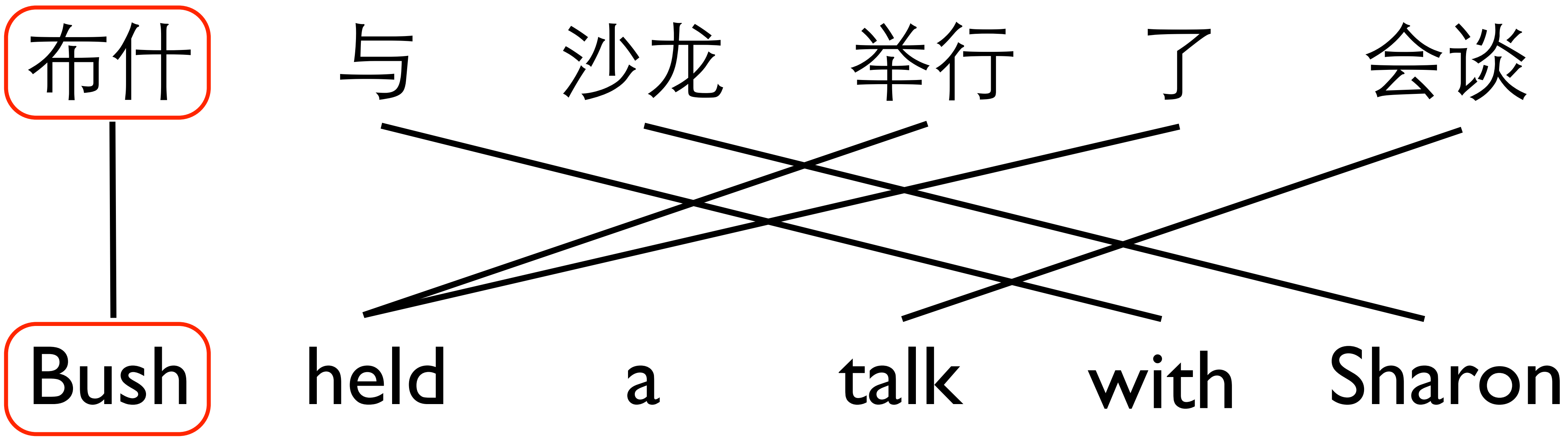
(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction



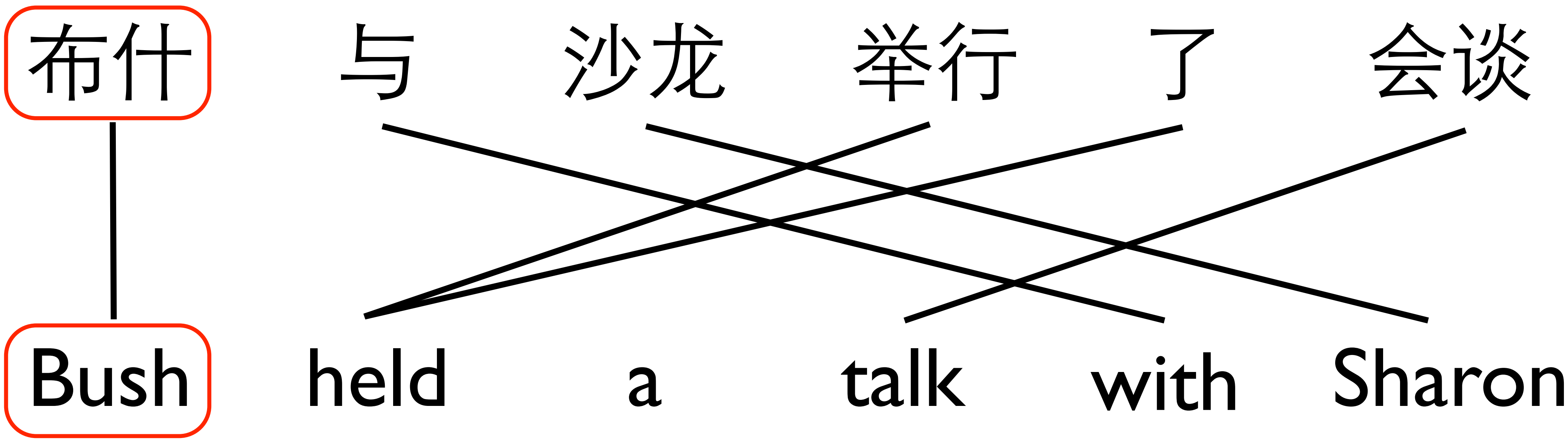
(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction



(Koehn et al., 2003; Och and Ney, 2004)

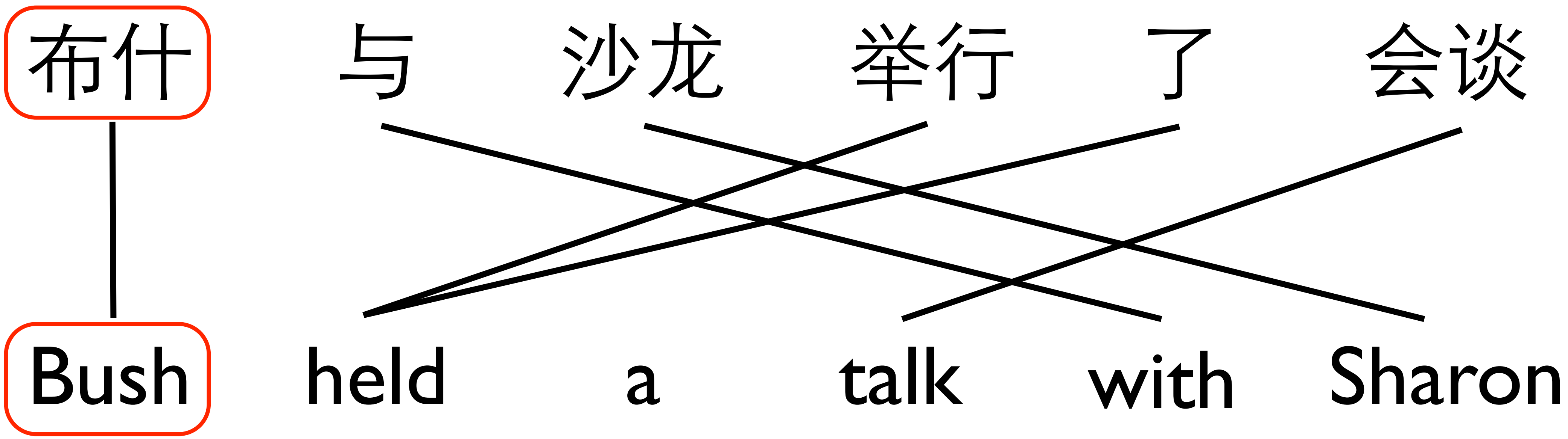
Phrase Extraction



(布什, Bush)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction

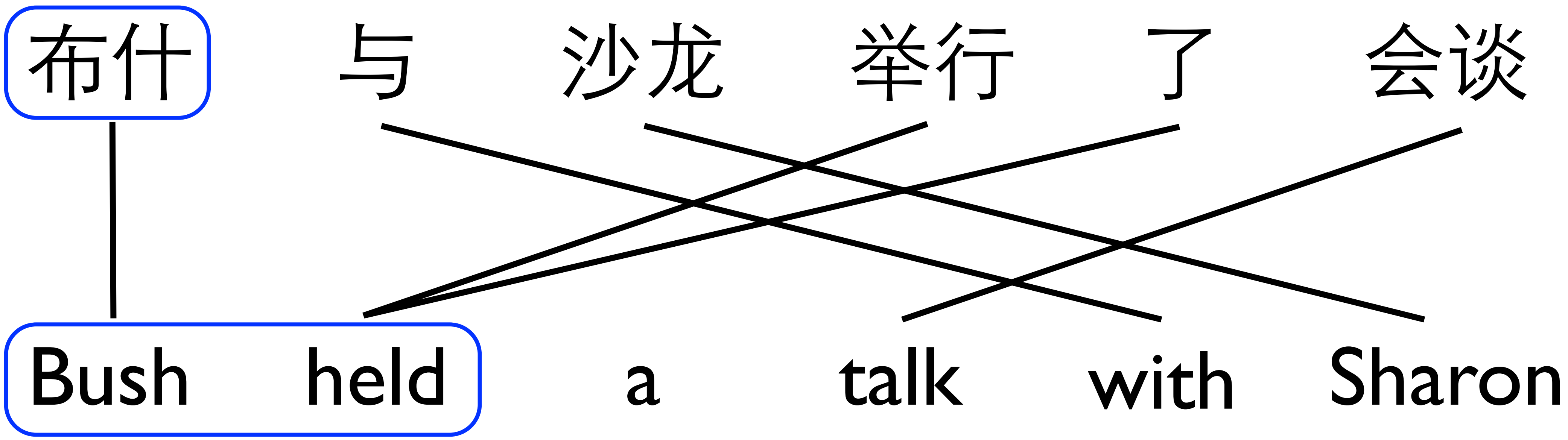


(布什, Bush)

all words in the source phrase are aligned to all words in the target phrase

(Koehn et al., 2003; Och and Ney, 2004)

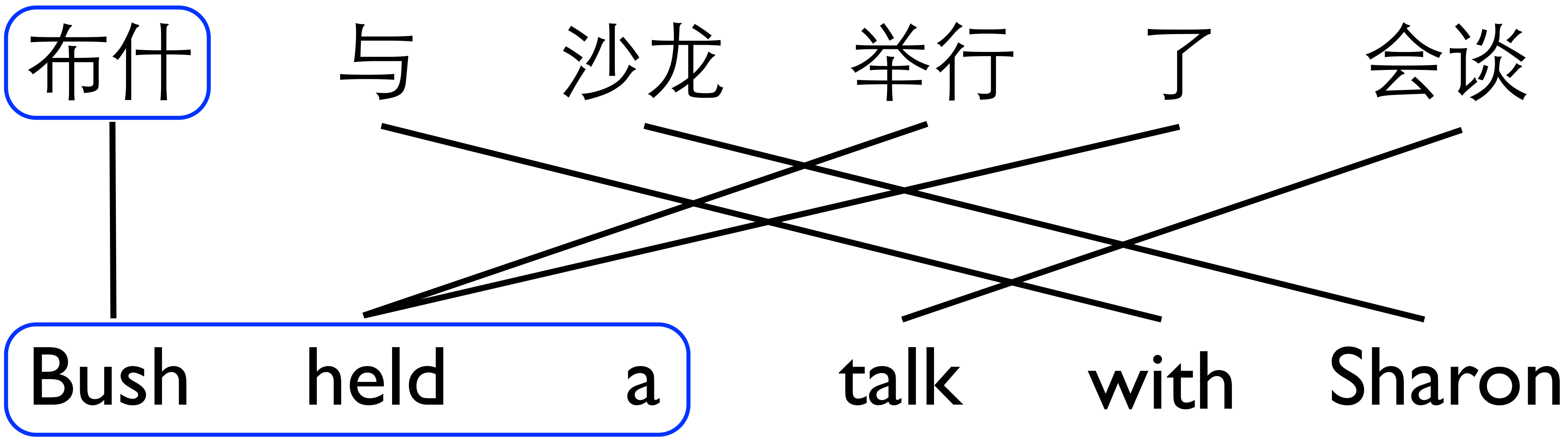
Phrase Extraction



(布什, Bush)

(Koehn et al., 2003; Och and Ney, 2004)

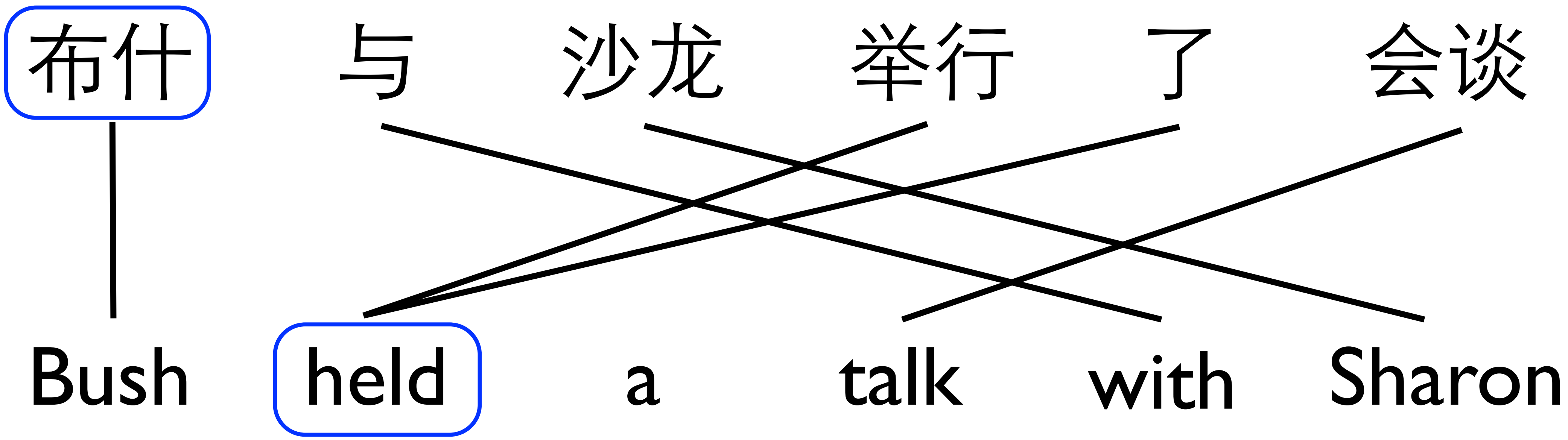
Phrase Extraction



(布什, Bush)

(Koehn et al., 2003; Och and Ney, 2004)

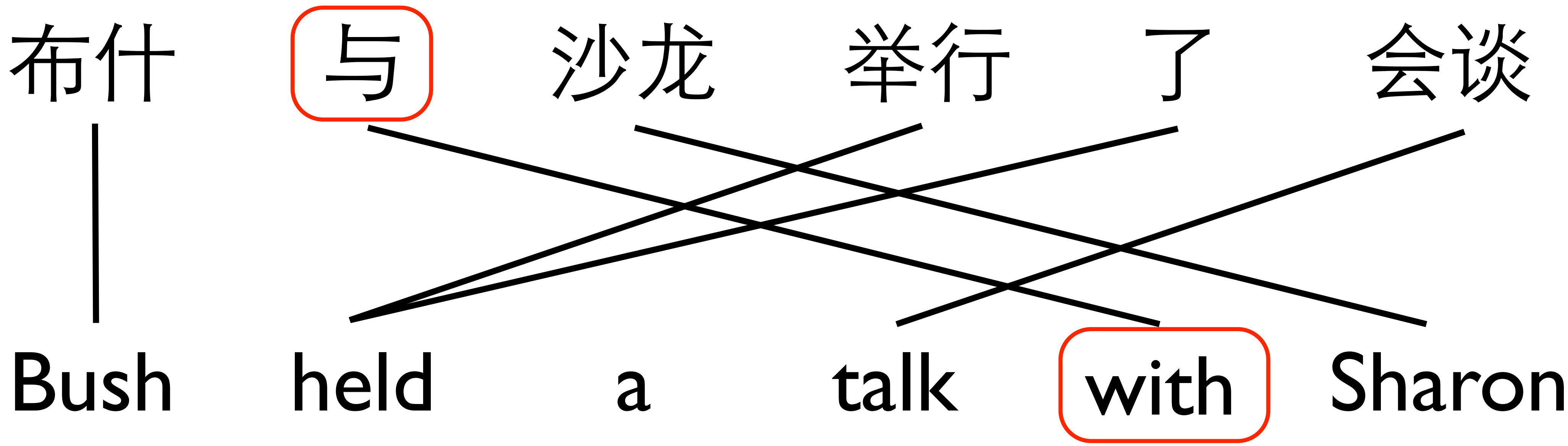
Phrase Extraction



(布什, Bush)

(Koehn et al., 2003; Och and Ney, 2004)

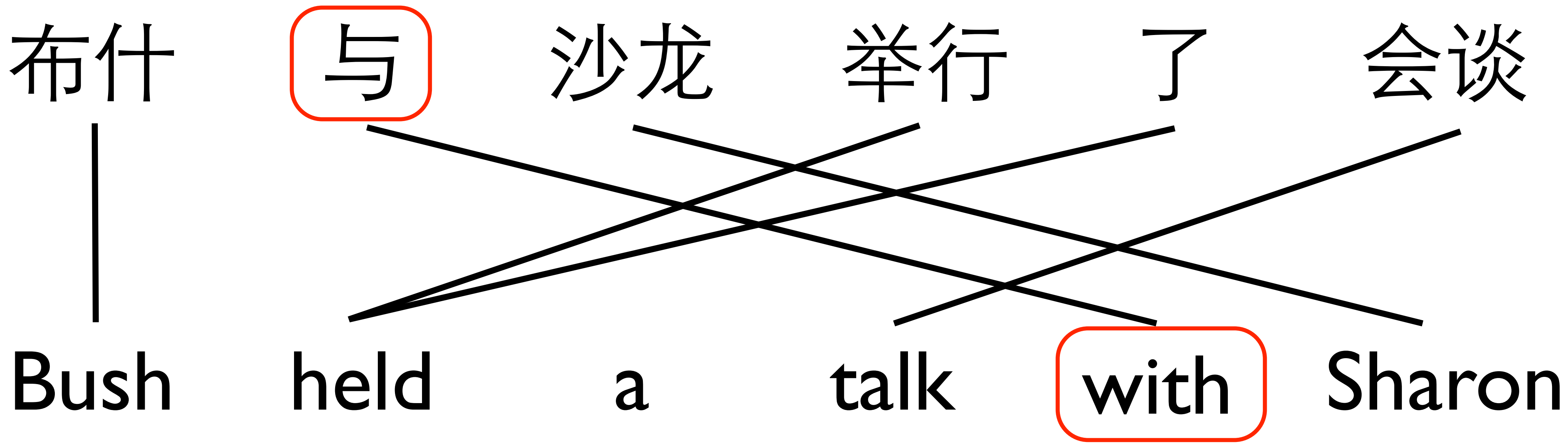
Phrase Extraction



(布什, Bush)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction

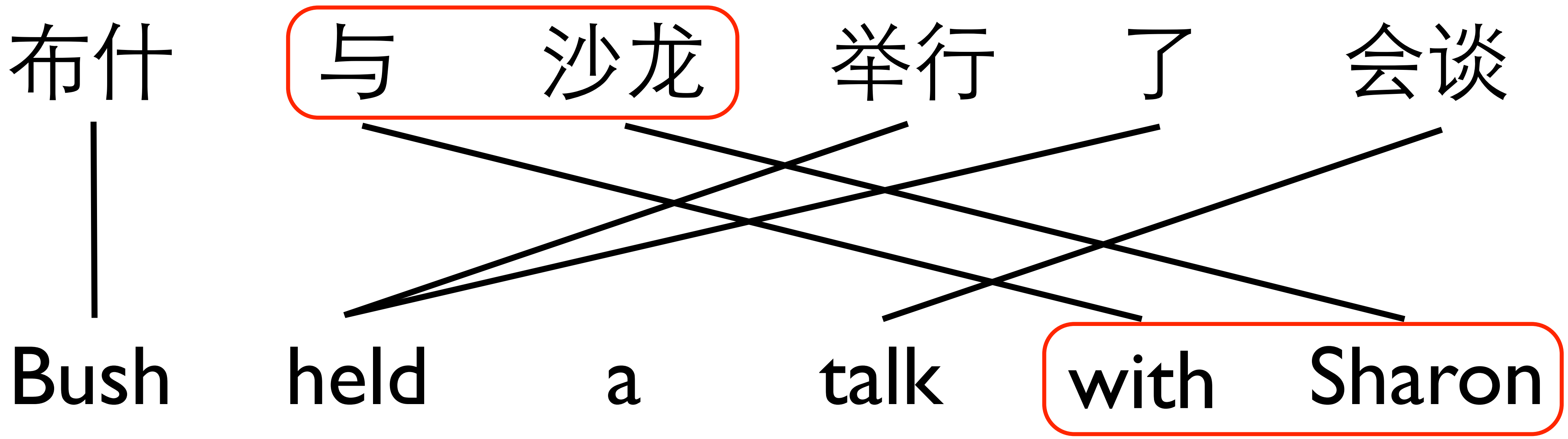


(布什, Bush)

(与, with)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction

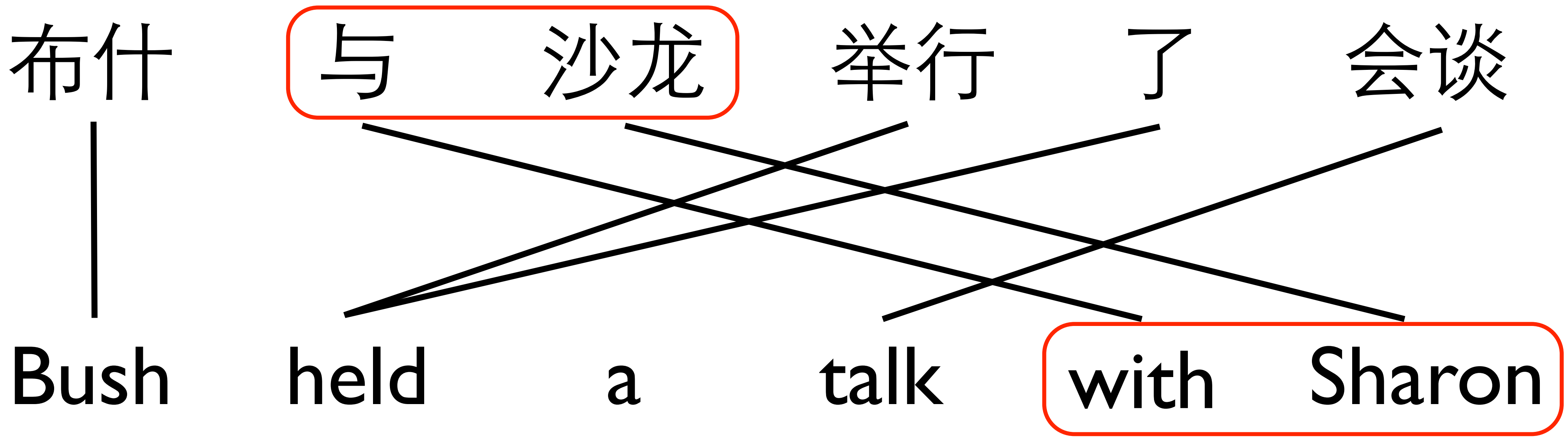


(布什, Bush)

(与, with)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction



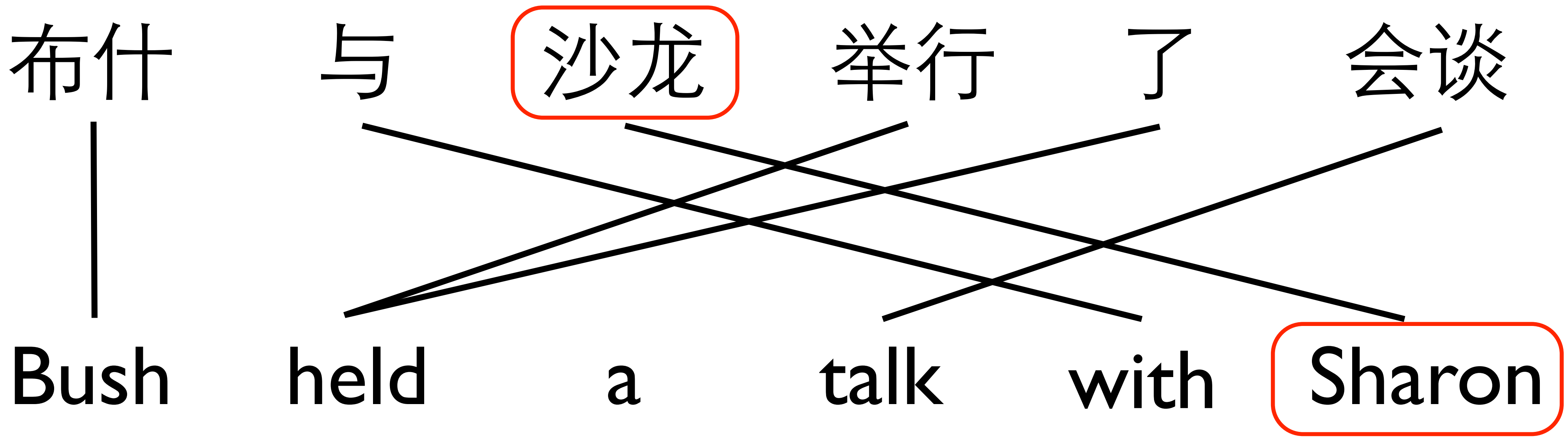
(布什, Bush)

(与, with)

(与 沙龙, with Sharon)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction



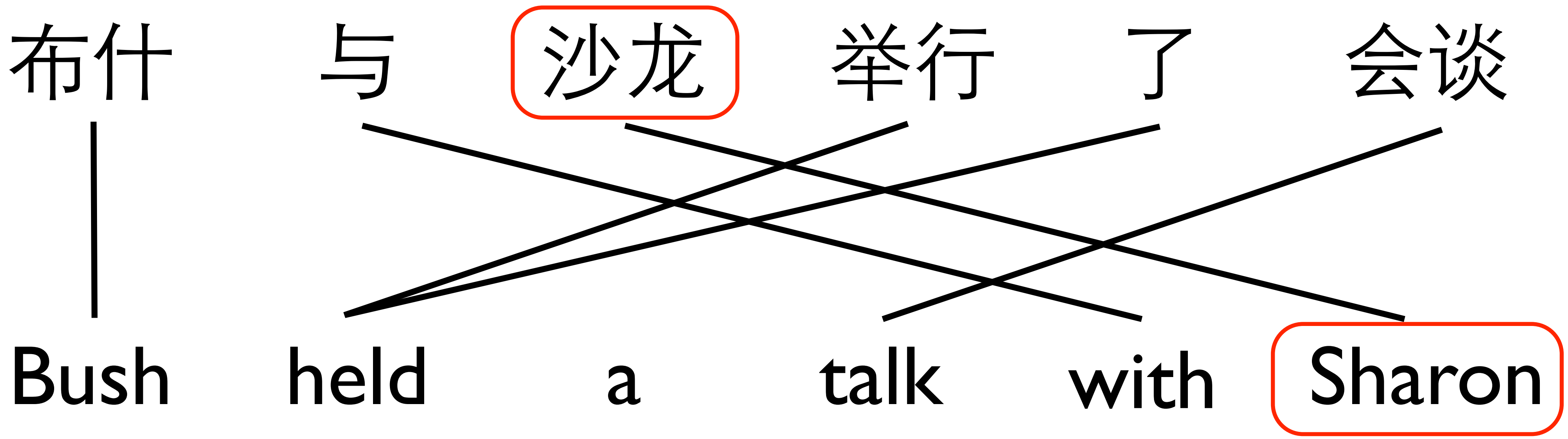
(布什, Bush)

(与, with)

(与 沙龙, with Sharon)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction



(布什, Bush)

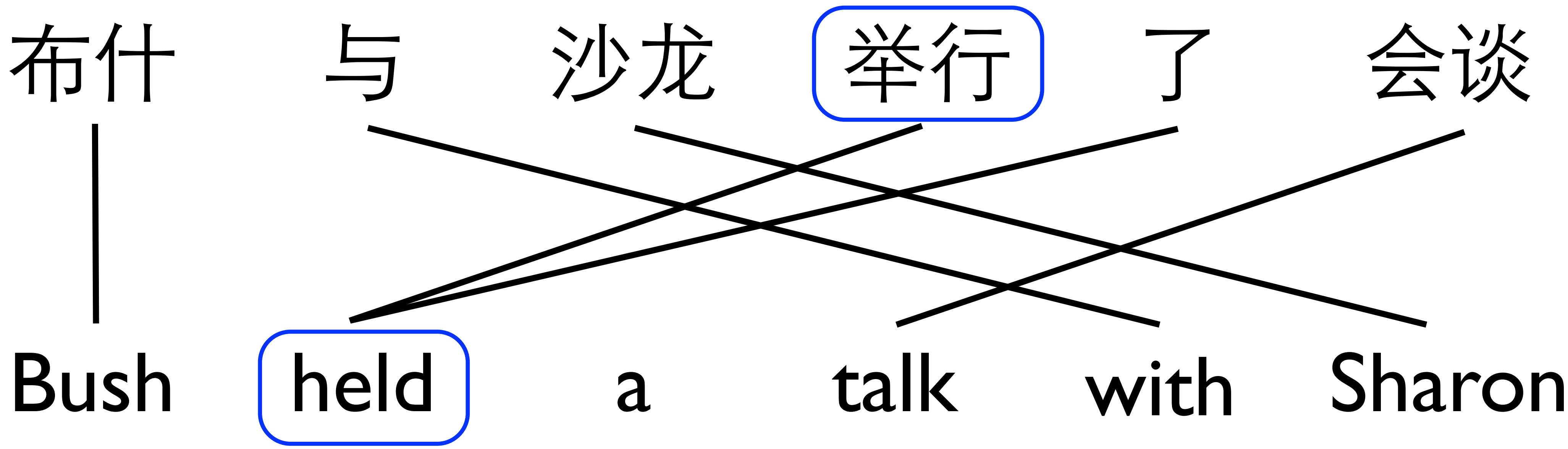
(与, with)

(与 沙龙, with Sharon)

(沙龙, Sharon)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction



(布什, Bush)

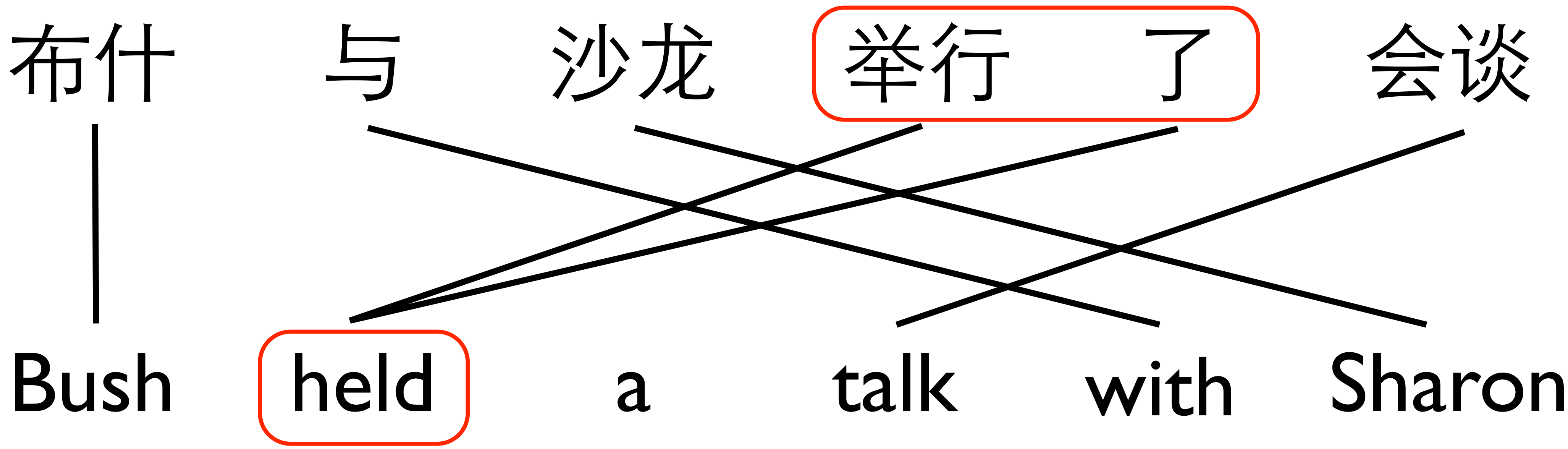
(与, with)

(与 沙龙, with Sharon)

(沙龙, Sharon)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction



(布什, Bush)

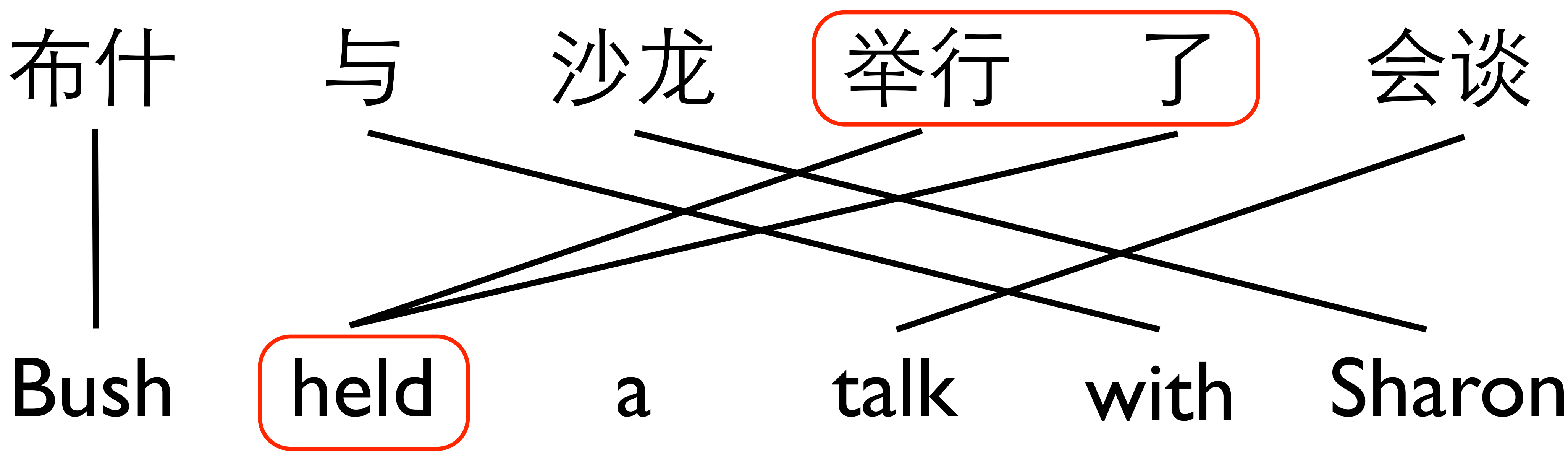
(与, with)

(与 沙龙, with Sharon)

(沙龙, Sharon)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction



(布什, Bush)

(与, with)

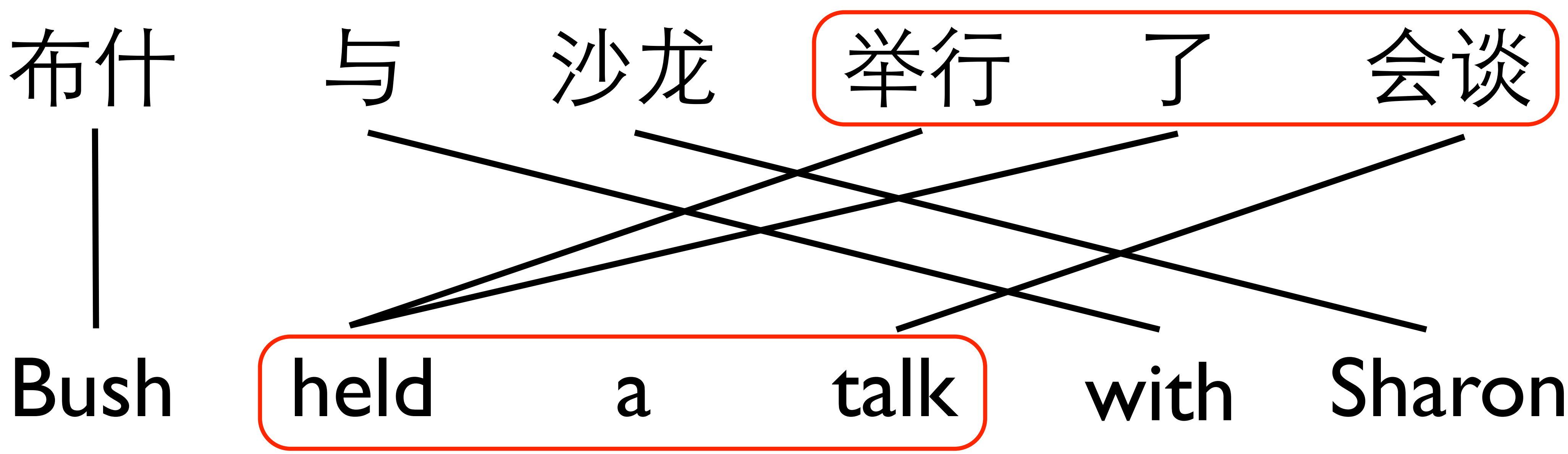
(与 沙龙, with Sharon)

(沙龙, Sharon)

(举行 了, held)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction



(布什, Bush)

(与, with)

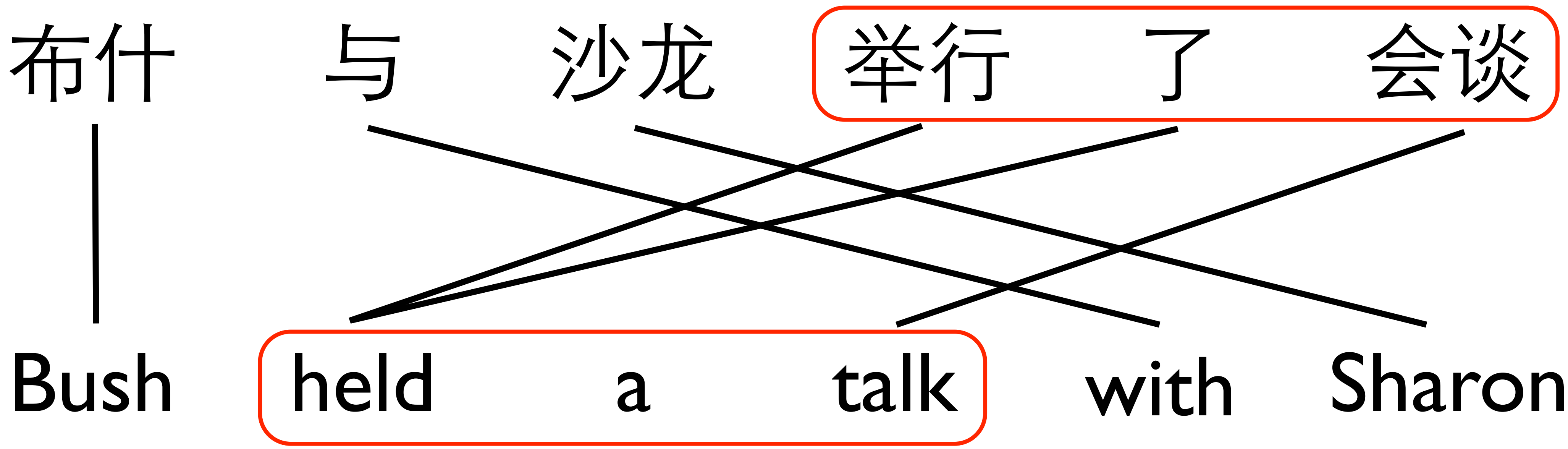
(与 沙龙, with Sharon)

(沙龙, Sharon)

(举行 了, held)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction



(布什, Bush)

(与, with)

(与 沙龙, with Sharon)

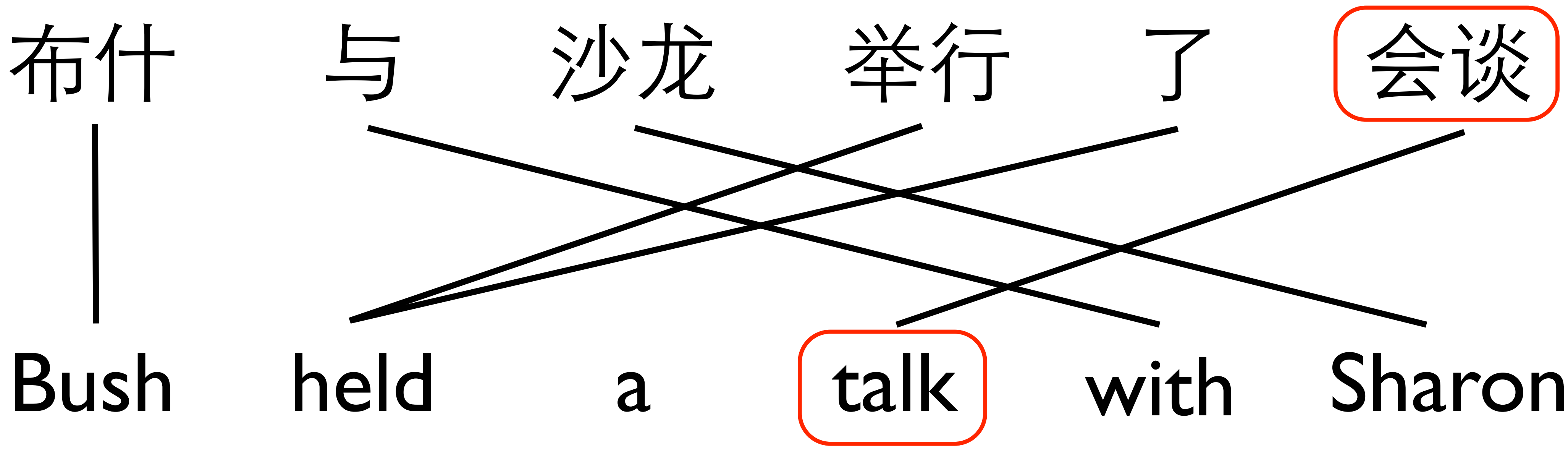
(沙龙, Sharon)

(举行 了, held)

(举行 了 会谈, held a talk)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction



(布什, Bush)

(与, with)

(与 沙龙, with Sharon)

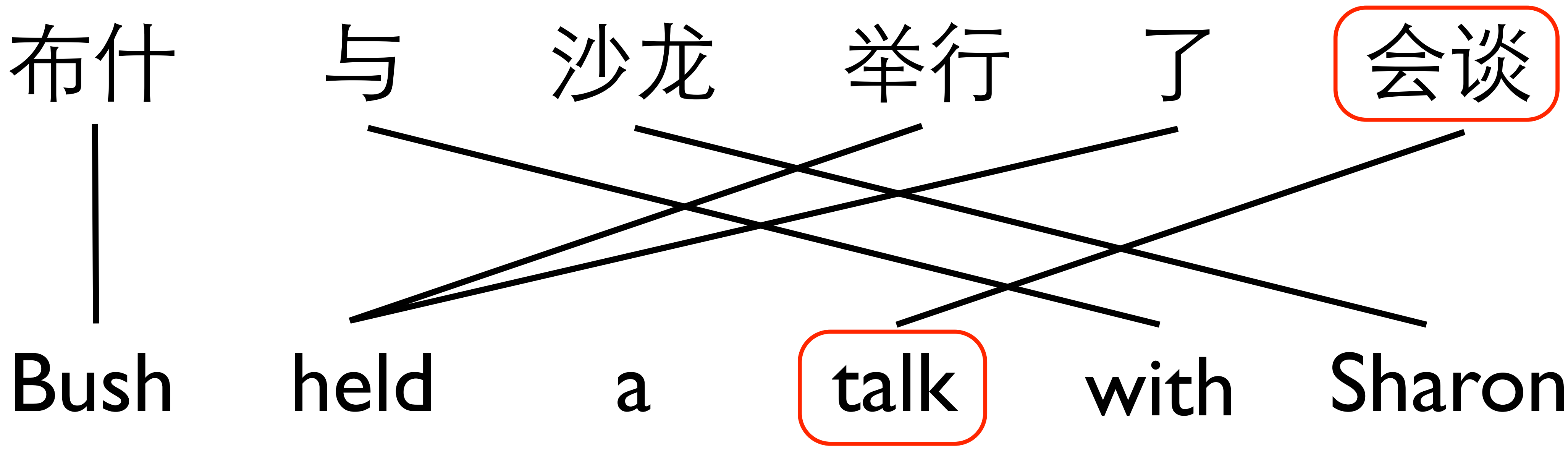
(沙龙, Sharon)

(举行 了, held)

(举行 了 会谈, held a talk)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction



(布什, Bush)

(与, with)

(与 沙龙, with Sharon)

(沙龙, Sharon)

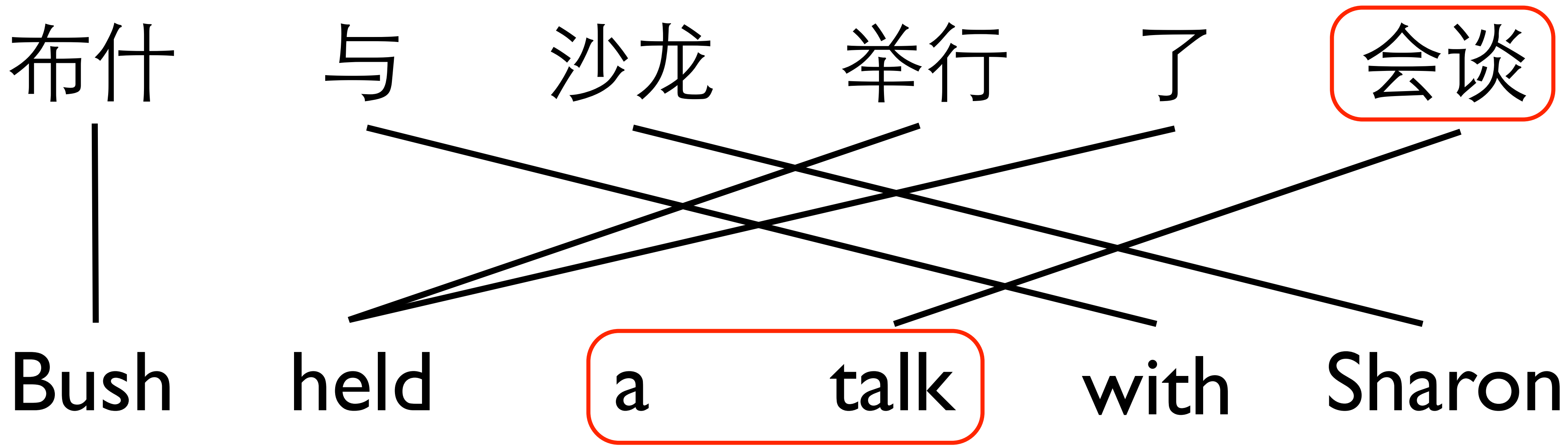
(举行 了, held)

(举行 了 会谈, held a talk)

(会谈, talk)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction



(布什, Bush)

(与, with)

(与 沙龙, with Sharon)

(沙龙, Sharon)

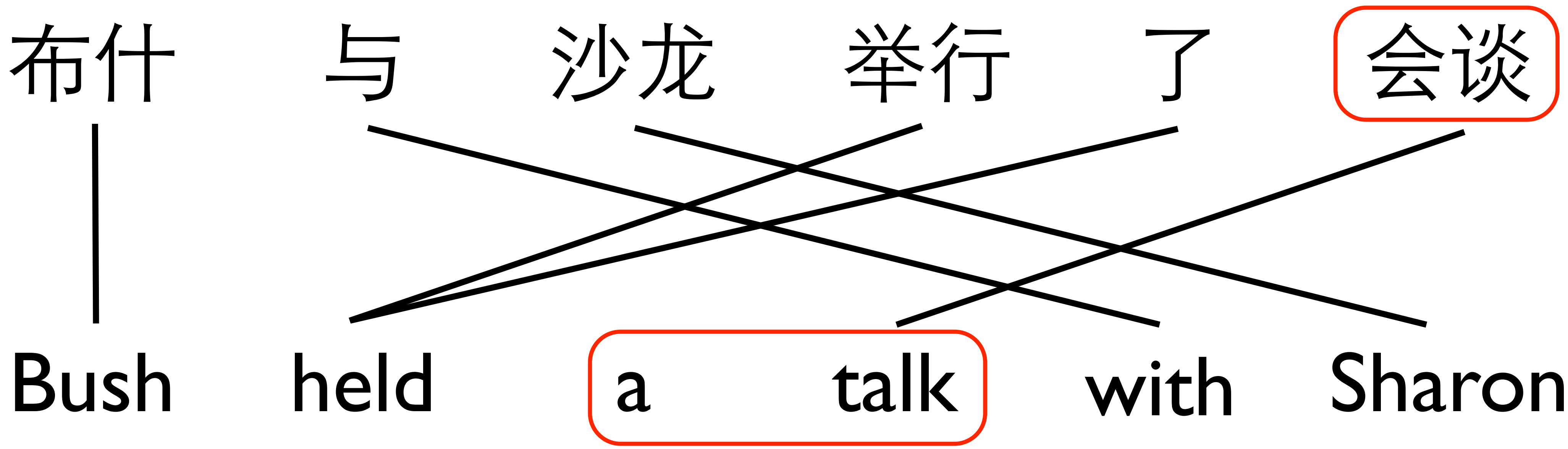
(举行 了, held)

(举行 了 会谈, held a talk)

(会谈, talk)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Extraction



(布什, Bush)

(与, with)

(与 沙龙, with Sharon)

(沙龙, Sharon)

(举行 了, held)

(举行 了 会谈, held a talk)

(会谈, talk)

(会谈, a talk)

(Koehn et al., 2003; Och and Ney, 2004)

Phrase Translation Probabilities

f	e	count	P(e|f)	P(f|e)
布什	Bush	1		
与	with	1		
与 沙龙	with Sharon	1		
沙龙	Sharon	1		
举行 了	held	1		
举行 了 会谈	held a talk	1		
会谈	talk	1		
会谈	a talk	1		

Phrase Translation Probabilities

$$P(e|f) = \frac{count(f,e)}{\sum_{e'} count(f,e')}$$

f	e	count	P(e|f)	P(f|e)
布什	Bush	1		
与	with	1		
与 沙龙	with Sharon	1		
沙龙	Sharon	1		
举行 了	held	1		
举行 了 会谈	held a talk	1		
会谈	talk	1		
会谈	a talk	1		

Phrase Translation Probabilities

$$P(e|f) = \frac{count(f,e)}{\sum_{e'} count(f,e')}$$

f	e	count	P(e|f)	P(f|e)
布什	Bush	1	1.0	
与	with	1	1.0	
与 沙龙	with Sharon	1	1.0	
沙龙	Sharon	1	1.0	
举行 了	held	1	1.0	
举行 了 会谈	held a talk	1	1.0	
会谈	talk	1	0.5	
会谈	a talk	1	0.5	

Phrase Translation Probabilities

$$P(e|f) = \frac{count(f,e)}{\sum_{e'} count(f,e')} \qquad P(f|e) = \frac{count(f,e)}{\sum_{f'} count(f',e)}$$

f	e	count	P(e|f)	P(f|e)
布什	Bush	1	1.0	
与	with	1	1.0	
与 沙龙	with Sharon	1	1.0	
沙龙	Sharon	1	1.0	
举行 了	held	1	1.0	
举行 了 会谈	held a talk	1	1.0	
会谈	talk	1	0.5	
会谈	a talk	1	0.5	

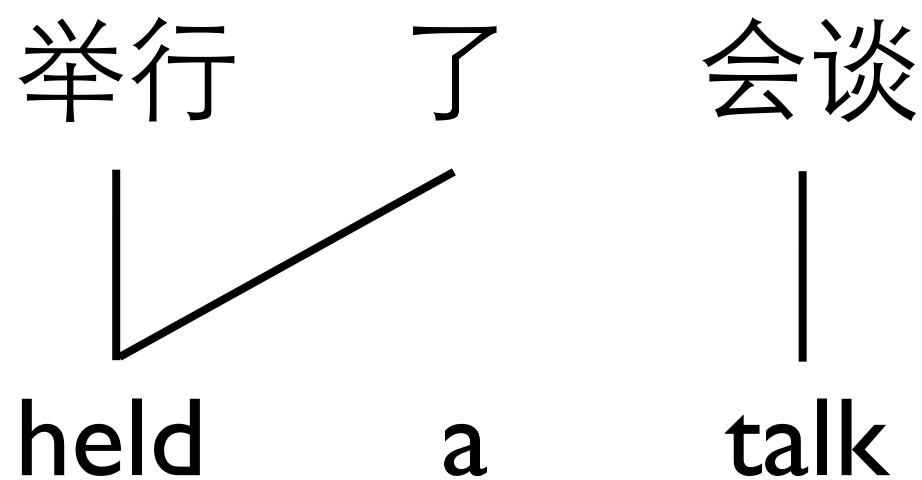
Phrase Translation Probabilities

$$P(e|f) = \frac{count(f,e)}{\sum_{e'} count(f,e')} \qquad P(f|e) = \frac{count(f,e)}{\sum_{f'} count(f',e)}$$

f	e	count	P(e|f)	P(f|e)
布什	Bush	1	1.0	1.0
与	with	1	1.0	1.0
与 沙龙	with Sharon	1	1.0	1.0
沙龙	Sharon	1	1.0	1.0
举行 了	held	1	1.0	1.0
举行 了 会谈	held a talk	1	1.0	1.0
会谈	talk	1	0.5	1.0
会谈	a talk	1	0.5	1.0

Lexical Weighting

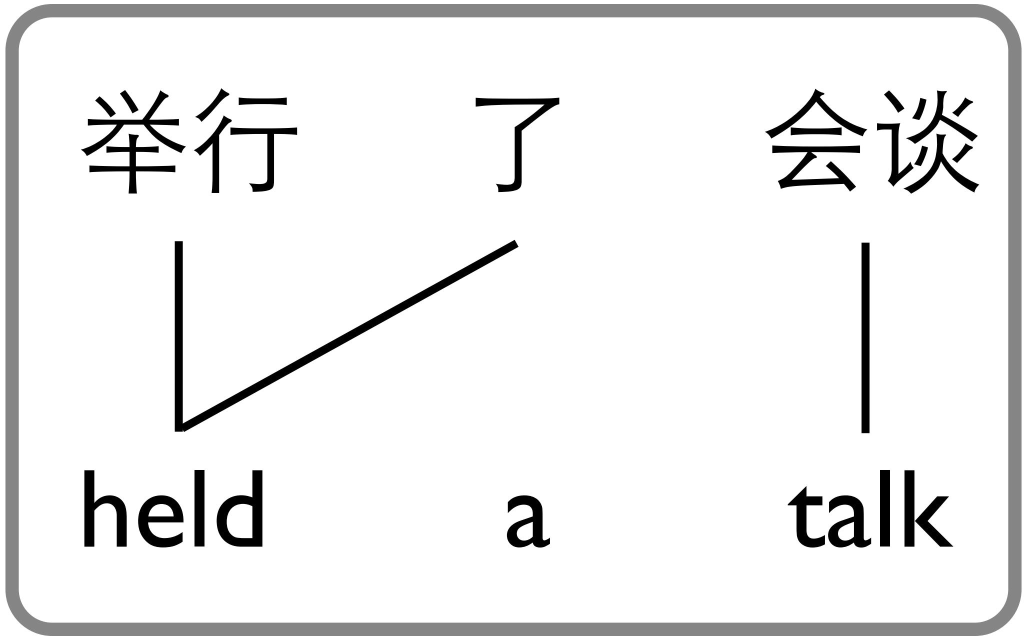
- estimating phrase translation probabilities using relative frequencies suffers from sparse data
- lexical weighting considers word alignment with phrase pairs



(Koehn et al., 2003)

Lexical Weighting

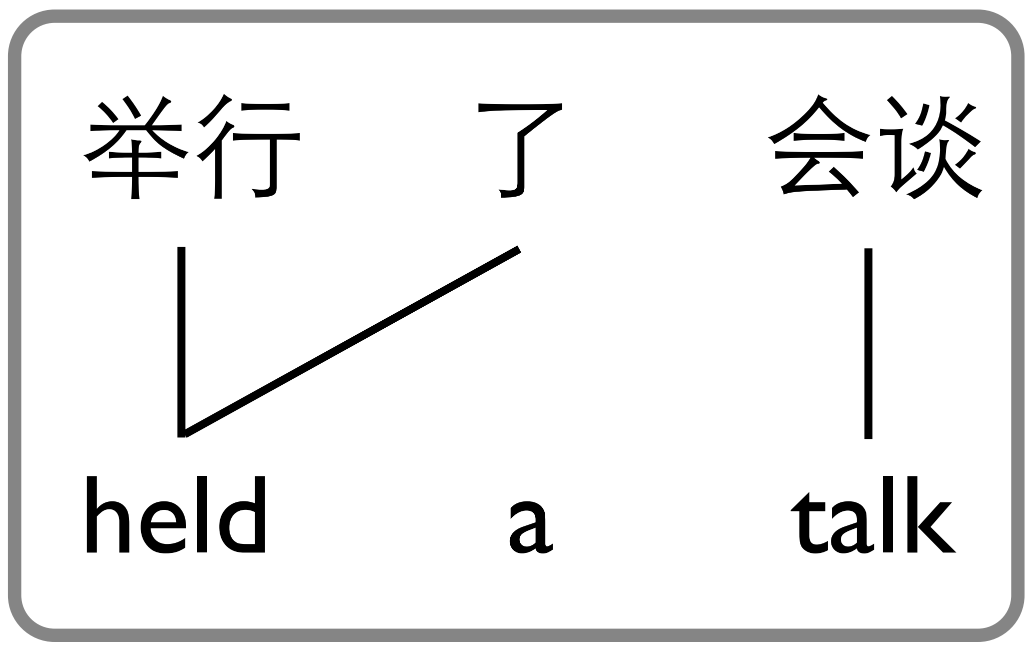
- estimating phrase translation probabilities using relative frequencies suffers from sparse data
- lexical weighting considers word alignment with phrase pairs



(Koehn et al., 2003)

Lexical Weighting

- estimating phrase translation probabilities using relative frequencies suffers from sparse data
- lexical weighting considers word alignment with phrase pairs



(w(“held”|“举行”)+w(“held”|“了”)) / 2 *
w(“a”|NULL) *
w(“talk”|“会谈”)

(Koehn et al., 2003)

Reordering is Hard

at

Prime Bush Minister

a the

US meeting

held President House

yesterday Sharon

with White Israeli

Reordering is Hard

at

Prime

Bush

Minister

a

the

US

meeting

President

held

House

yesterday

Sharon

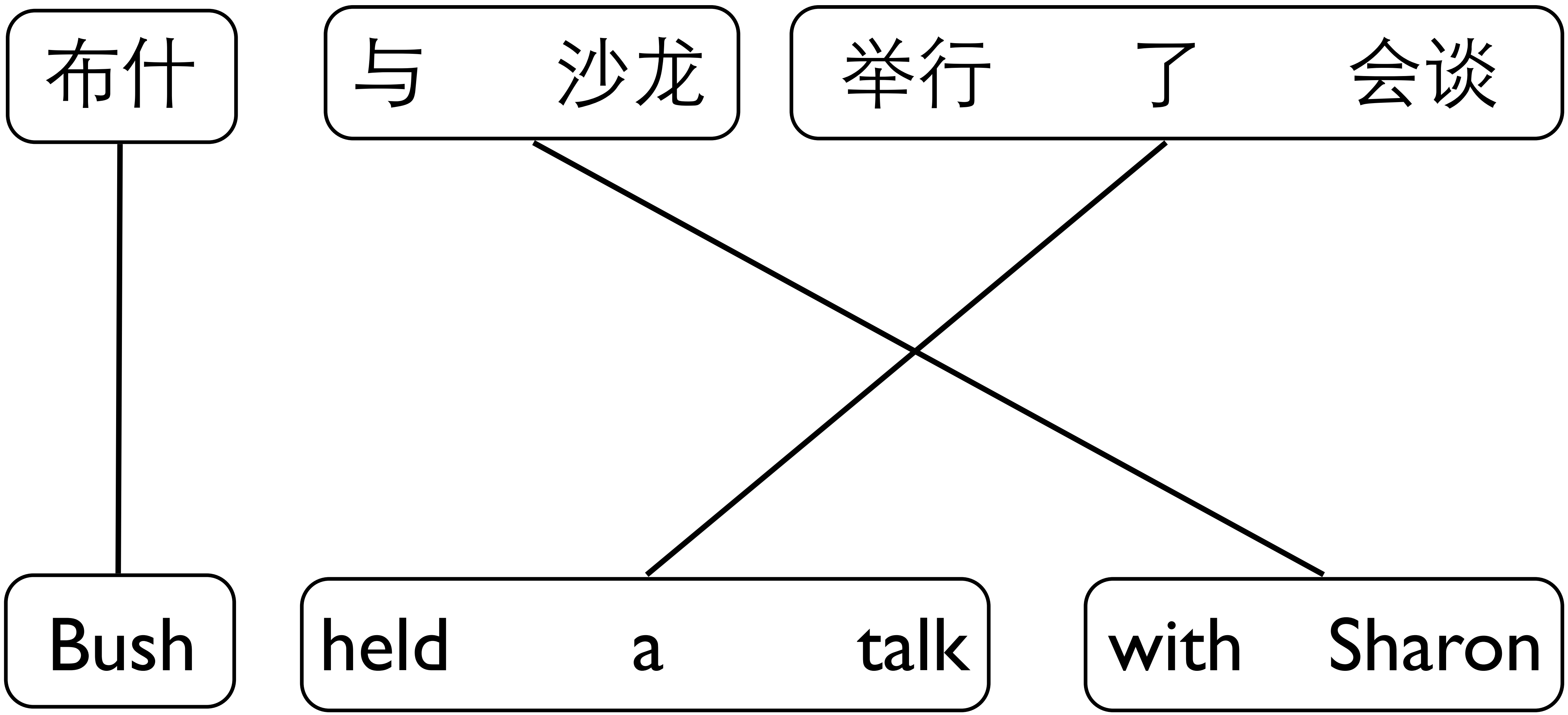
with

White

Israeli

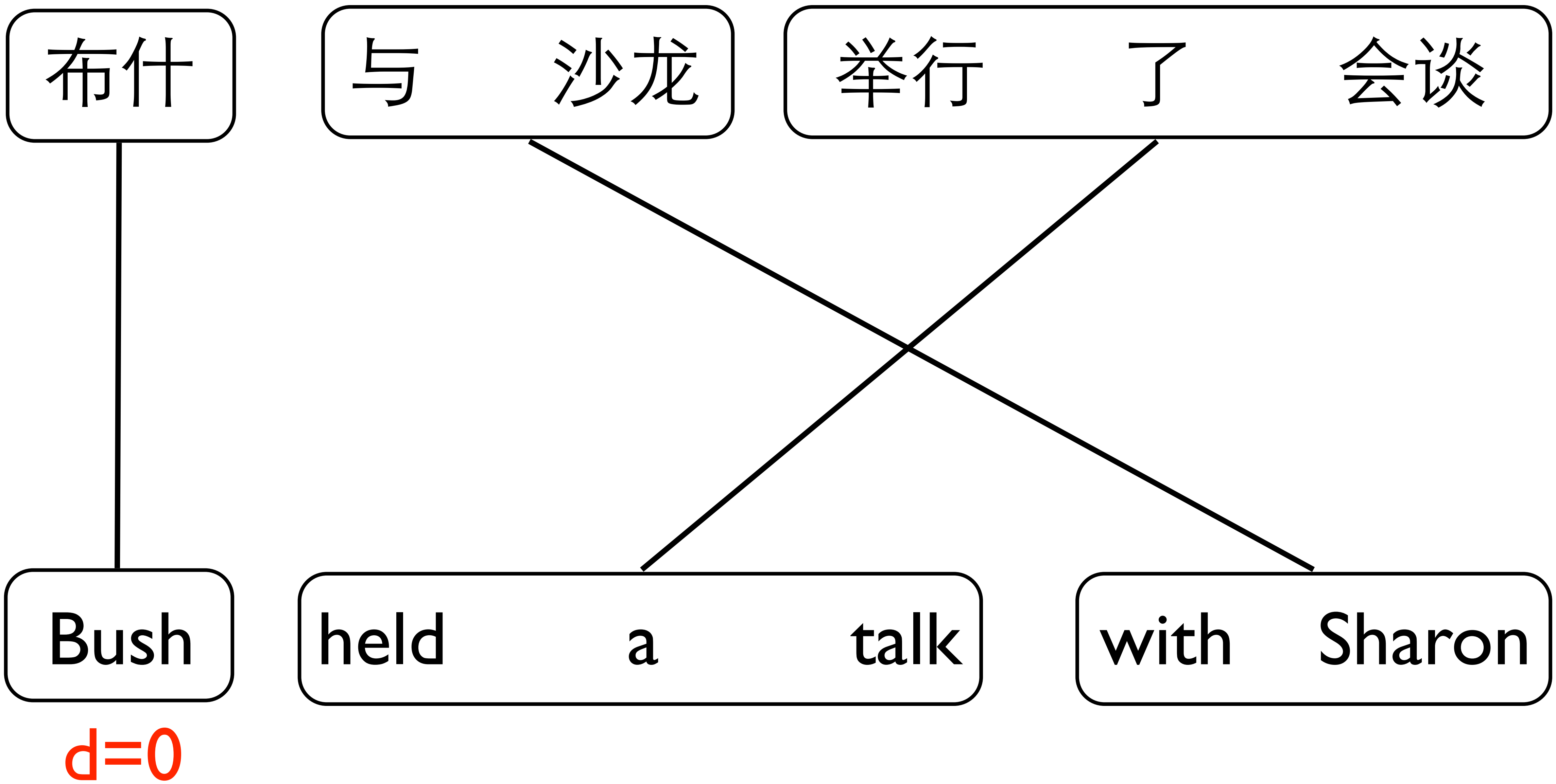
Q: can you figure out a sentence using these words?

Distance-based Reordering Model



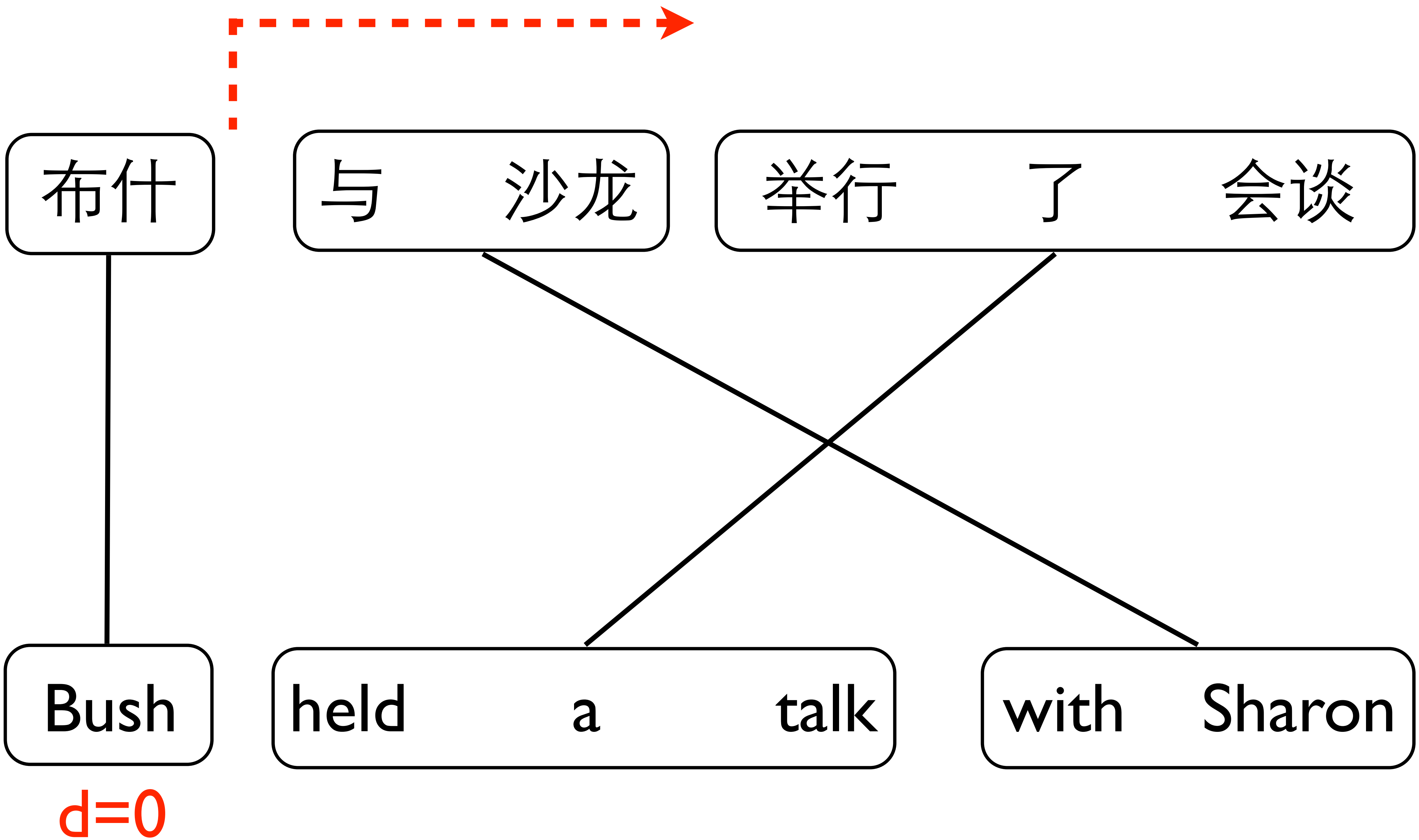
(Koehn et al., 2003; Och and Ney, 2004)

Distance-based Reordering Model



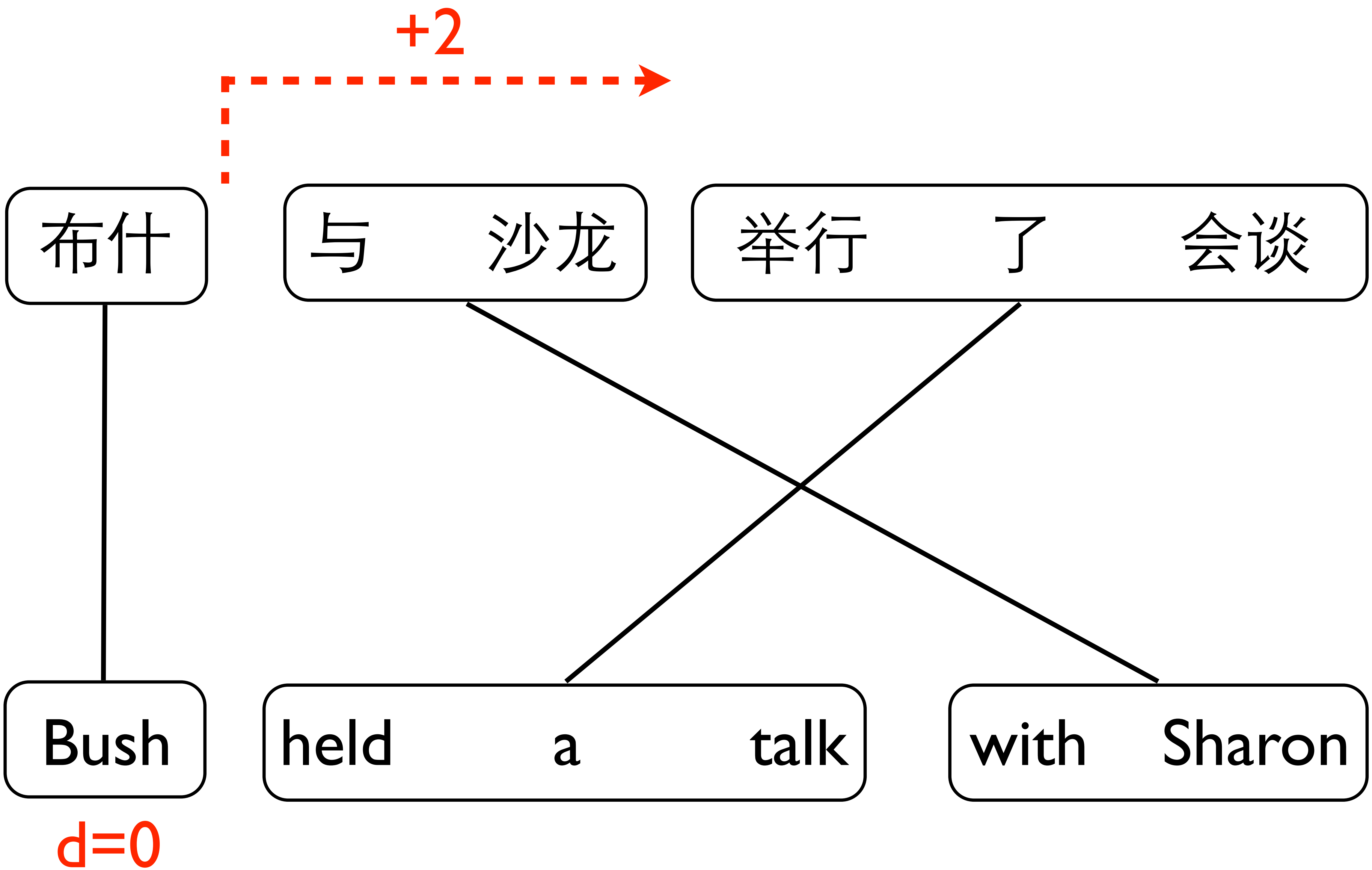
(Koehn et al., 2003; Och and Ney, 2004)

Distance-based Reordering Model



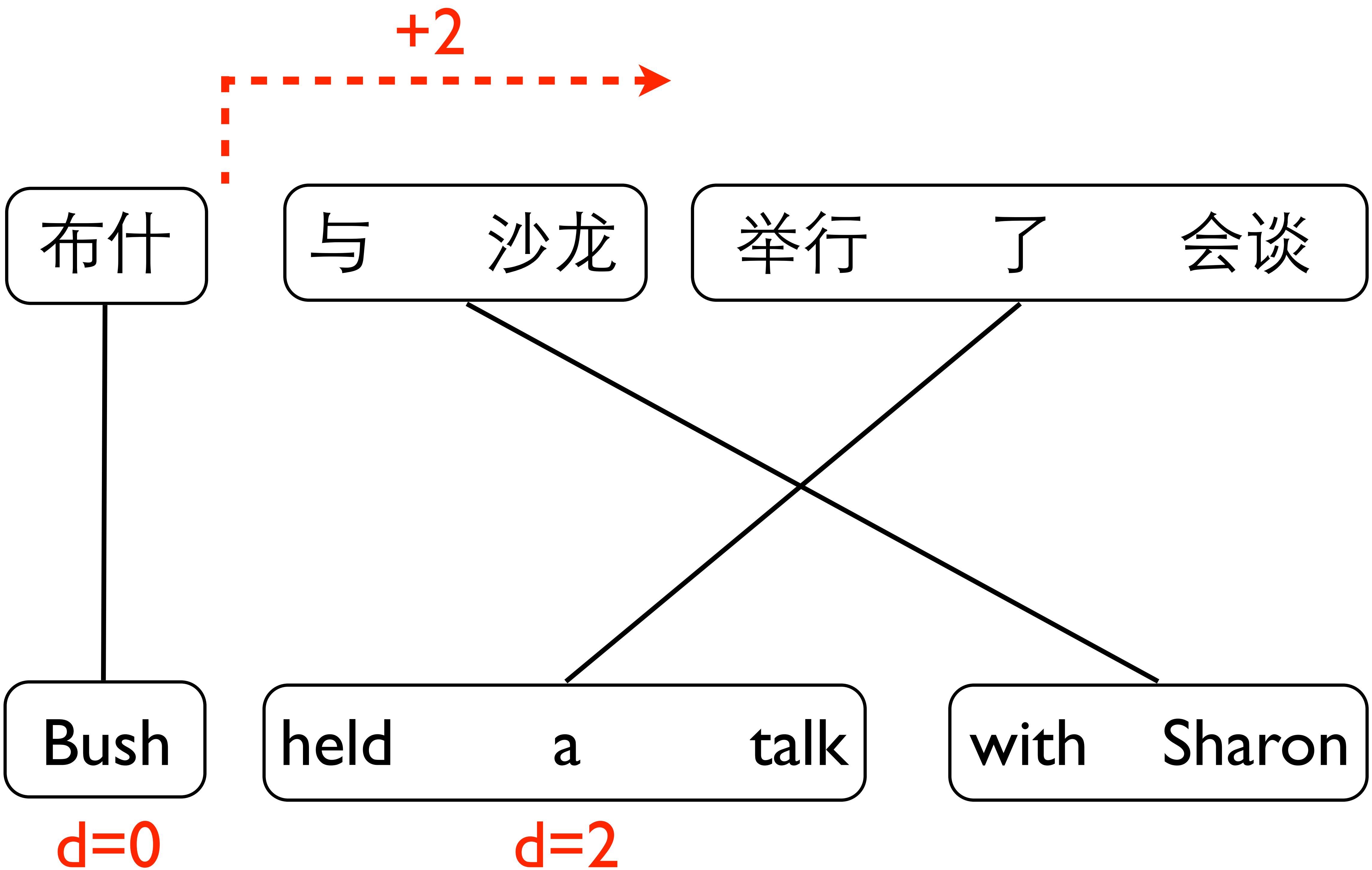
(Koehn et al., 2003; Och and Ney, 2004)

Distance-based Reordering Model



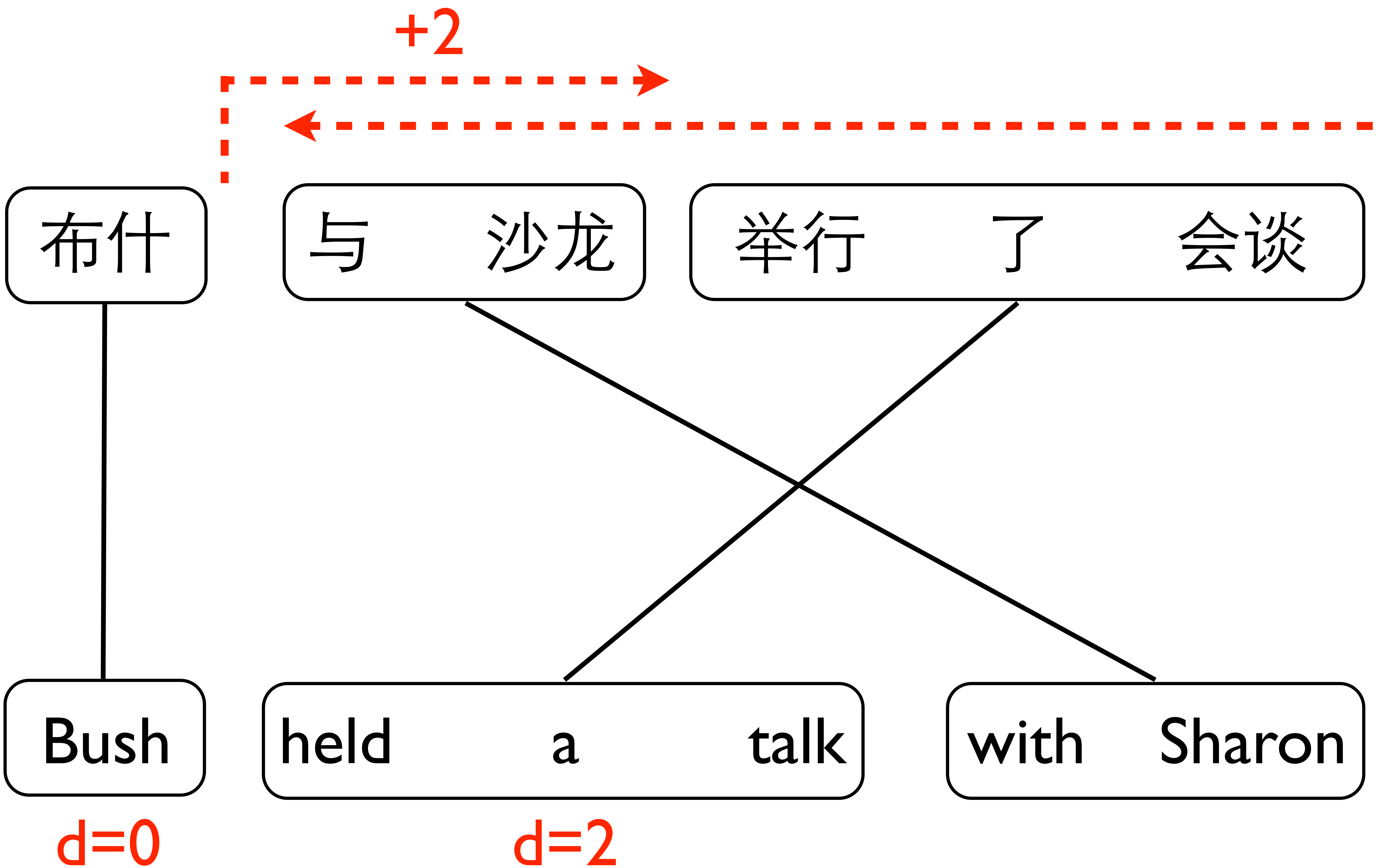
(Koehn et al., 2003; Och and Ney, 2004)

Distance-based Reordering Model



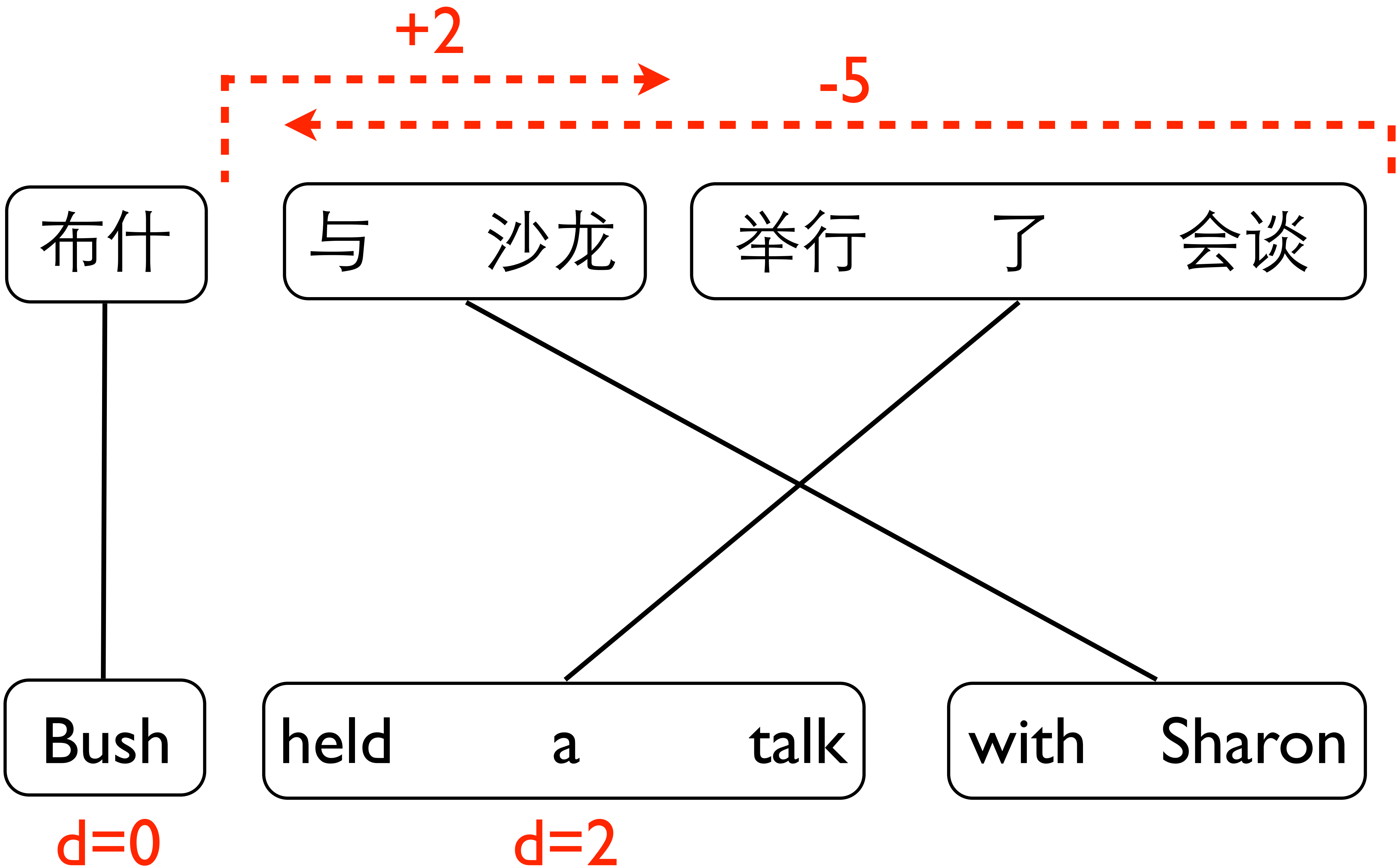
(Koehn et al., 2003; Och and Ney, 2004)

Distance-based Reordering Model



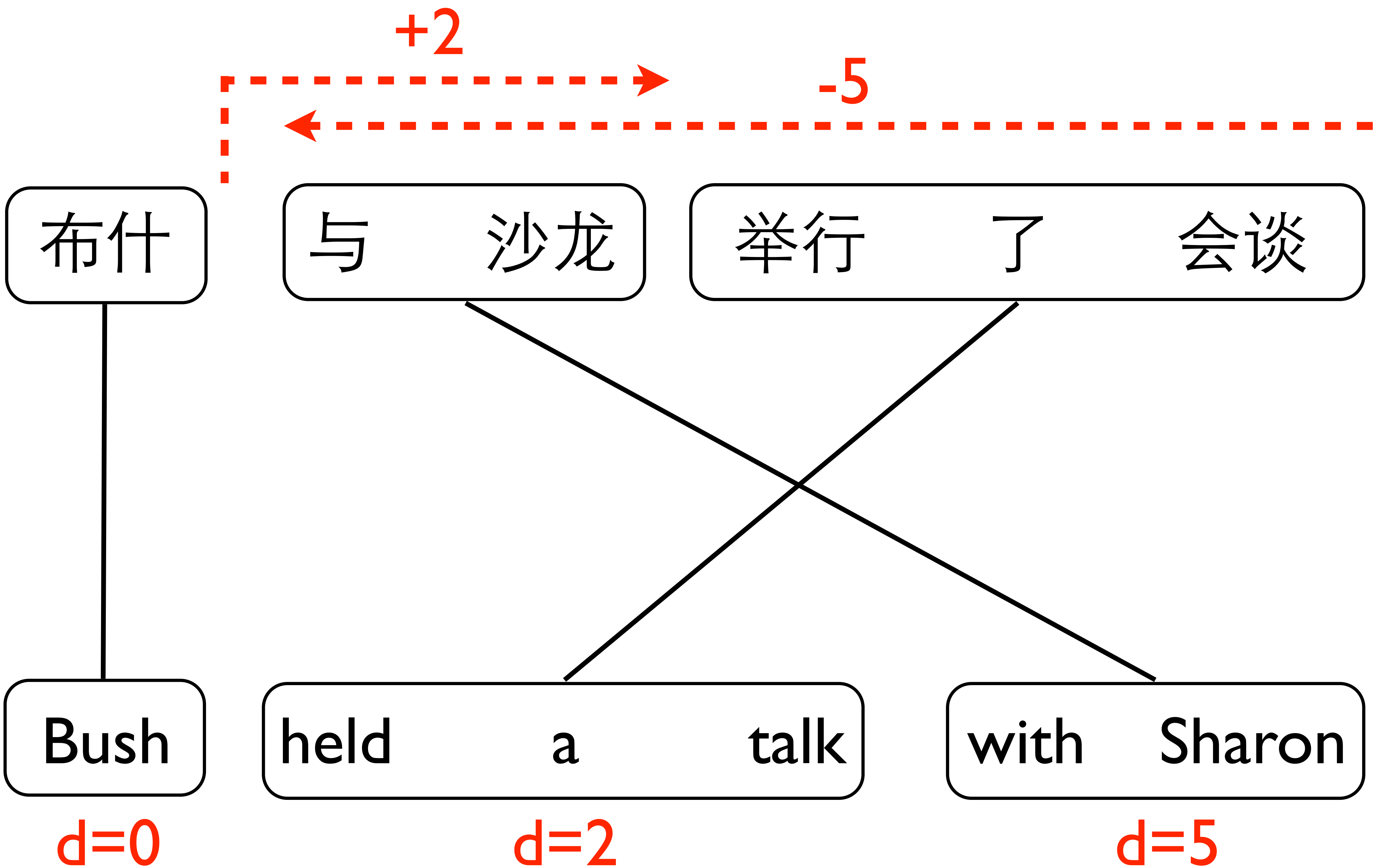
(Koehn et al., 2003; Och and Ney, 2004)

Distance-based Reordering Model



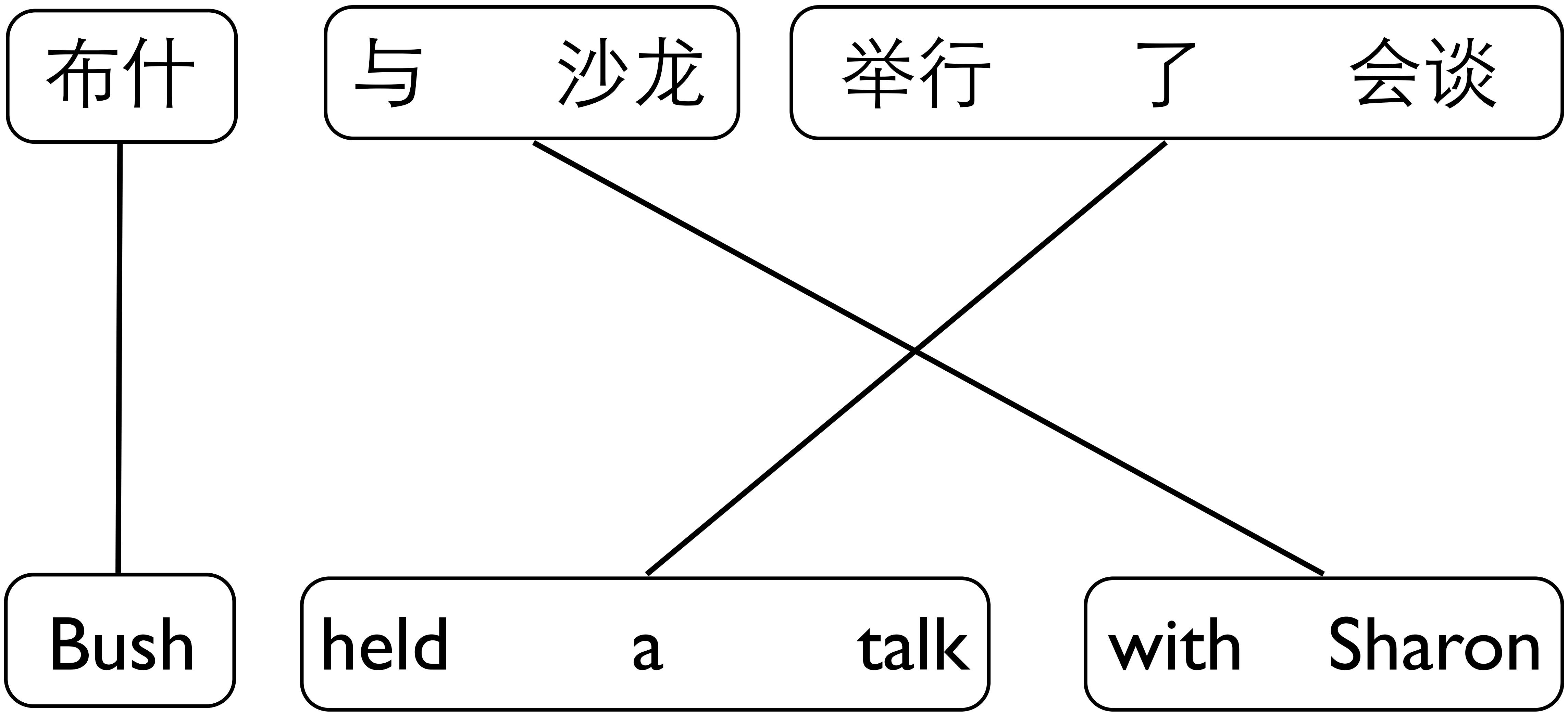
(Koehn et al., 2003; Och and Ney, 2004)

Distance-based Reordering Model



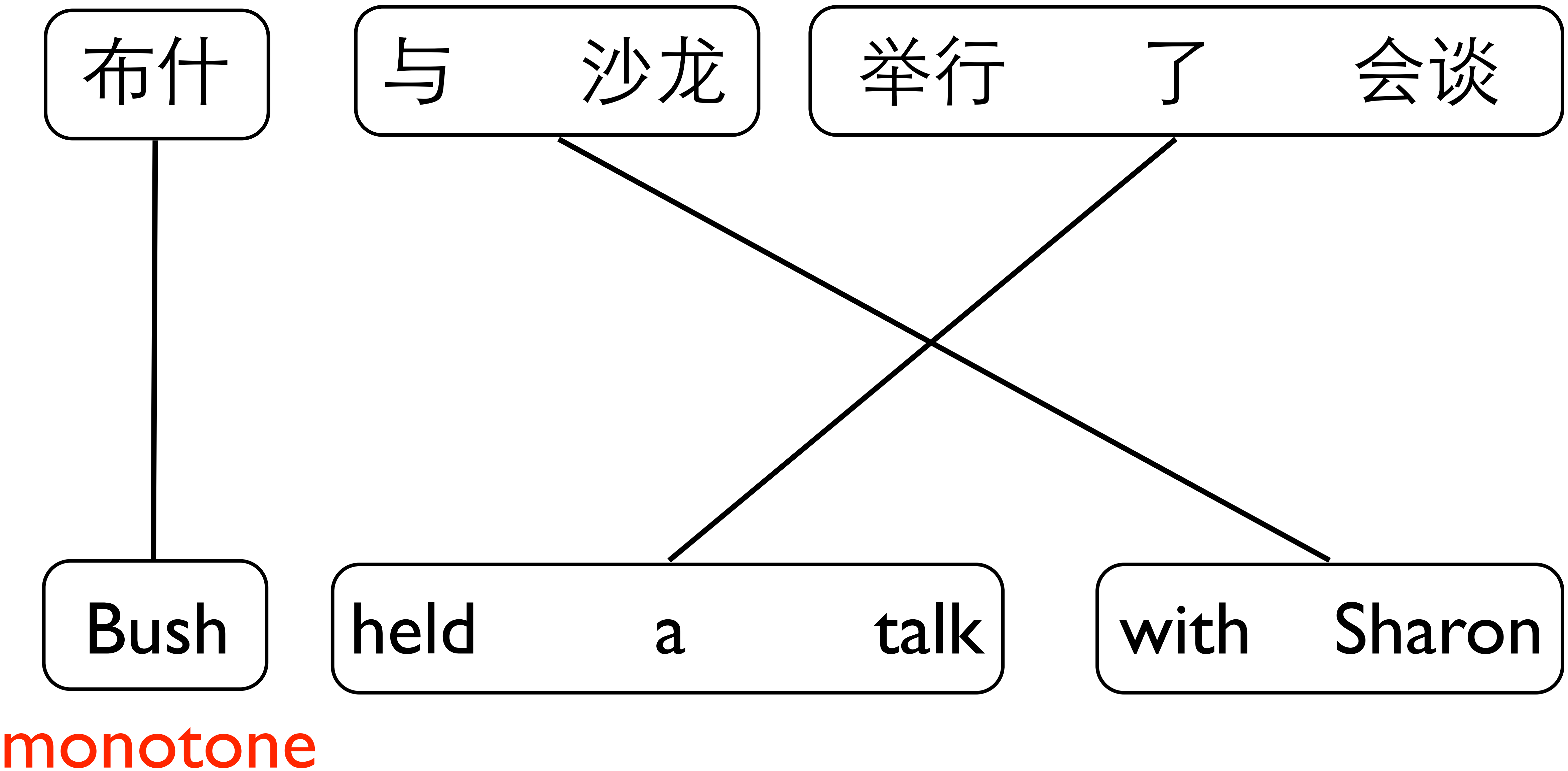
(Koehn et al., 2003; Och and Ney, 2004)

Lexicalized Reordering Model



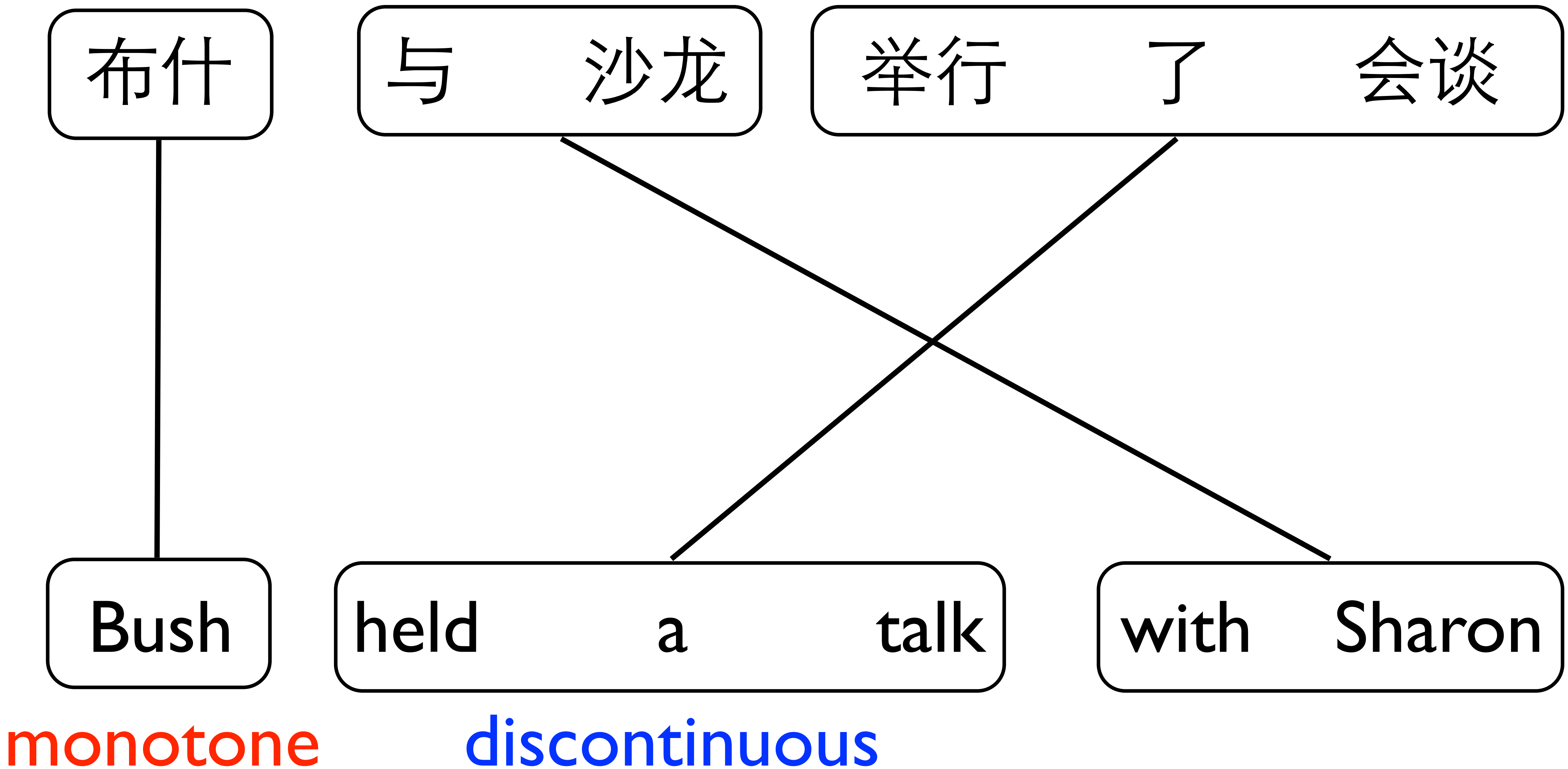
(Koehn et al., 2007)

Lexicalized Reordering Model



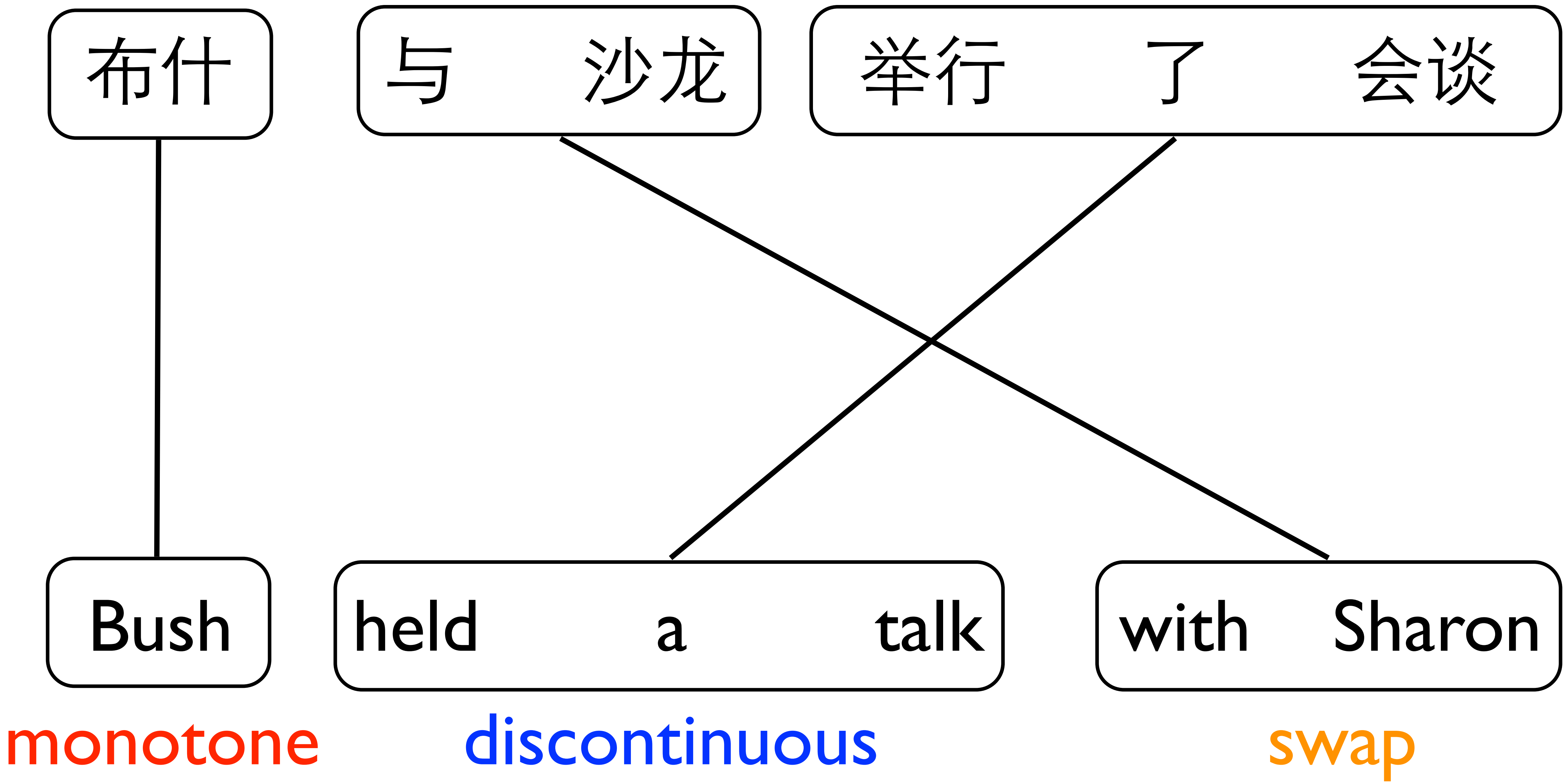
(Koehn et al., 2007)

Lexicalized Reordering Model



(Koehn et al., 2007)

Lexicalized Reordering Model



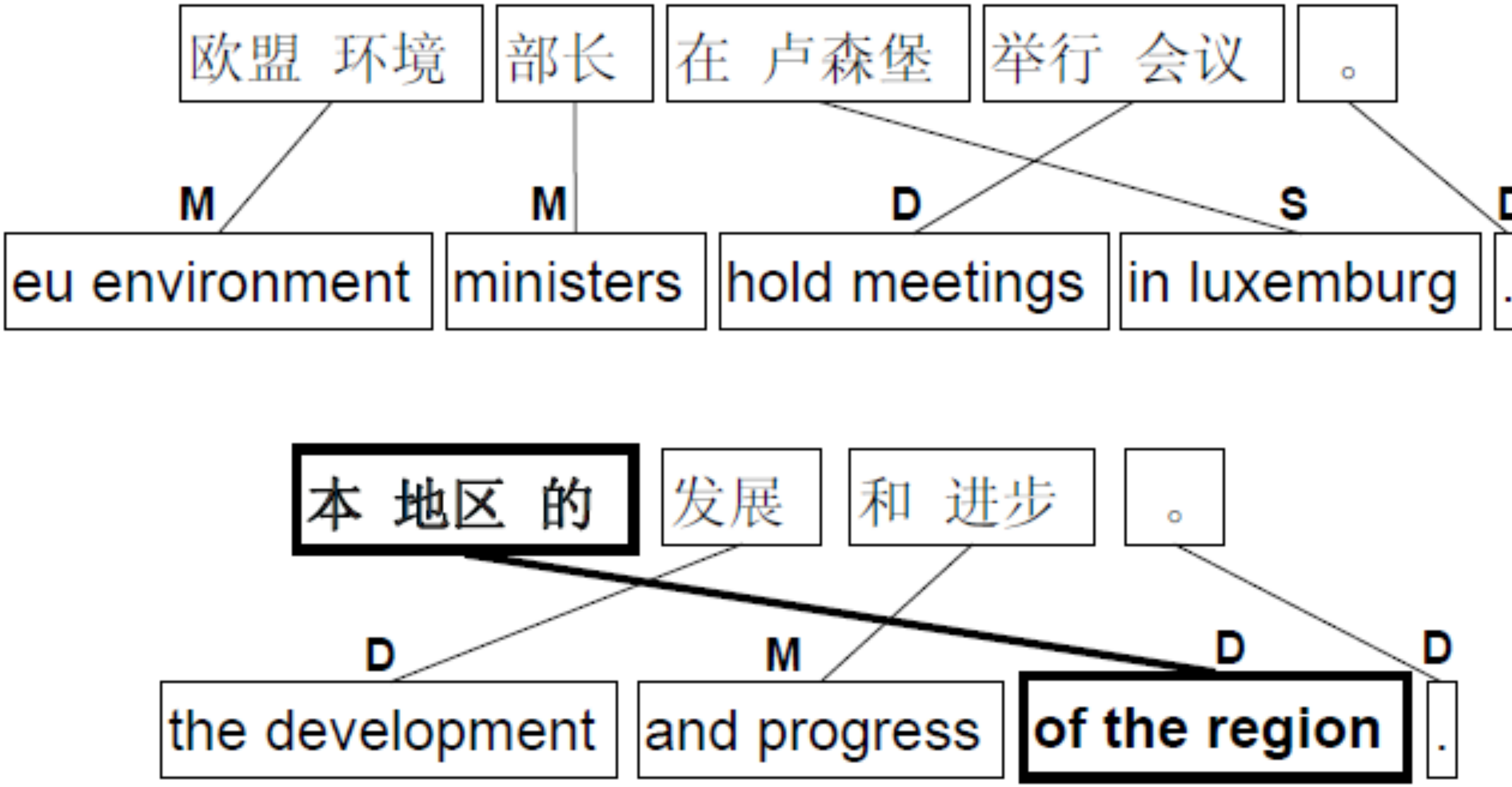
(Koehn et al., 2007)

Lexicalized Reordering Model

f	e	M	S	D
布什	Bush	0.4	0.3	0.3
与	with	0.6	0.1	0.3
与 沙龙	with Sharon	0.3	0.5	0.2
沙龙	Sharon	0.4	0.3	0.3
举行 了	held	0.8	0.1	0.1
举行 了 会谈	held a talk	0.4	0.4	0.2
会谈	talk	0.3	0.3	0.4
会谈	a talk	0.3	0.4	0.3

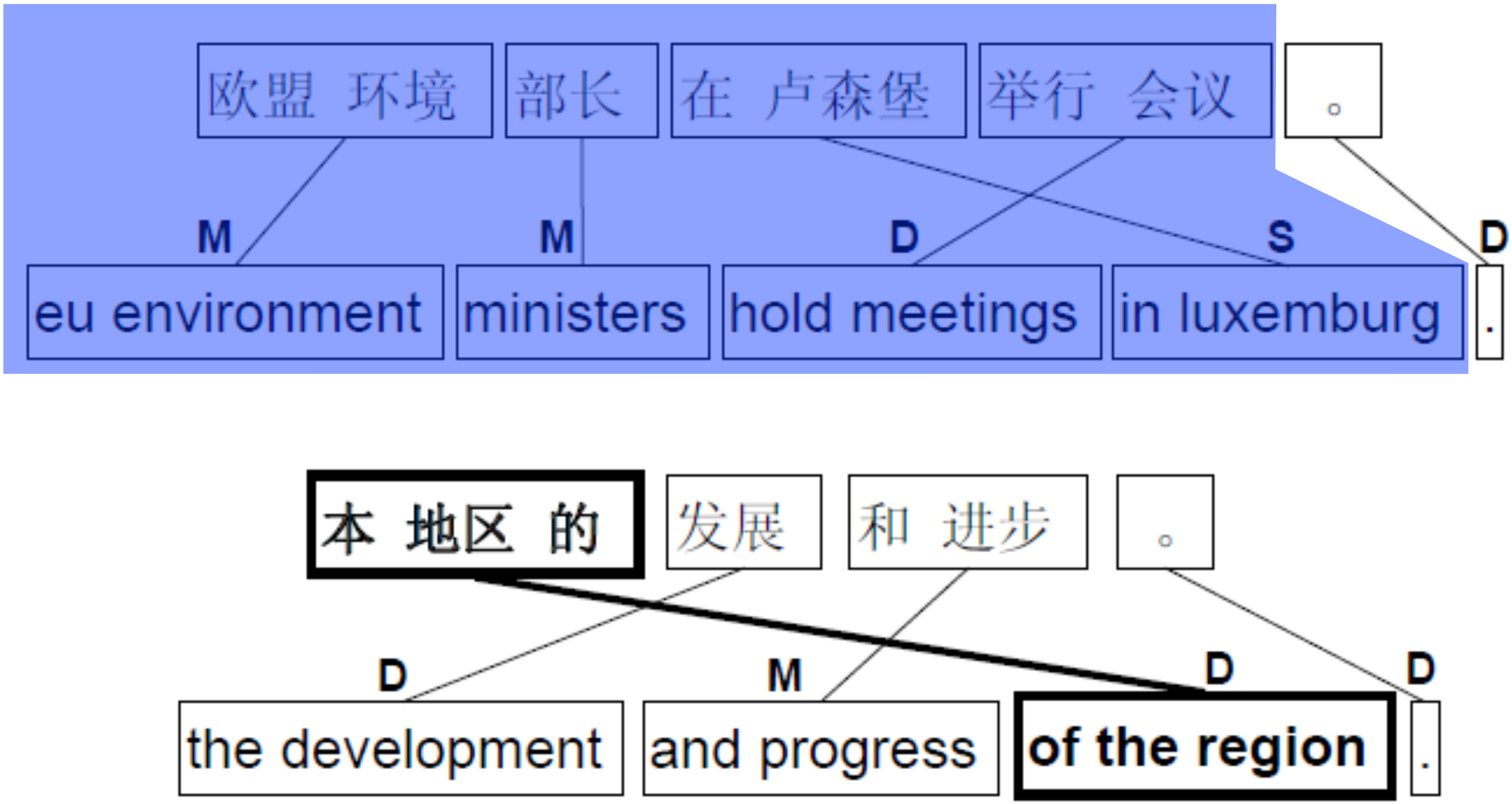
(Koehn et al., 2007)

Hierarchical Reordering Model



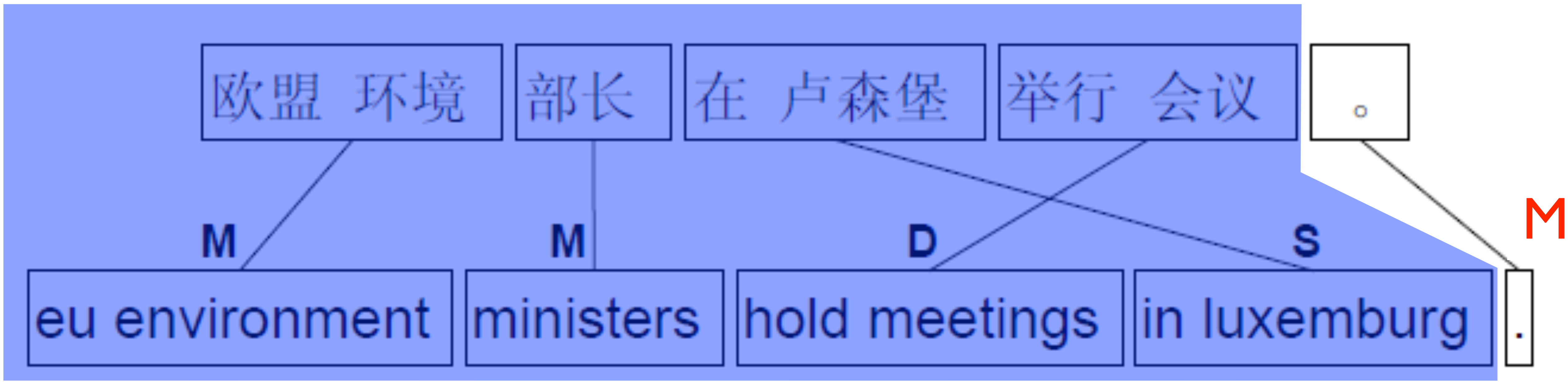
(Galley and Manning, 2008)

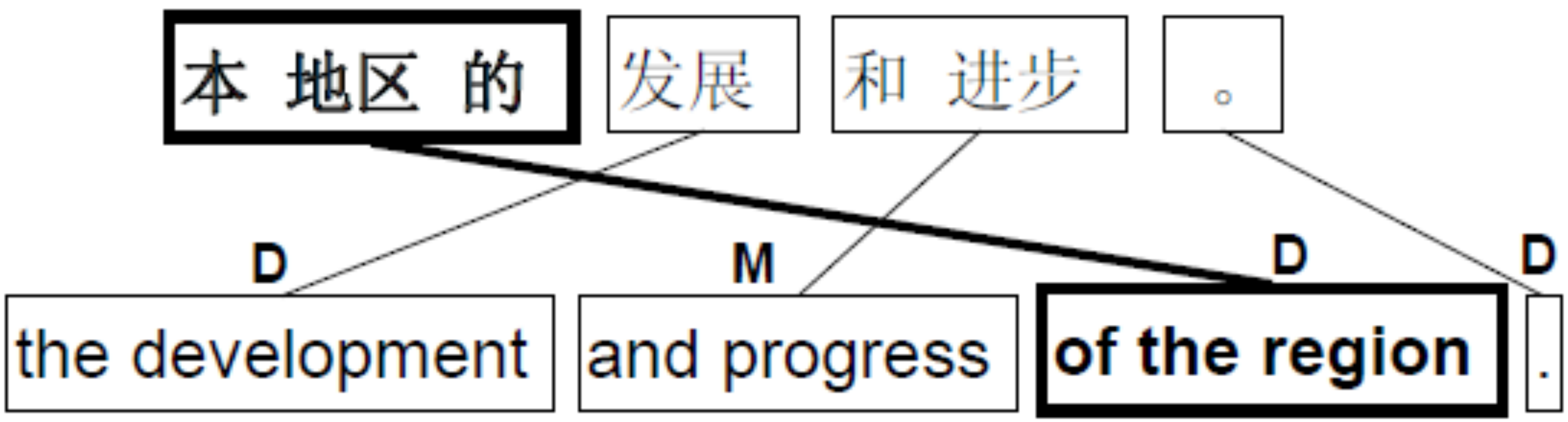
Hierarchical Reordering Model



(Galley and Manning, 2008)

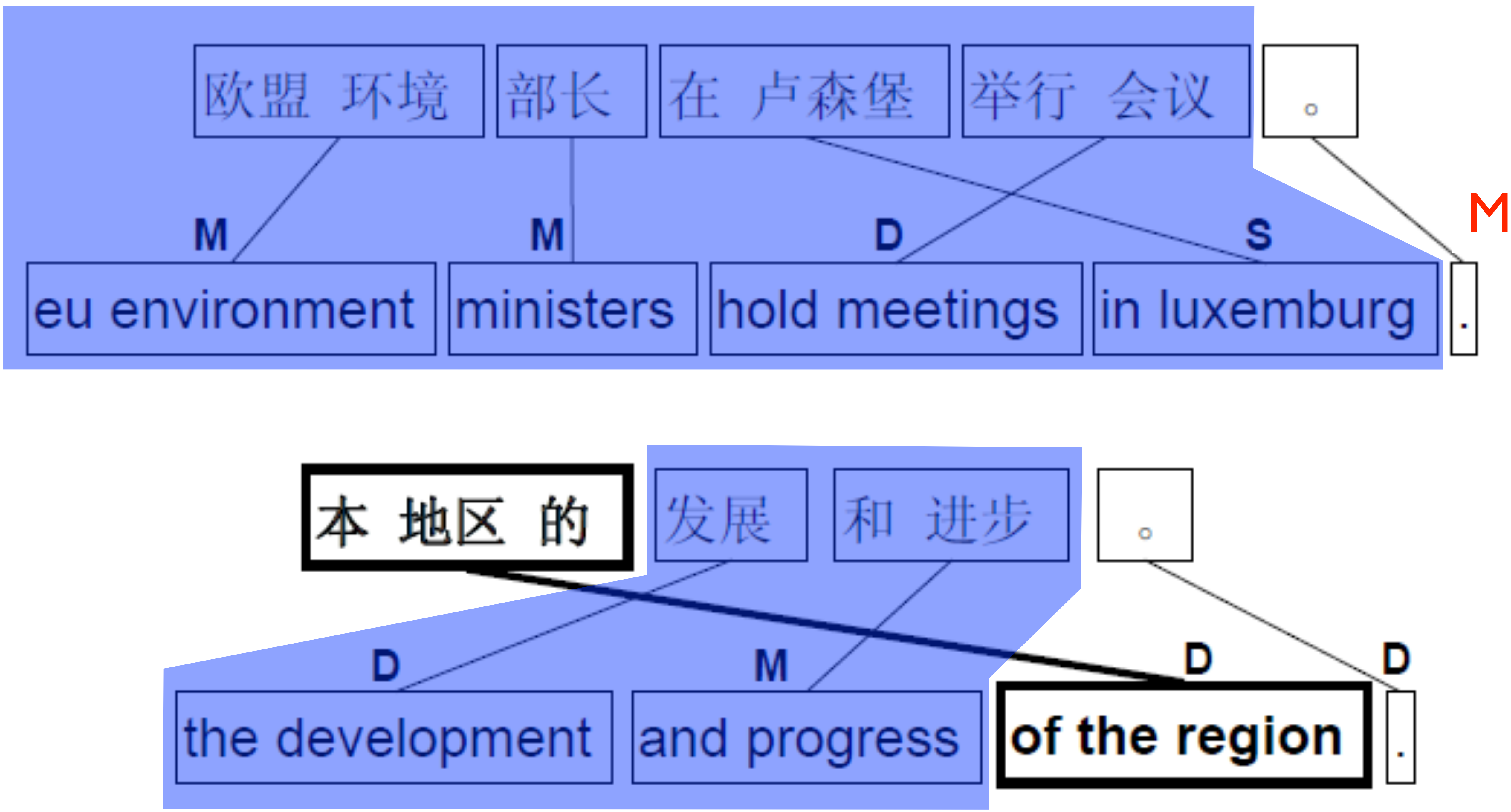
Hierarchical Reordering Model





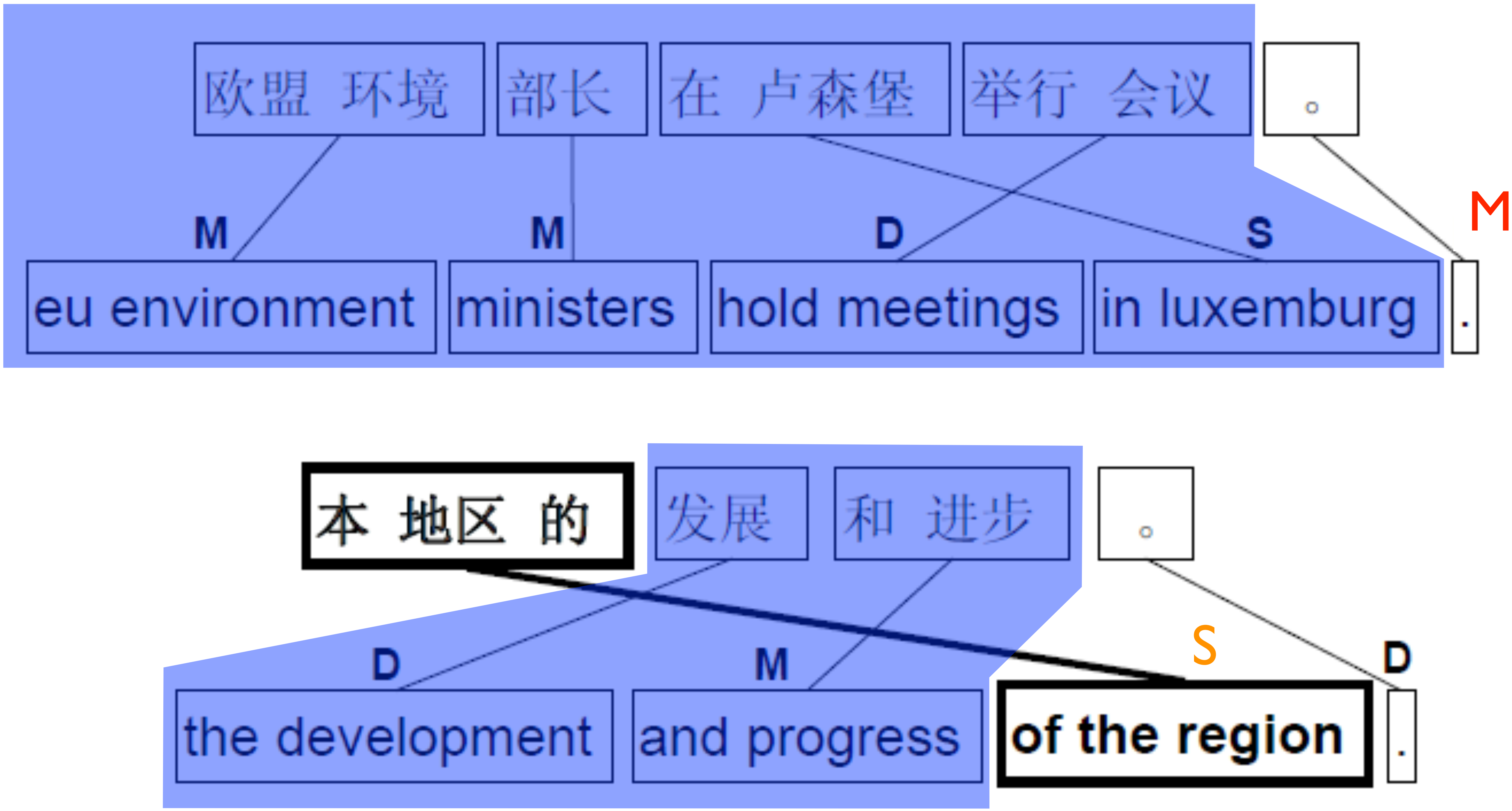
(Galley and Manning, 2008)

Hierarchical Reordering Model



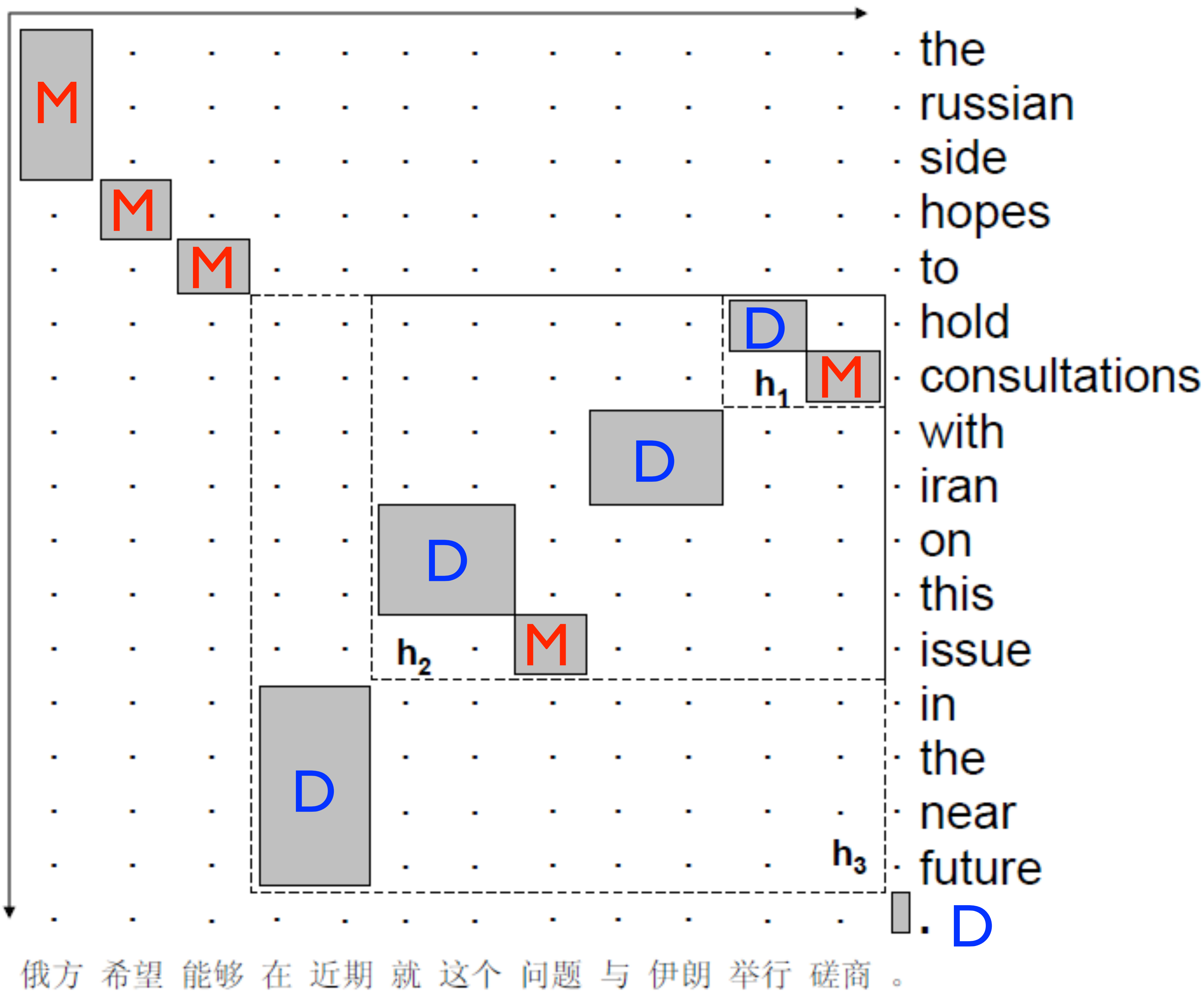
(Galley and Manning, 2008)

Hierarchical Reordering Model



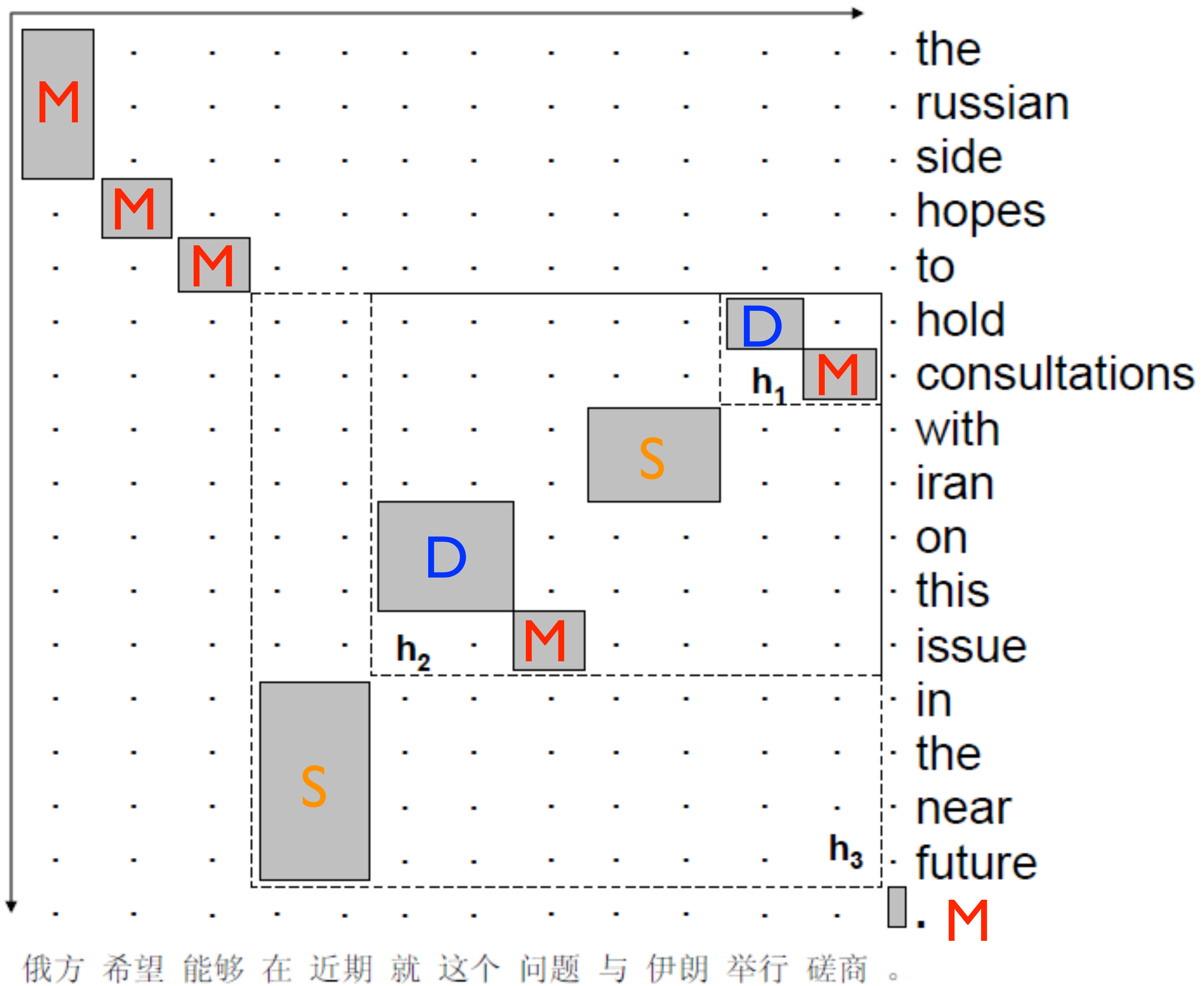
(Galley and Manning, 2008)

Lexicalized Reordering Model



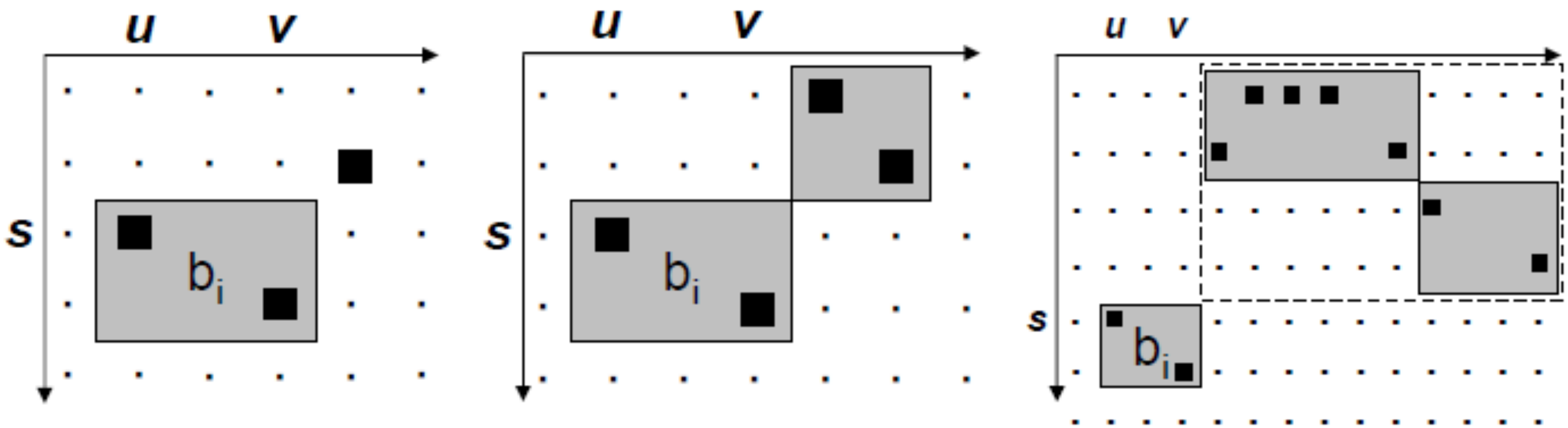
(Galley and Manning, 2008)

Hierarchical Reordering Model



(Galley and Manning, 2008)

Word, Phrase and Hierarchical



word-based phrase-based hierarchical

Orientation Model	$o_i = M$	$o_i = S$	$o_i = D$
word-based (Moses)	0.1750	0.0159	0.8092
phrase-based	0.3192	0.0704	0.6104
hierarchical	0.4878	0.1004	0.4116

(Galley and Manning, 2008)

Pre-Reordering Model

布什　与　沙龙　举行　了　会谈

(Li et al., 2007)

Pre-Reordering Model

布什 与 沙龙 举行 了 会谈

$\Downarrow$

布什 举行 了 会谈 与 沙龙

(Li et al., 2007)

Pre-Reordering Model

布什 与 沙龙 举行 了 会谈

$\Downarrow$

布什 举行 了 会谈 与 沙龙

$\Downarrow$

Bush held a talk with Sharon

(Li et al., 2007)

Generative Model Revisited

$$P(\mathbf{e}|\mathbf{f}) = \frac{P(\mathbf{e}) \times P(\mathbf{f}|\mathbf{e})}{P(\mathbf{f})}$$

Generative Model Revisited

$$P(\mathbf{e}|\mathbf{f}) = \frac{P(\mathbf{e}) \times P(\mathbf{f}|\mathbf{e})}{P(\mathbf{f})}$$

Is it possible to directly model $P(\mathrm{e}|\mathbf{f})$?

Interview

Interview

Interview

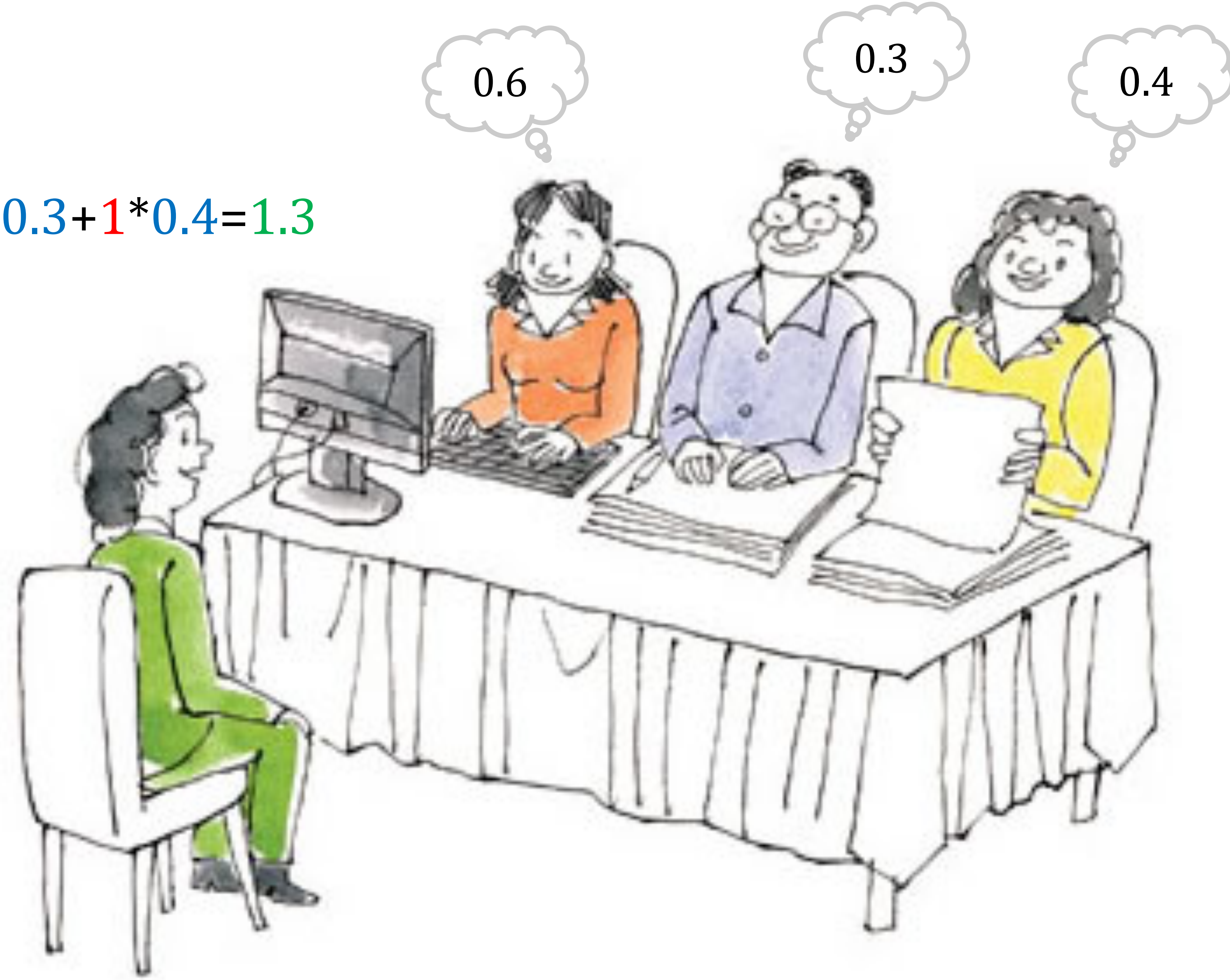
Interview

Interview

0.6+0.3+0.4=1.3

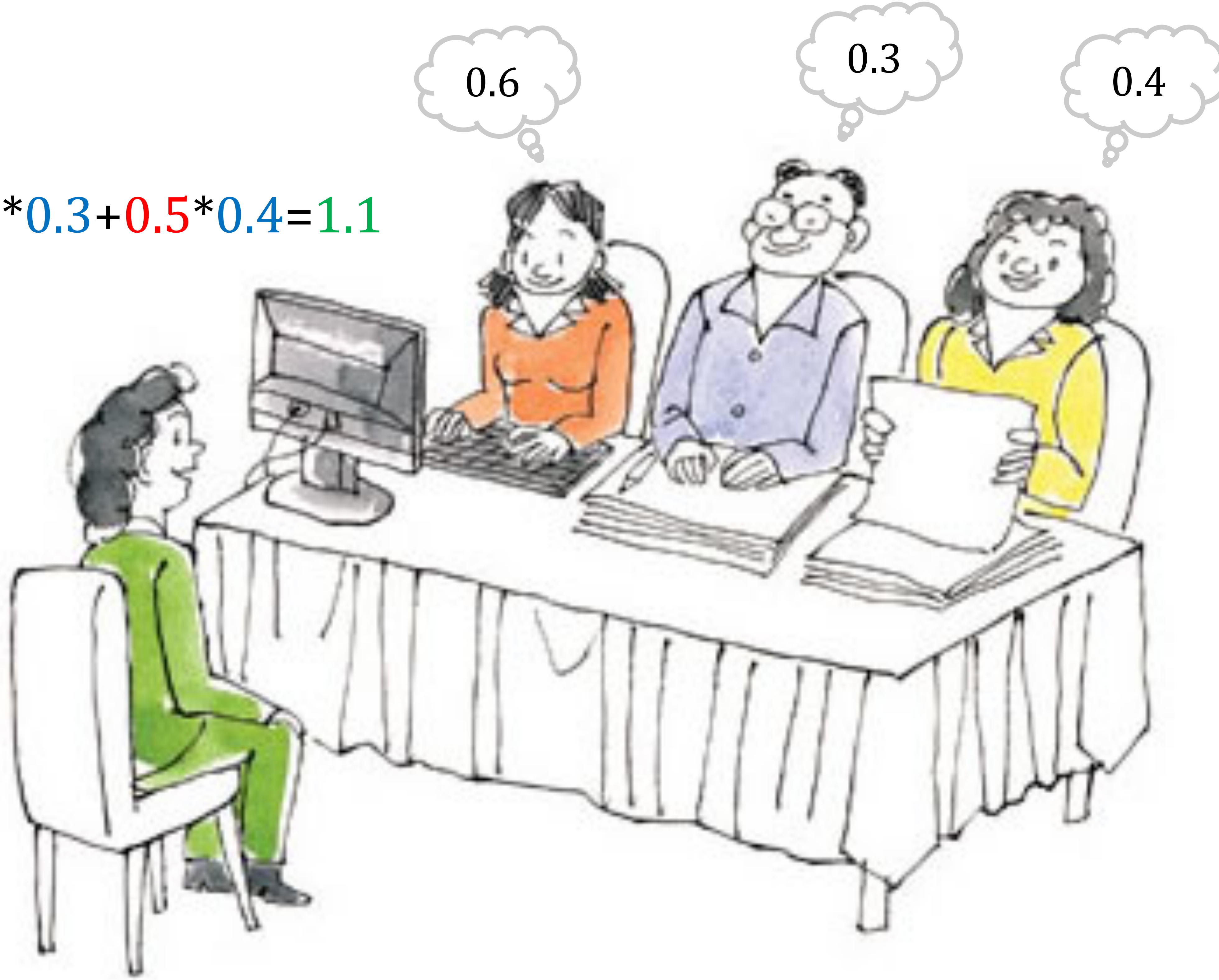
Interview

$1*0.6+1*0.3+1*0.4=1.3$



Interview

$0.5*0.6+2*0.3+0.5*0.4=1.1$



Generative Vs. Discriminative

(Och and Ney, 2002; Och, 2003)

Generative Vs. Discriminative

$$P(\mathbf{e}|\mathbf{f}) = \frac{P(\mathbf{e}) \times P(\mathbf{f}|\mathbf{e})}{P(\mathbf{f})}$$

(Och and Ney, 2002; Och, 2003)

Generative Vs. Discriminative

generative
model

$$P(\mathbf{e}|\mathbf{f}) = \frac{P(\mathbf{e}) \times P(\mathbf{f}|\mathbf{e})}{P(\mathbf{f})}$$

(Och and Ney, 2002; Och, 2003)

Generative Vs. Discriminative

generative model

$$P(\mathbf{e}|\mathbf{f}) = \frac{P(\mathbf{e}) \times P(\mathbf{f}|\mathbf{e})}{P(\mathbf{f})}$$

$$P(\mathbf{e}|\mathbf{f}) = \frac{\exp\left(\sum_{k=1}^{K} \theta_k h_k(\mathbf{f}, \mathbf{e}) \right)}{\sum_{\mathbf{e}'} \exp\left(\sum_{k=1}^{K} \theta_k h_k(\mathbf{f}, \mathbf{e}') \right)}$$

(Och and Ney, 2002; Och, 2003)

Generative Vs. Discriminative

generative model

$$P(\mathbf{e}|\mathbf{f}) = \frac{P(\mathbf{e}) \times P(\mathbf{f}|\mathbf{e})}{P(\mathbf{f})}$$

discriminative model

$$P(\mathbf{e}|\mathbf{f}) = \frac{\exp\Big(\sum_{k=1}^{K} \theta_k h_k(\mathbf{f}, \mathbf{e})\Big)}{\sum_{\mathbf{e}'} \exp\Big(\sum_{k=1}^{K} \theta_k h_k(\mathbf{f}, \mathbf{e}')\Big)}$$

(Och and Ney, 2002; Och, 2003)

Generative Vs. Discriminative

generative model

$$P(\mathbf{e}|\mathbf{f}) = \frac{P(\mathbf{e}) \times P(\mathbf{f}|\mathbf{e})}{P(\mathbf{f})}$$

discriminative model

$$P(\mathbf{e}|\mathbf{f}) = \frac{\exp\Big(\sum_{k=1}^{K} \theta_k h_k(\mathbf{f}, \mathbf{e})\Big)}{\sum_{\mathbf{e}'} \exp\Big(\sum_{k=1}^{K} \theta_k h_k(\mathbf{f}, \mathbf{e}')\Big)}$$

$$score(\mathbf{e}, \mathbf{f}) = \sum_{k=1}^{K} \theta_k h_k(\mathbf{f}, \mathbf{e})$$

(Och and Ney, 2002; Och, 2003)

Generative Vs. Discriminative

generative model

$$P(\mathbf{e}|\mathbf{f}) = \frac{P(\mathbf{e}) \times P(\mathbf{f}|\mathbf{e})}{P(\mathbf{f})}$$

discriminative model

$$P(\mathbf{e}|\mathbf{f}) = \frac{\exp\Big(\sum_{k=1}^{K} \theta_k h_k(\mathbf{f}, \mathbf{e})\Big)}{\sum_{\mathbf{e}'} \exp\Big(\sum_{k=1}^{K} \theta_k h_k(\mathbf{f}, \mathbf{e}')\Big)}$$

discriminant function

$$score(\mathbf{e}, \mathbf{f}) = \sum_{k=1}^{K} \theta_k h_k(\mathbf{f}, \mathbf{e})$$

(Och and Ney, 2002; Och, 2003)

Generative Vs. Discriminative

- the advantages of discriminative models include
  - accessible to arbitrary overlapping knowledge sources
  - distinguish the contributions between different knowledge sources
- generative models are a special case of discriminative models

Features

- The following features are widely used in phrase-based discriminative translation models:
  - phrase translation probabilities
  - phrase lexical weights
  - phrase penalty
  - reordering models
  - language models
  - word penalty

(Och and Ney, 2002; Och, 2003)

Optimizing Feature Weights

θ_1	θ_2	θ_3
1.0	1.0	1.0

cand	h_1	h_2	h_3	$score$	$eval$
e_1	-85	4	10		
e_2	-89	3	12		
e_3	-93	6	11		

(Och, 2003)

Optimizing Feature Weights

θ_1	θ_2	θ_3
1.0	1.0	1.0

cand	h_1	h_2	h_3	$score$	$eval$
e_1	-85	4	10	-71	
e_2	-89	3	12	-74	
e_3	-93	6	11	-76	

(Och, 2003)

Optimizing Feature Weights

θ_1	θ_2	θ_3
1.0	1.0	1.0

cand	h_1	h_2	h_3	$score$	$eval$
e_1	-85	4	10	-71	0.7
e_2	-89	3	12	-74	0.9
e_3	-93	6	11	-76	0.6

(Och, 2003)

Optimizing Feature Weights

θ_1	θ_2	θ_3
1.0	1.0	1.0

cand	h_1	h_2	h_3	$score$	$eval$
e_1	-85	4	10	-71	0.7
e_2	-89	3	12	-74	0.9
e_3	-93	6	11	-76	0.6

(Och, 2003)

Optimizing Feature Weights

θ_1	θ_2	θ_3
1.0	-2.0	-2.0

cand	h_1	h_2	h_3	$score$	$eval$
e_1	-85	4	10		0.7
e_2	-89	3	12		0.9
e_3	-93	6	11		0.6

(Och, 2003)

Optimizing Feature Weights

θ_1	θ_2	θ_3
1.0	-2.0	-2.0

cand	h_1	h_2	h_3	$score$	$eval$
e_1	-85	4	10	-73	0.7
e_2	-89	3	12	-71	0.9
e_3	-93	6	11	-83	0.6

(Och, 2003)

Optimizing Feature Weights

θ_1	θ_2	θ_3
1.0	1.0	1.0

cand	h_1	h_2	h_3	$score$	$eval$
e_1	-85	4	10	-73	0.7
e_2	-89	3	12	-71	0.9
e_3	-93	6	11	-83	0.6

(Och, 2003)

Optimizing Feature Weights

θ_1	θ_2	θ_3
1.0	1.0	1.0

⇧

cand	h_1	h_2	h_3	$score$	$eval$
e_1	-85	4	10	-73	0.7
e_2	-89	3	12	-71	0.9
e_3	-93	6	11	-83	0.6

(Och, 2003)

Optimizing Feature Weights

line 1: $-81 + 10x$

θ_1	θ_2	θ_3
1.0	1.0	1.0

cand	h_1	h_2	h_3	$score$	$eval$
e_1	-85	4	10	-73	0.7
e_2	-89	3	12	-71	0.9
e_3	-93	6	11	-83	0.6

(Och, 2003)

Optimizing Feature Weights

$$\text{line 1:} \quad -81 + 10x$$
$$\text{line 2:} \quad -86 + 12x$$

θ_1	θ_2	θ_3
1.0	1.0	1.0

cand	h_1	h_2	h_3	*score*	*eval*
e_1	-85	4	10	-73	0.7
e_2	-89	3	12	-71	0.9
e_3	-93	6	11	-83	0.6

(Och, 2003)

Optimizing Feature Weights

$$\text{line 1:} \quad -81 + 10x$$
$$\text{line 2:} \quad -86 + 12x$$
$$\text{line 3:} \quad -87 + 11x$$

θ_1	θ_2	θ_3
1.0	1.0	1.0

⇧

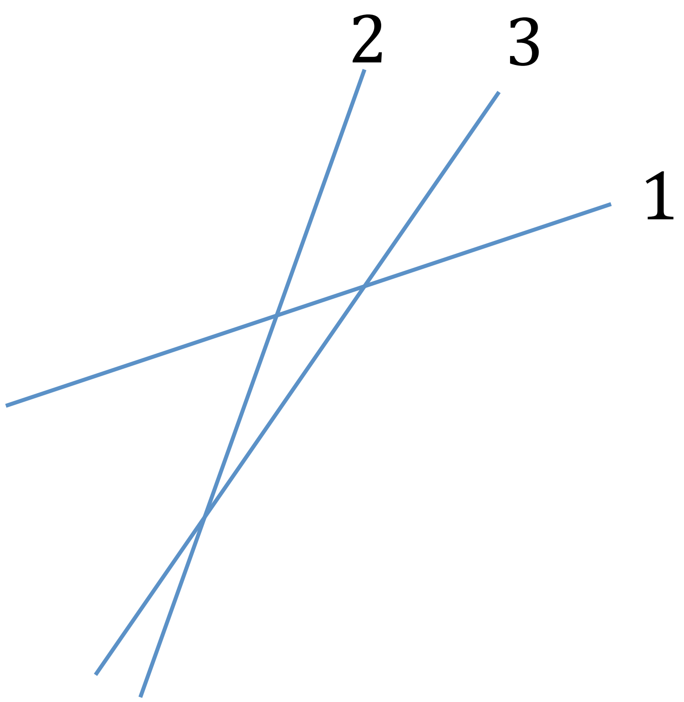
cand	h_1	h_2	h_3	*score*	*eval*
e_1	-85	4	10	-73	0.7
e_2	-89	3	12	-71	0.9
e_3	-93	6	11	-83	0.6

(Och, 2003)

Optimizing Feature Weights

$$\text{line 1:} \quad -81 + 10x$$
$$\text{line 2:} \quad -86 + 12x$$
$$\text{line 3:} \quad -87 + 11x$$

θ_1	θ_2	θ_3
1.0	1.0	1.0



cand	h_1	h_2	h_3	$score$	$eval$
e_1	-85	4	10	-73	0.7
e_2	-89	3	12	-71	0.9
e_3	-93	6	11	-83	0.6

(Och, 2003)

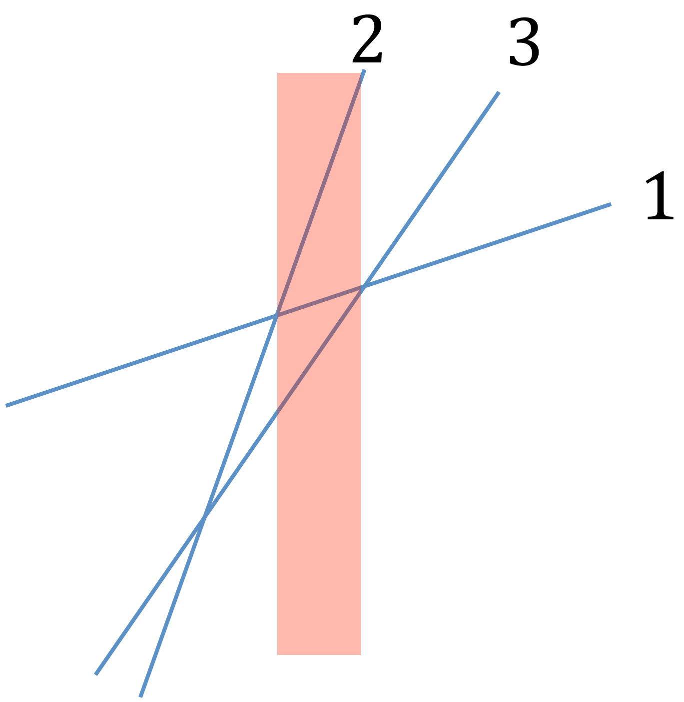
Optimizing Feature Weights

line 1: $-81 + 10x$

line 2: $-86 + 12x$

line 3: $-87 + 11x$

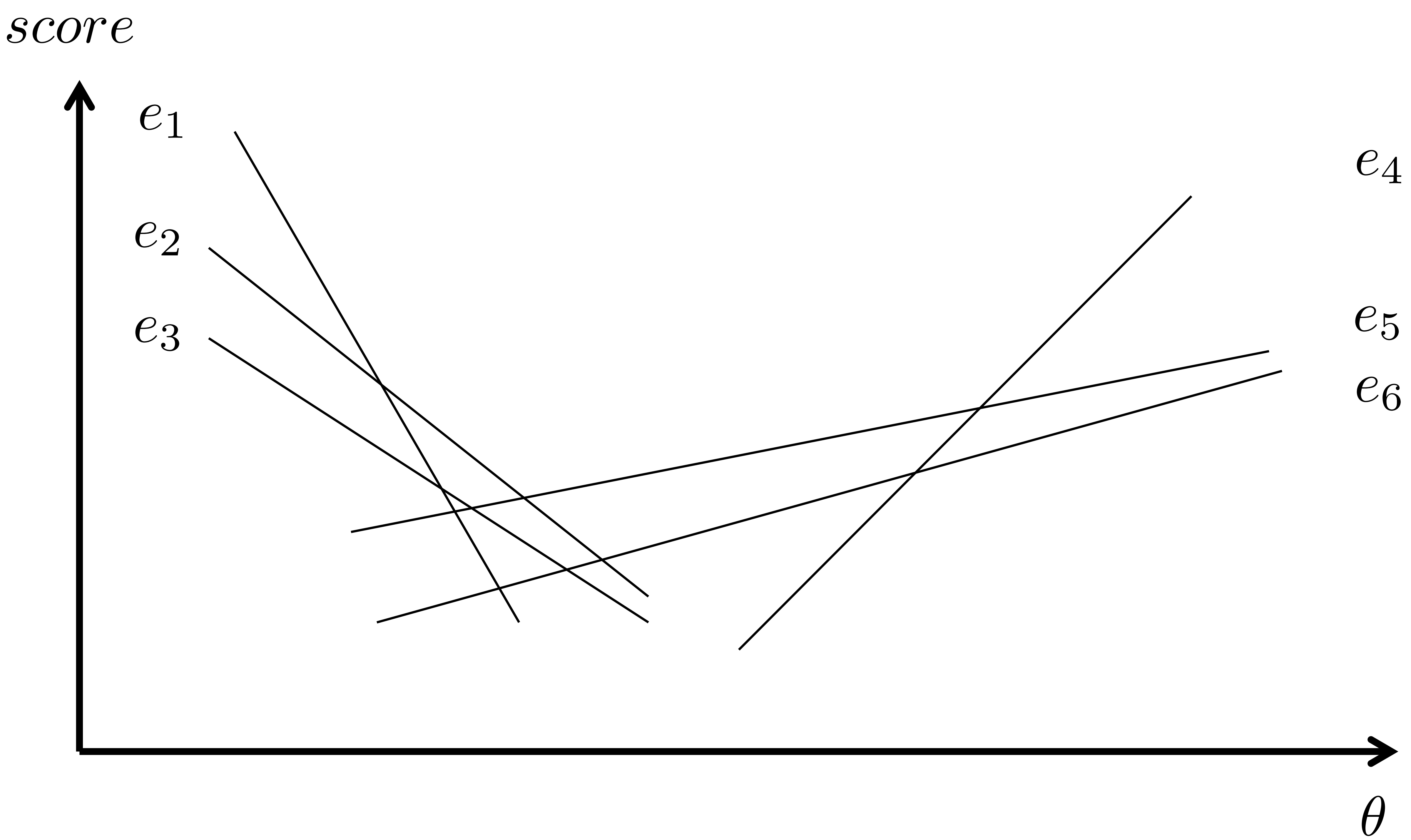
θ_1	θ_2	θ_3
1.0	1.0	1.0



cand	h_1	h_2	h_3	$score$	$eval$
e_1	-85	4	10	-73	0.7
e_2	-89	3	12	-71	0.9
e_3	-93	6	11	-83	0.6

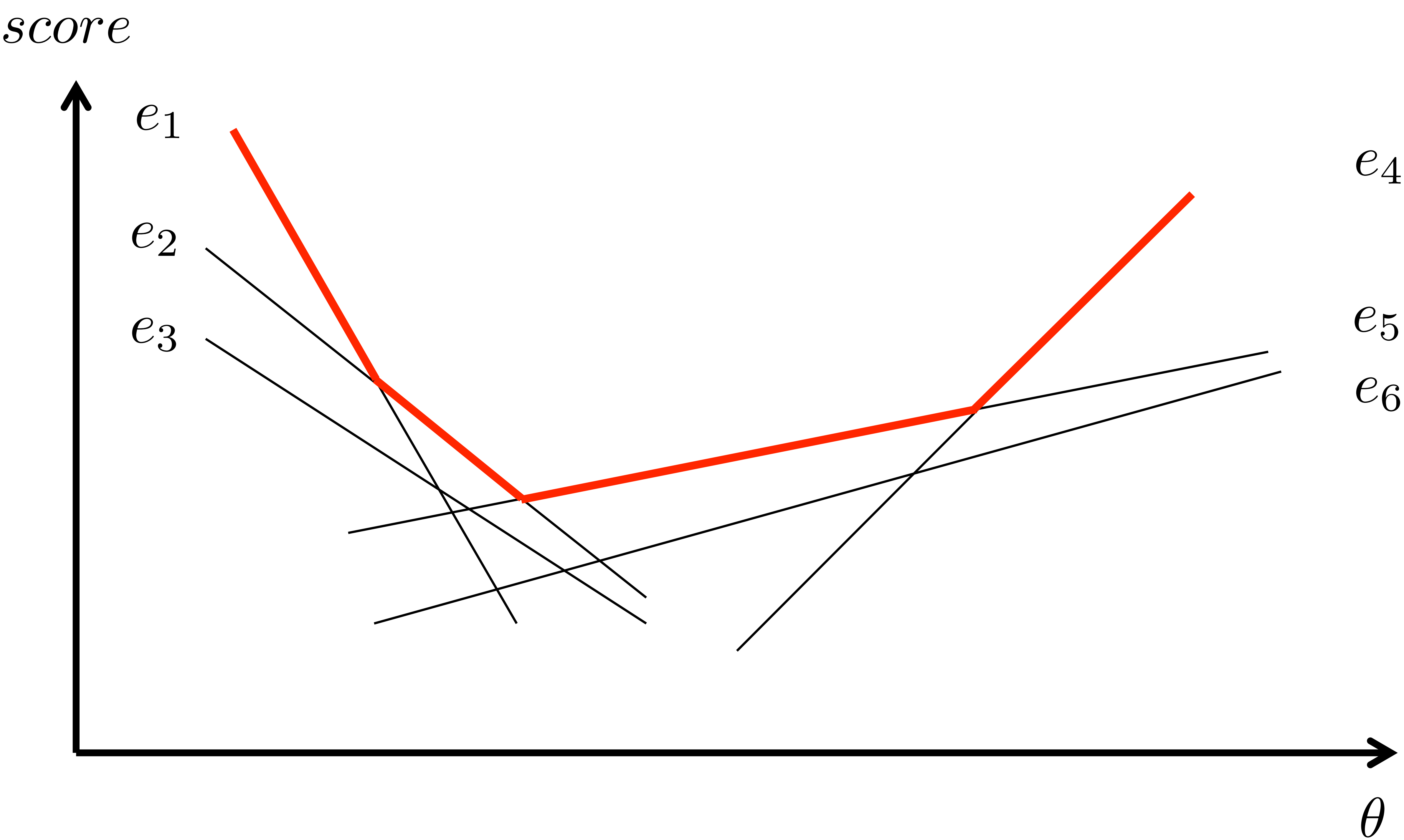
(Och, 2003)

Minimum Error Rate Training



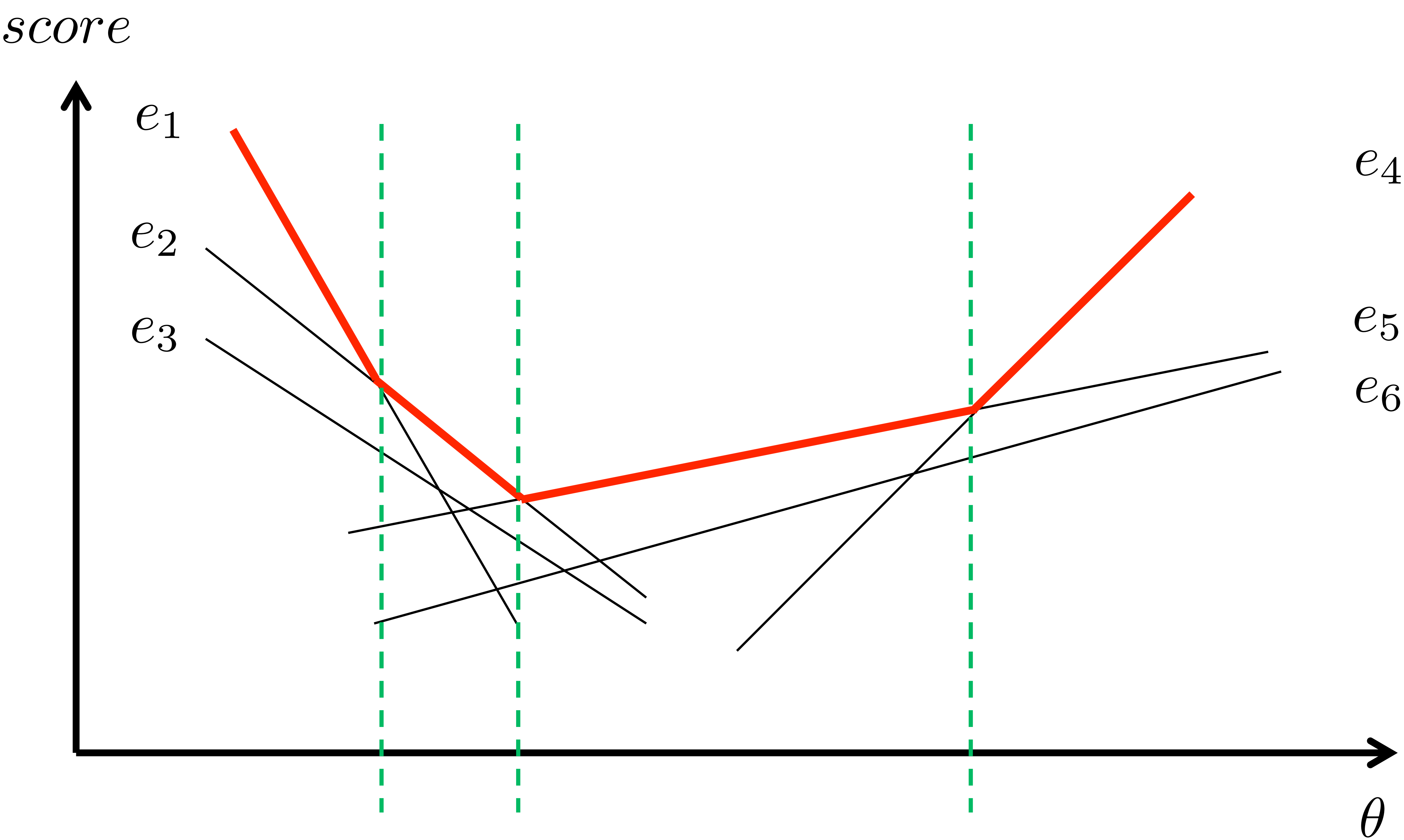
(Och, 2003)

Minimum Error Rate Training



(Och, 2003)

Minimum Error Rate Training



(Och, 2003)

Decoding

布什	与	沙龙	举行	了	会谈

(Koehn et al., 2007)

Decoding

布什	与	沙龙	举行	了	会谈
Bush	with	Sharon	hold	have	talk
	and		held		a talk

(Koehn et al., 2007)

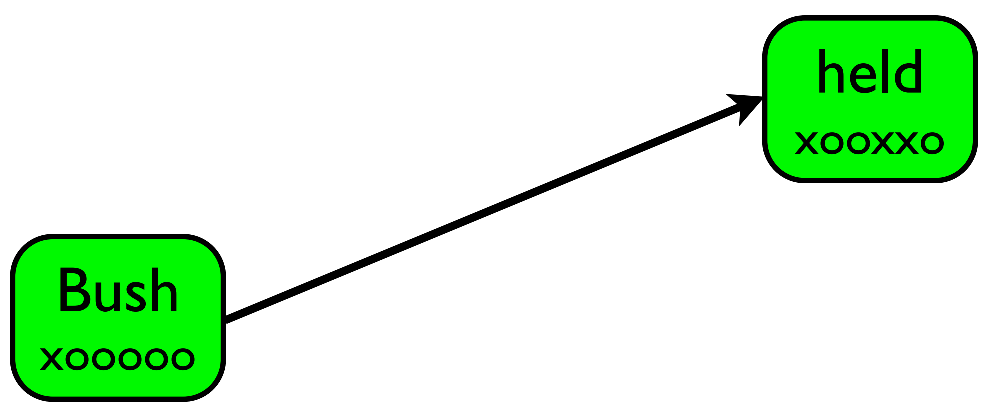
Decoding

布什	与	沙龙	举行	了	会谈
Bush	with	Sharon	hold	have	talk
	and		held		a talk

(Koehn et al., 2007)

Decoding

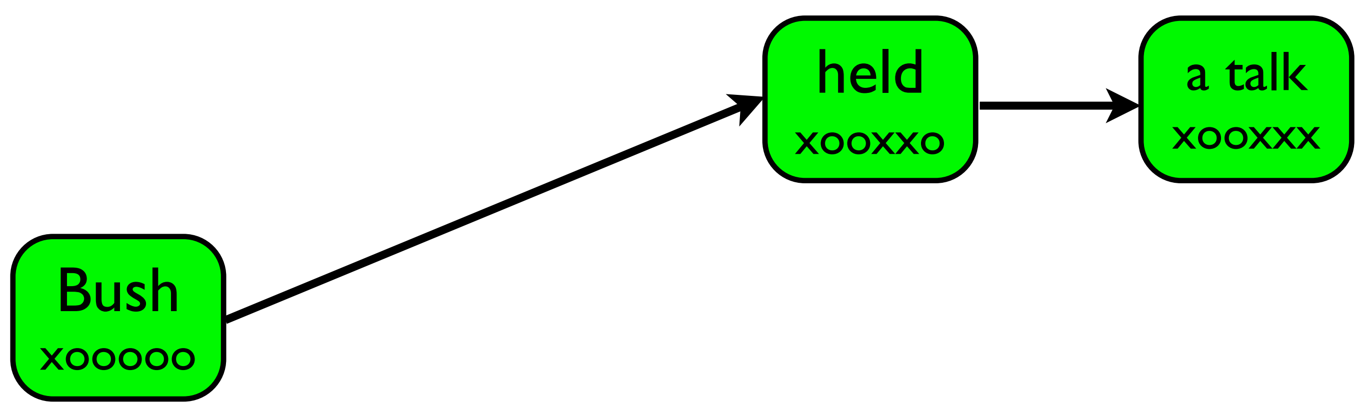
布什	与	沙龙	举行	了	会谈
Bush	with	Sharon	hold	have	talk
	and		held		a talk



(Koehn et al., 2007)

Decoding

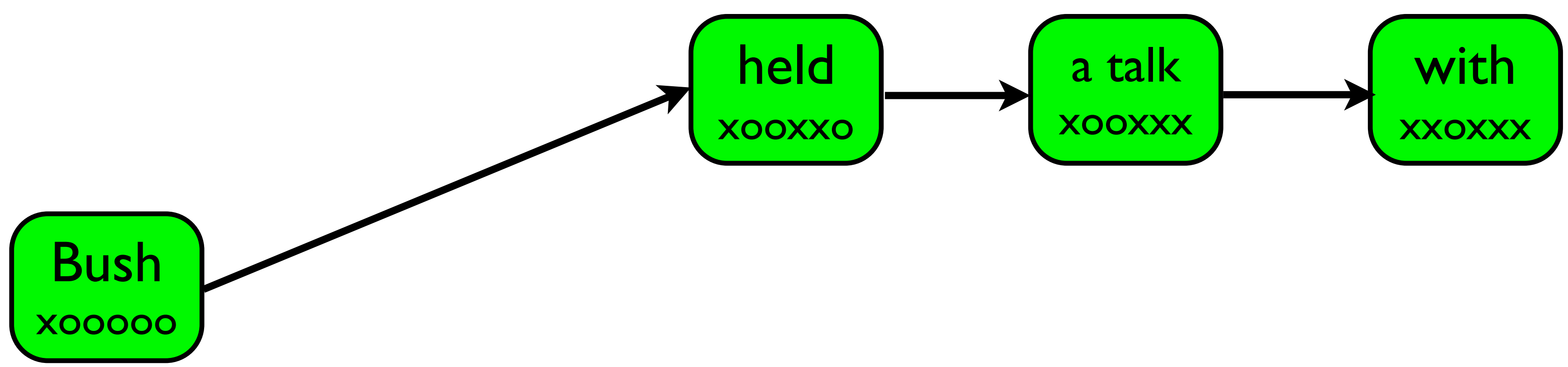
布什	与	沙龙	举行	了	会谈
Bush	with	Sharon	hold	have	talk
	and		held		a talk



(Koehn et al., 2007)

Decoding

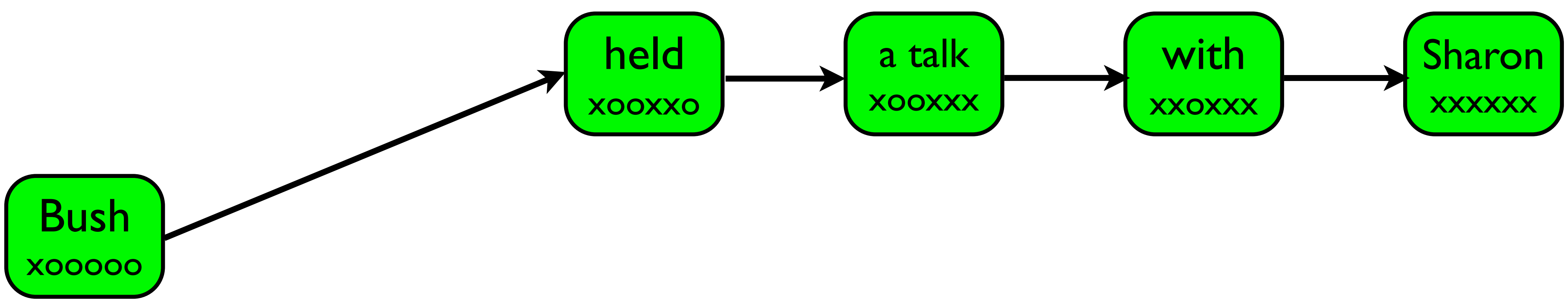
布什	与	沙龙	举行	了	会谈
Bush	with	Sharon	hold	have	talk
	and		held		a talk



(Koehn et al., 2007)

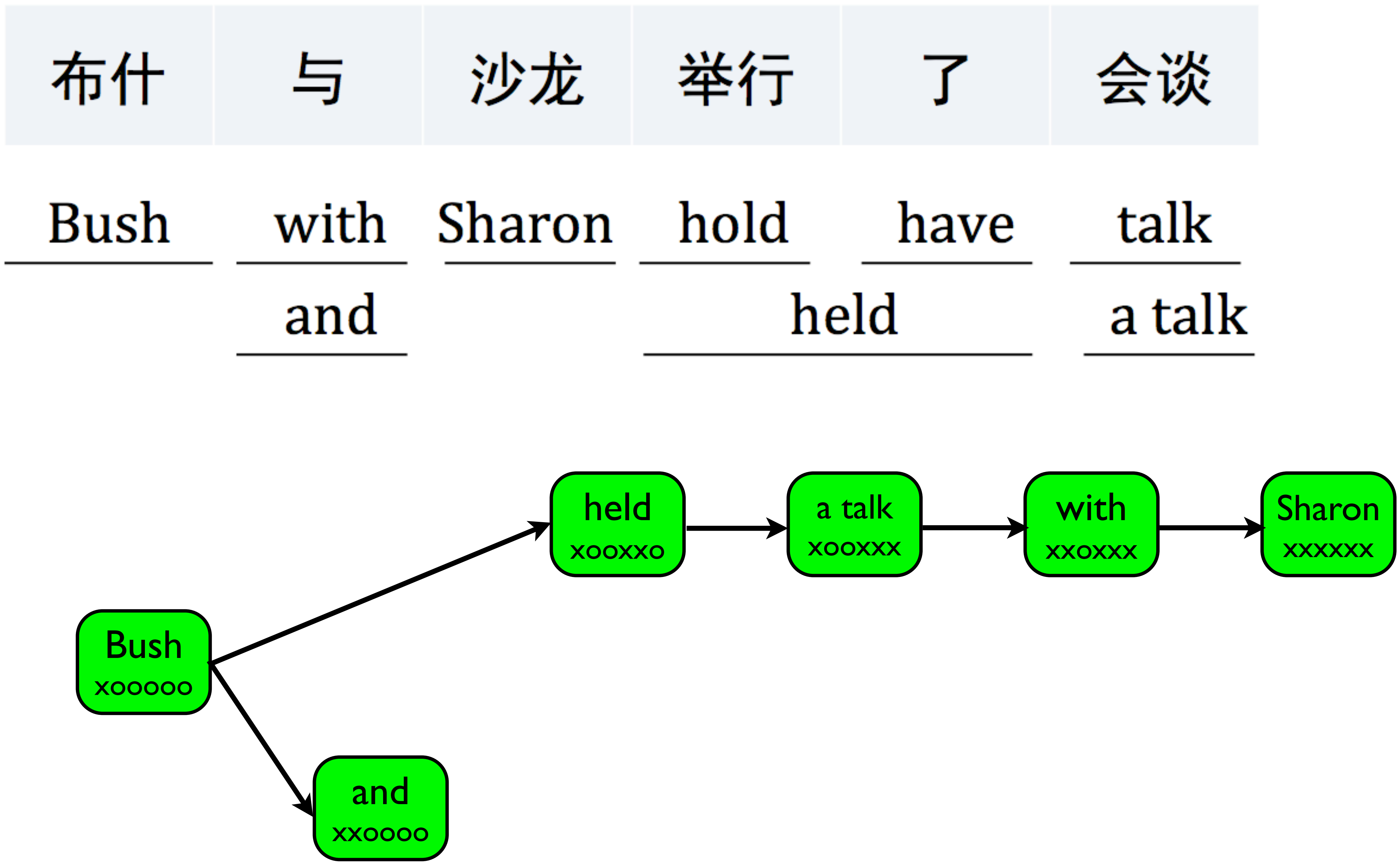
Decoding

布什	与	沙龙	举行	了	会谈
Bush	with	Sharon	hold	have	talk
	and		held		a talk



(Koehn et al., 2007)

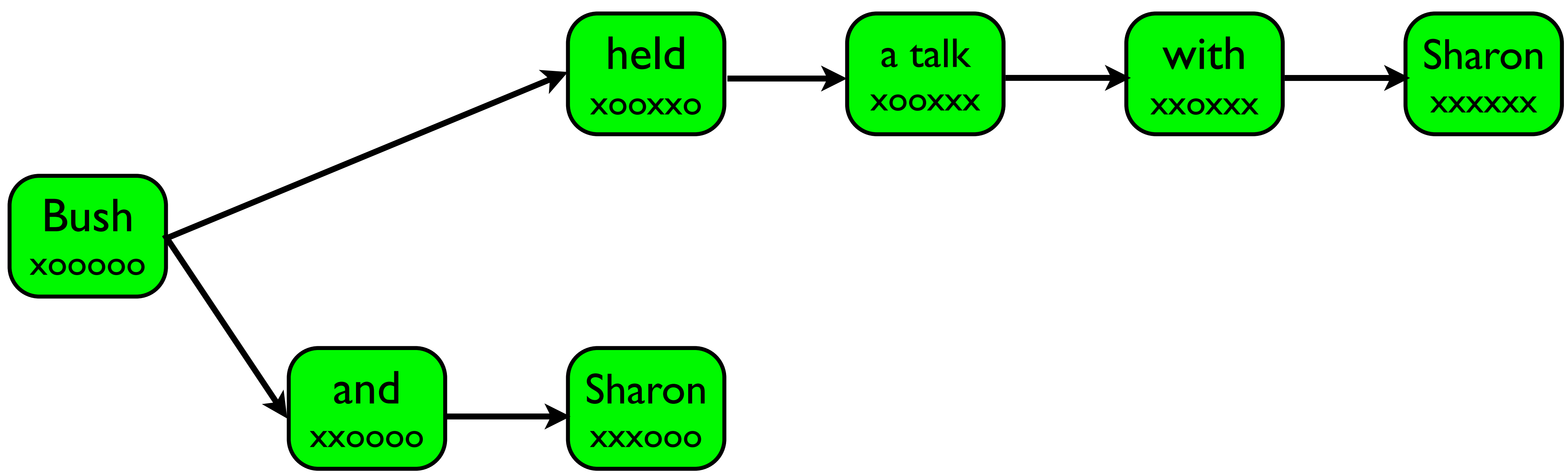
Decoding



(Koehn et al., 2007)

Decoding

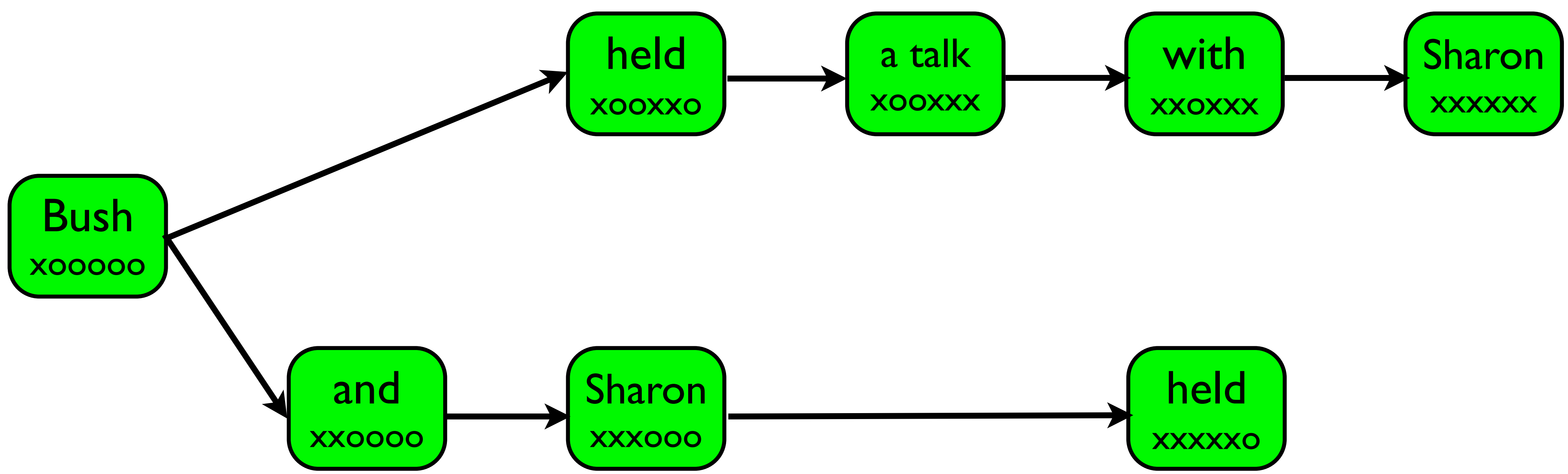
布什	与	沙龙	举行	了	会谈
Bush	with	Sharon	hold	have	talk
	and		held		a talk



(Koehn et al., 2007)

Decoding

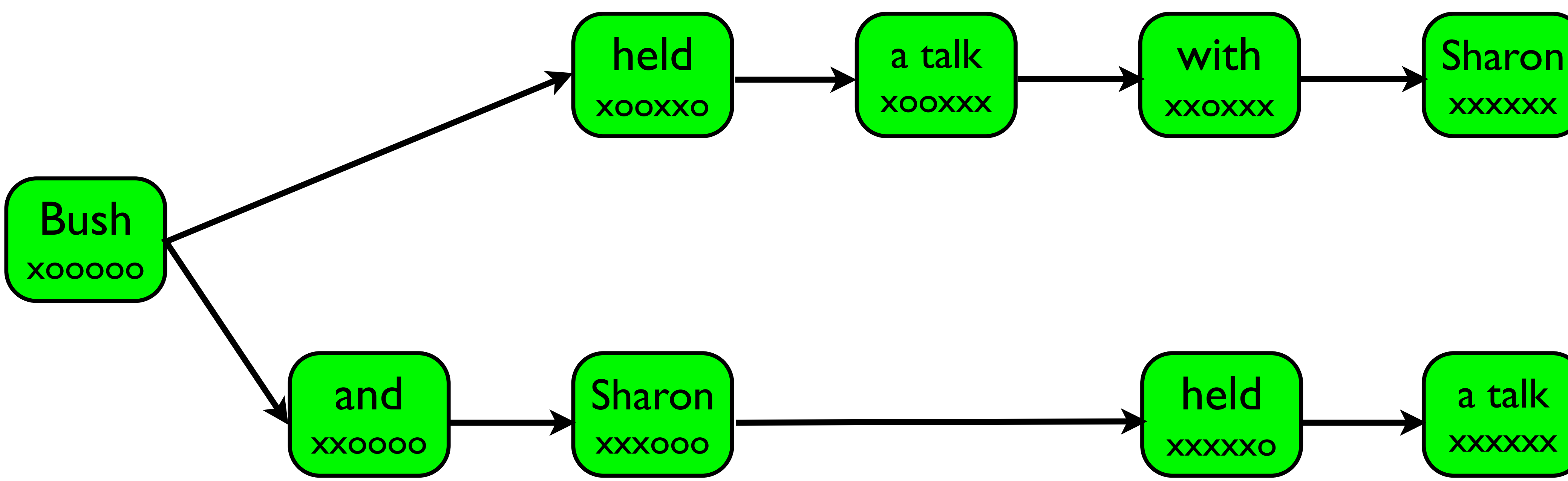
布什	与	沙龙	举行	了	会谈
Bush	with	Sharon	hold	have	talk
	and		held		a talk



(Koehn et al., 2007)

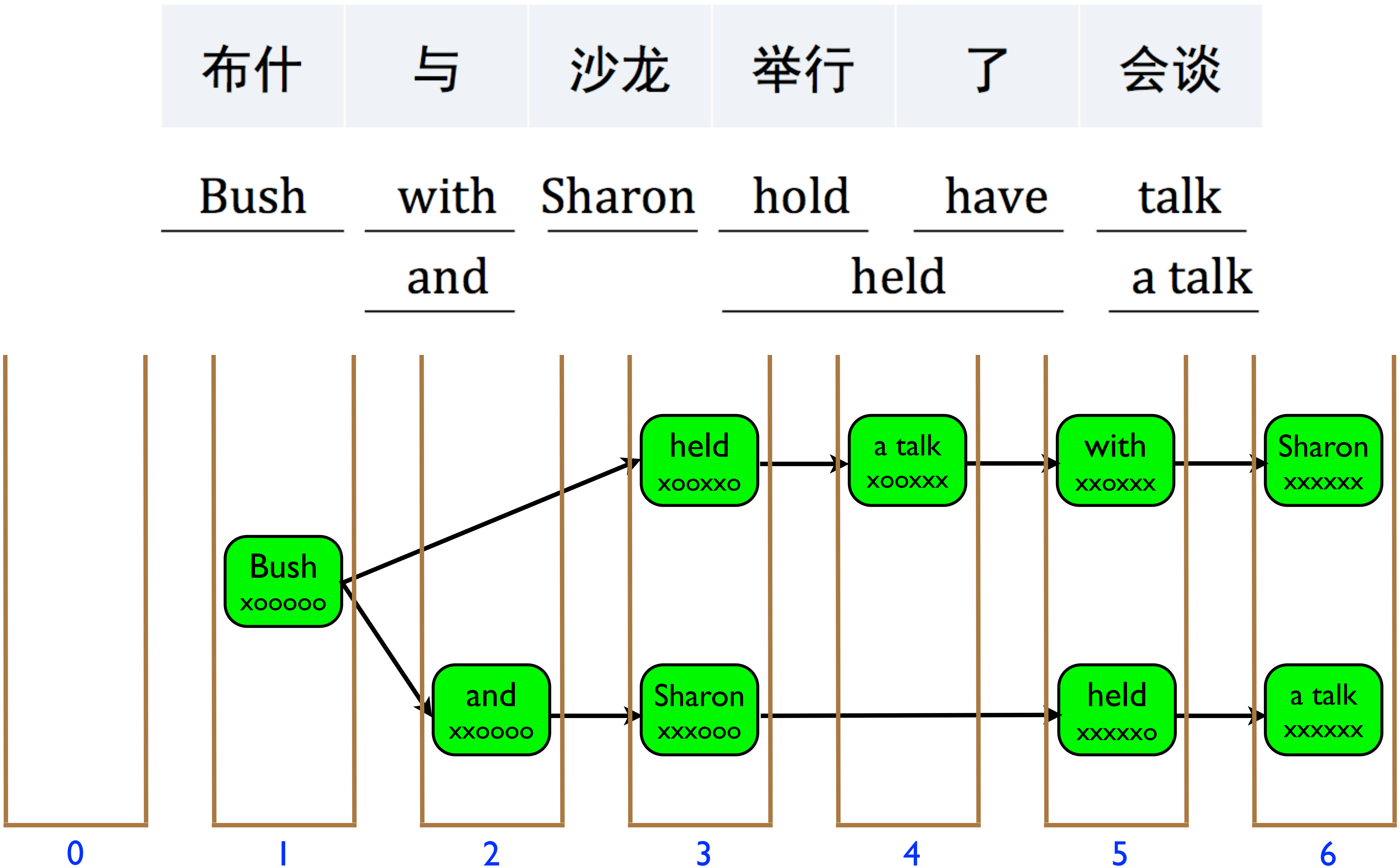
Decoding

布什	与	沙龙	举行	了	会谈
Bush	with	Sharon	hold	have	talk
	and		held		a talk



(Koehn et al., 2007)

Decoding



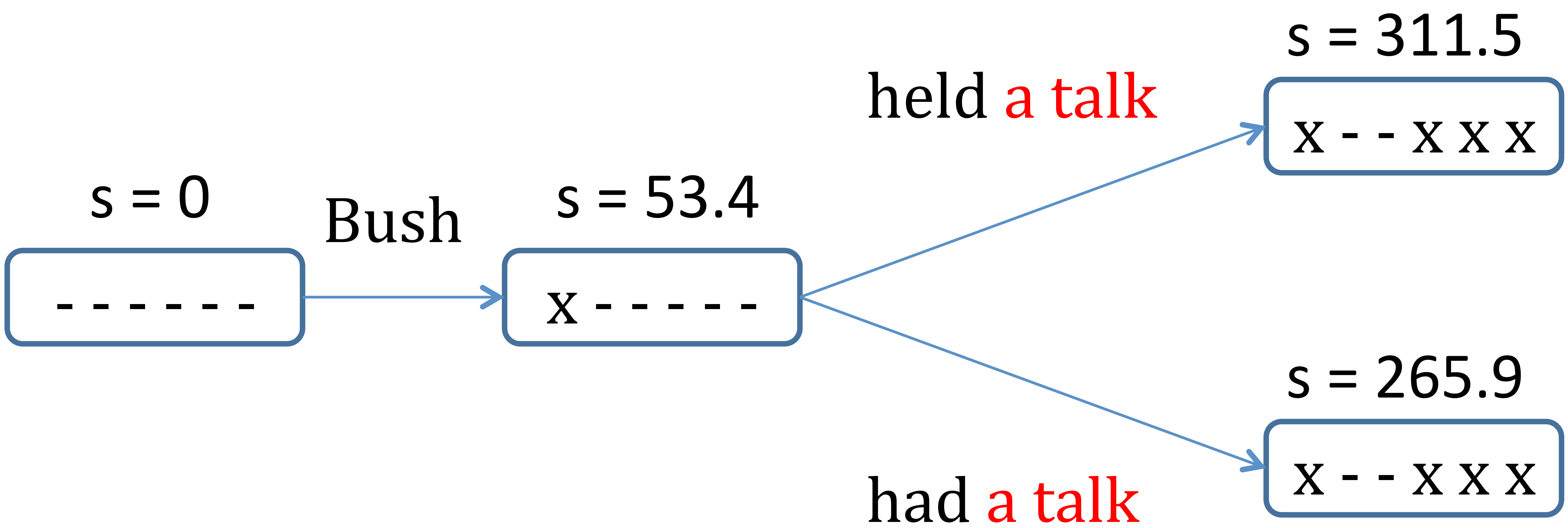
(Koehn et al., 2007)

Hypothesis

- a hypothesis (partial translation) consists of the following information
    - phrase pair ID
    - pointer to the previous hypothesis
    - coverage
    - last n-1 target words
    - the end of the last translated source phrase
    - feature value vector
    - current score
    - the estimate of future score
    - overall score
    - recombined hypotheses

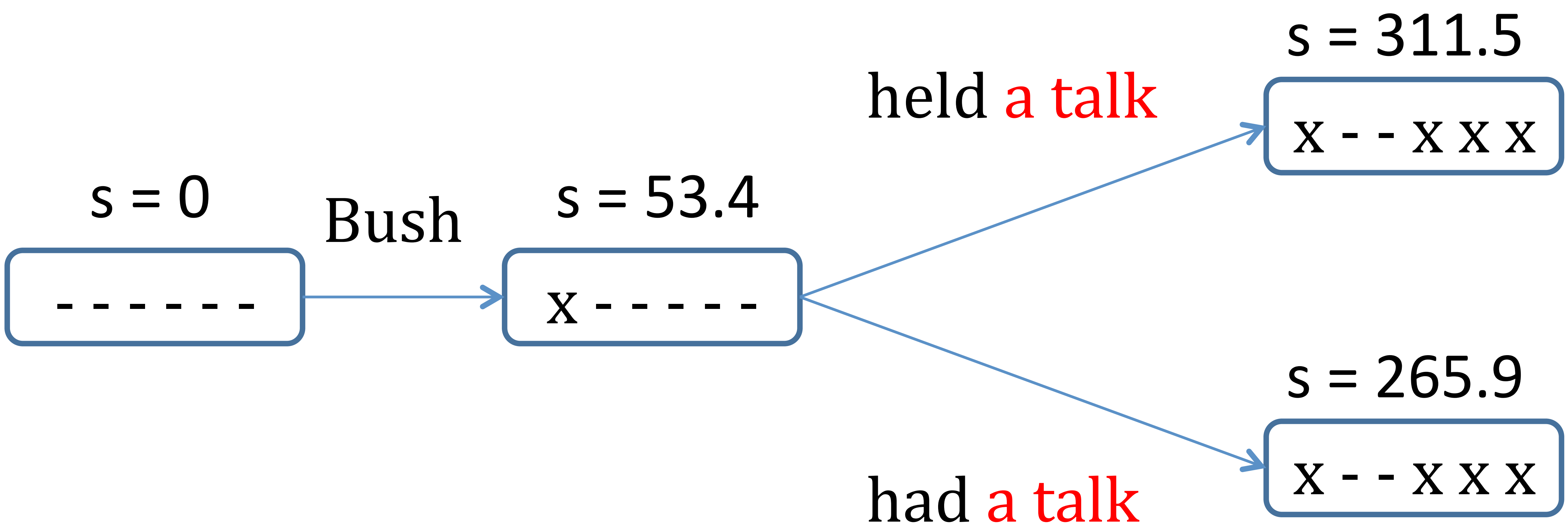
(Koehn et al., 2007)

Hypothesis Recombination



(Koehn et al., 2007)

Hypothesis Recombination



can be safely pruned!

(Koehn et al., 2007)

Hypothesis Recombination

- Two hypotheses can be recombined iff the following items are identical
  - coverage (translation model, phrase/word penalty)
  - last *n*-1 target words (language model)
  - the end of the last translated source phrase (reordering model)

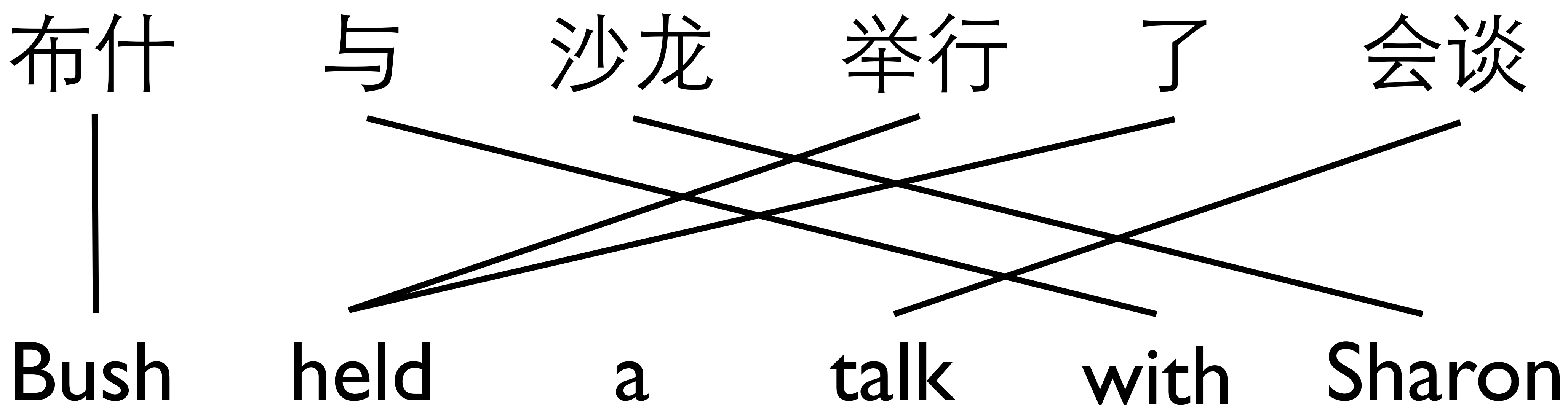
(Koehn et al., 2007)

Pruning

- Hypothesis recombination is risk-free pruning
- Two aggressive pruning methods are widely used to maintain a reasonable stack size:
  - retain at most *a* hypotheses in a stack
  - discard hypotheses *b* times worse than the best hypothesis in a stack

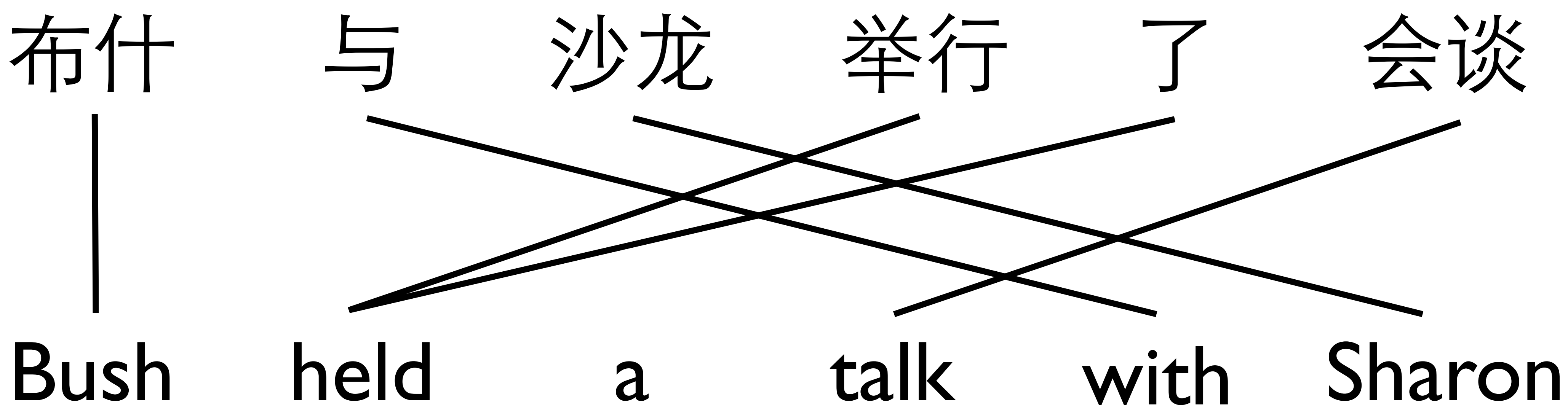
(Koehn et al., 2007)

Discontinuous Phrase-Based Model



(Galley and Manning, 2010)

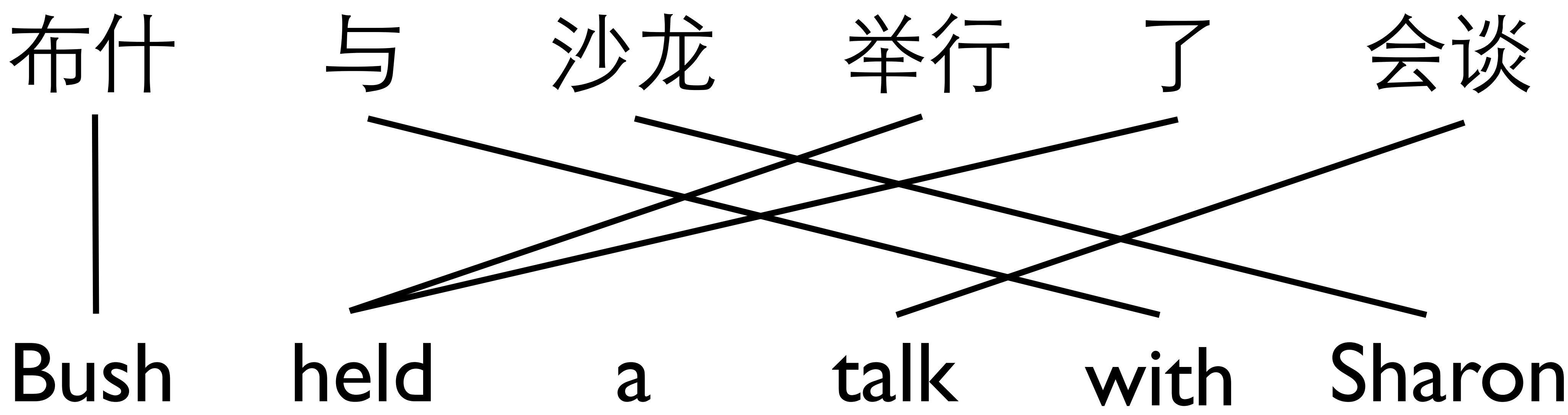
Discontinuous Phrase-Based Model



(布什 与, Bush ... with)

(Galley and Manning, 2010)

Discontinuous Phrase-Based Model

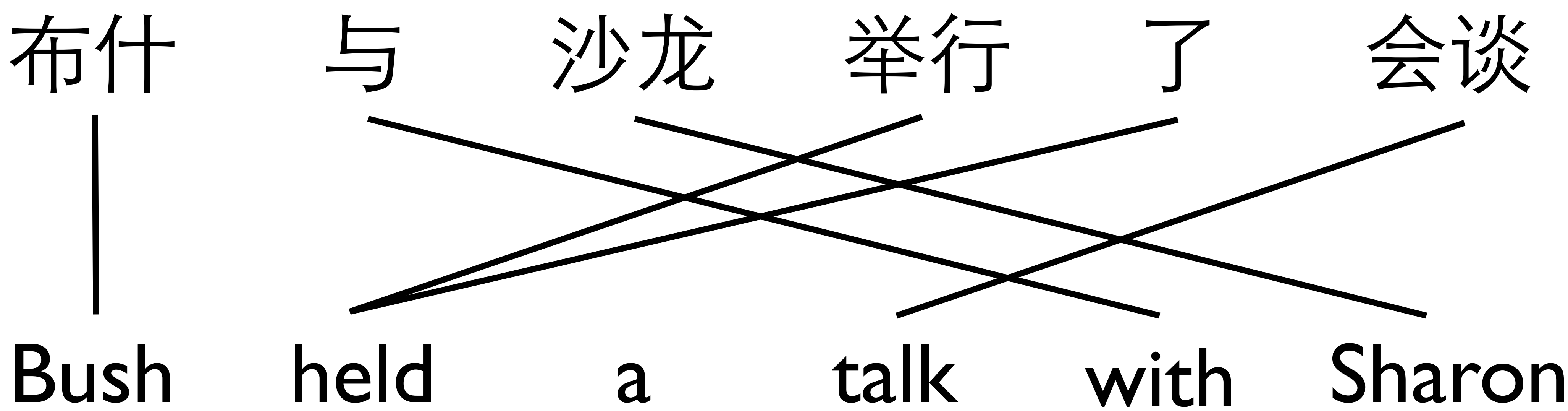


(布什 与, Bush ... with)

(布什 ... 举行 了, Bush held)

(Galley and Manning, 2010)

Discontinuous Phrase-Based Model



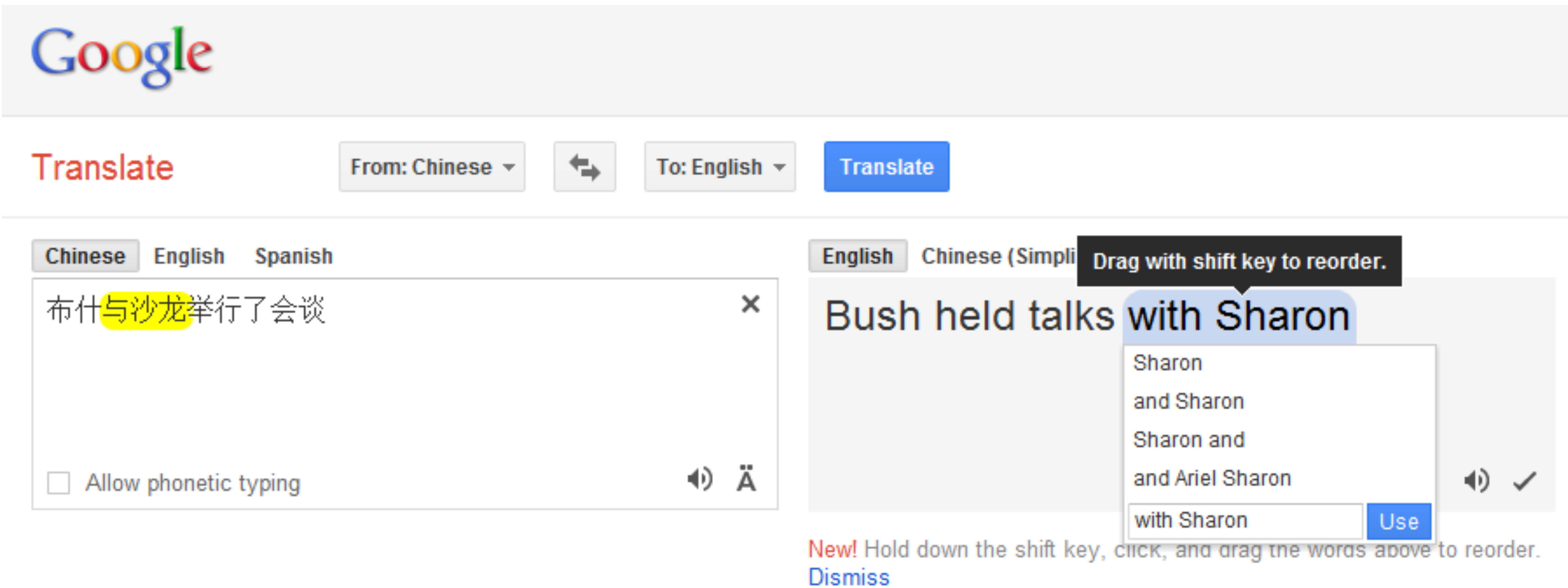
(布什 与, Bush ... with)

(布什 ... 举行 了, Bush held)

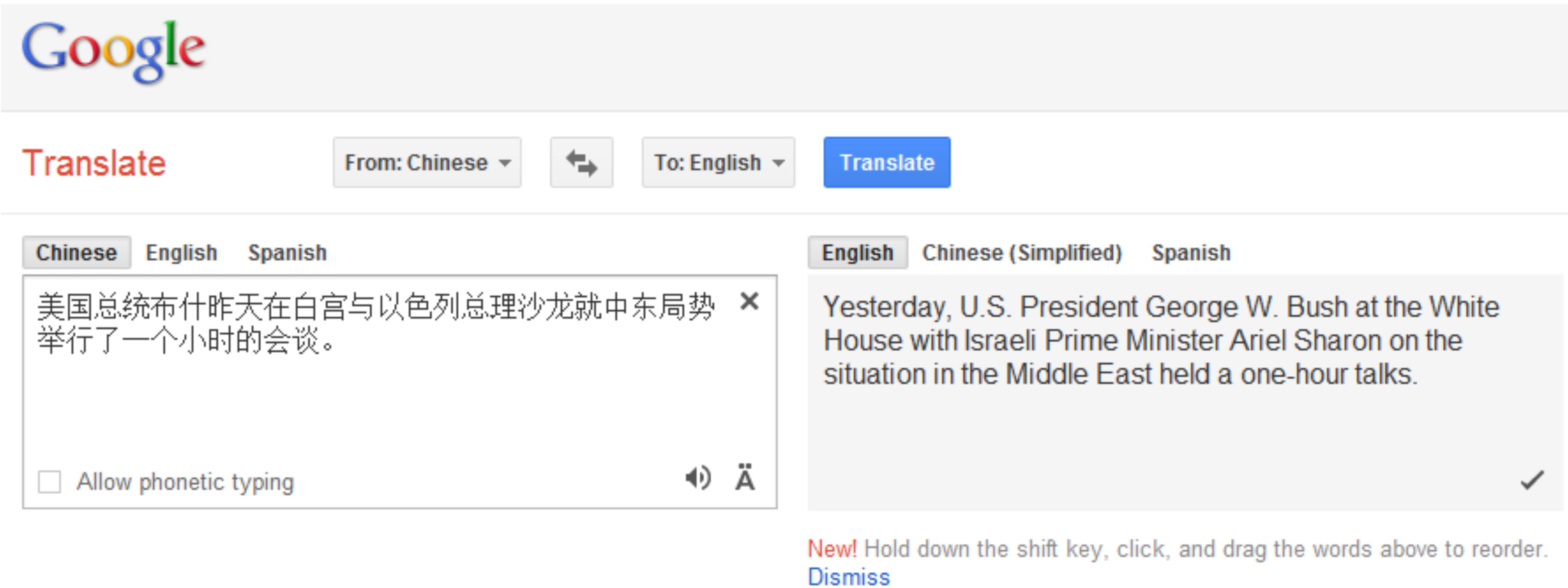
(与 ... 举行 了, held ... with)

(Galley and Manning, 2010)

Google Translate



Google Translate



Part 4: Syntax-based MT

Regularities in Natural Languages

- The way people say things has regularities

Every boy likes a car
The girl saw a dog
Prof. Wang gave a talk

Regularities in Natural Languages

- The way people say things has regularities

Every boy likes a car
The girl saw a dog
Prof. Wang gave a talk

Regularities in Natural Languages

- The way people say things has regularities

Every boy likes a car
The girl saw a dog
Prof. Wang gave a talk

How are the sentences generated?

Context-Free Grammar

- Context-free grammar describes how natural language sentences are generated

Context-Free Grammar

- Context-free grammar describes how natural language sentences are generated

lexical rules

Context-Free Grammar

- Context-free grammar describes how natural language sentences are generated

lexical rules

$$\text{NNP} \rightarrow \text{Bush}$$
$$\text{VBD} \rightarrow \text{held}$$
$$\text{DT} \rightarrow \text{a}$$
$$\text{NN} \rightarrow \text{talk}$$
$$\text{IN} \rightarrow \text{with}$$
$$\text{NNP} \rightarrow \text{Sharon}$$

Context-Free Grammar

- Context-free grammar describes how natural language sentences are generated

lexical rules

syntactic rules

$$\begin{aligned}
\text{NNP} &\rightarrow \text{Bush} \\
\text{VBD} &\rightarrow \text{held} \\
\text{DT} &\rightarrow \text{a} \\
\text{NN} &\rightarrow \text{talk} \\
\text{IN} &\rightarrow \text{with} \\
\text{NNP} &\rightarrow \text{Sharon}
\end{aligned}$$

Context-Free Grammar

- Context-free grammar describes how natural language sentences are generated

lexical rules	syntactic rules
$\mathrm{NNP} \rightarrow \mathrm{Bush}$	$\mathrm{NP} \rightarrow \mathrm{NNP}$
$\mathrm{VBD} \rightarrow \mathrm{held}$	$\mathrm{NP} \rightarrow \mathrm{DT\ NN}$
$\mathrm{DT} \rightarrow \mathrm{a}$	$\mathrm{PP} \rightarrow \mathrm{IN\ NP}$
$\mathrm{NN} \rightarrow \mathrm{talk}$	$\mathrm{VP} \rightarrow \mathrm{VBD\ NP\ PP}$
$\mathrm{IN} \rightarrow \mathrm{with}$	$\mathrm{S} \rightarrow \mathrm{NP\ VP}$
$\mathrm{NNP} \rightarrow \mathrm{Sharon}$	

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$\mathrm{S} \rightarrow \mathrm{NP\ VP}$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$\mathrm{S} \rightarrow \mathrm{NP}\ \mathrm{VP}$ $\quad\quad\quad$ $\mathrm{S} \Rightarrow \mathrm{NP}\ \mathrm{VP}$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$\mathrm{S} \rightarrow \mathrm{NP}\ \mathrm{VP}$ $\qquad$ $\mathrm{S} \Rightarrow \mathrm{NP}\ \mathrm{VP}$

$\mathrm{NP} \rightarrow \mathrm{NNP}$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$$\mathrm{S} \rightarrow \mathrm{NP\ VP}$$

$$\mathrm{NP} \rightarrow \mathrm{NNP}$$

$$\begin{aligned} \mathrm{S} &\Rightarrow \mathrm{NP\ VP} \\ &\Rightarrow \mathrm{NNP\ VP} \end{aligned}$$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$$\mathrm{S} \rightarrow \mathrm{NP\ VP}$$

$$\mathrm{NP} \rightarrow \mathrm{NNP}$$

$$\mathrm{NNP} \rightarrow \mathrm{Bush}$$

$$\begin{aligned}\mathrm{S} &\Rightarrow \mathrm{NP\ VP}\\ &\Rightarrow \mathrm{NNP\ VP}\end{aligned}$$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$$\mathrm{S} \rightarrow \mathrm{NP\ VP}$$
$$\mathrm{NP} \rightarrow \mathrm{NNP}$$
$$\mathrm{NNP} \rightarrow \mathrm{Bush}$$

$$\begin{aligned} \mathrm{S} &\Rightarrow \mathrm{NP\ VP} \\ &\Rightarrow \mathrm{NNP\ VP} \\ &\Rightarrow \mathrm{Bush\ VP} \end{aligned}$$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$$\begin{aligned} \mathrm{S} &\rightarrow \mathrm{NP\ VP} \\ \mathrm{NP} &\rightarrow \mathrm{NNP} \\ \mathrm{NNP} &\rightarrow \mathrm{Bush} \\ \mathrm{VP} &\rightarrow \mathrm{VBD\ NP\ PP} \end{aligned}$$

$$\begin{aligned} \mathrm{S} &\Rightarrow \mathrm{NP\ VP} \\ &\Rightarrow \mathrm{NNP\ VP} \\ &\Rightarrow \mathrm{Bush\ VP} \end{aligned}$$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$$\begin{array}{ll} \mathrm{S} \rightarrow \mathrm{NP\ VP} & \mathrm{S} \Rightarrow \mathrm{NP\ VP} \\ \mathrm{NP} \rightarrow \mathrm{NNP} & \quad \Rightarrow \mathrm{NNP\ VP} \\ \mathrm{NNP} \rightarrow \mathrm{Bush} & \quad \Rightarrow \mathrm{Bush\ VP} \\ \mathrm{VP} \rightarrow \mathrm{VBD\ NP\ PP} & \quad \Rightarrow \mathrm{Bush\ VBD\ NP\ PP} \end{array}$$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$$\mathrm{S} \rightarrow \mathrm{NP\ VP}$$
$$\mathrm{NP} \rightarrow \mathrm{NNP}$$
$$\mathrm{NNP} \rightarrow \mathrm{Bush}$$
$$\mathrm{VP} \rightarrow \mathrm{VBD\ NP\ PP}$$
$$\mathrm{VBD} \rightarrow \mathrm{held}$$

$$\begin{aligned} \mathrm{S} &\Rightarrow \mathrm{NP\ VP} \\ &\Rightarrow \mathrm{NNP\ VP} \\ &\Rightarrow \mathrm{Bush\ VP} \\ &\Rightarrow \mathrm{Bush\ VBD\ NP\ PP} \end{aligned}$$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$$
\begin{array}{ll}
\text{S} \rightarrow \text{NP VP} & \text{S} \Rightarrow \text{NP VP} \\
\text{NP} \rightarrow \text{NNP} & \quad \Rightarrow \text{NNP VP} \\
\text{NNP} \rightarrow \text{Bush} & \quad \Rightarrow \text{Bush VP} \\
\text{VP} \rightarrow \text{VBD NP PP} & \quad \Rightarrow \text{Bush VBD NP PP} \\
\text{VBD} \rightarrow \text{held} & \quad \Rightarrow \text{Bush held NP PP}
\end{array}
$$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$$\begin{array}{ll} \mathrm{S} \rightarrow \mathrm{NP\ VP} & \mathrm{S} \Rightarrow \mathrm{NP\ VP} \\ \mathrm{NP} \rightarrow \mathrm{NNP} & \quad \Rightarrow \mathrm{NNP\ VP} \\ \mathrm{NNP} \rightarrow \mathrm{Bush} & \quad \Rightarrow \mathrm{Bush\ VP} \\ \mathrm{VP} \rightarrow \mathrm{VBD\ NP\ PP} & \quad \Rightarrow \mathrm{Bush\ VBD\ NP\ PP} \\ \mathrm{VBD} \rightarrow \mathrm{held} & \quad \Rightarrow \mathrm{Bush\ held\ NP\ PP} \\ \mathrm{NP} \rightarrow \mathrm{DT\ NN} & \end{array}$$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$\text{S} \rightarrow \text{NP VP}$	$\text{S} \Rightarrow \text{NP VP}$
$\text{NP} \rightarrow \text{NNP}$	$\Rightarrow \text{NNP VP}$
$\text{NNP} \rightarrow \text{Bush}$	$\Rightarrow \text{Bush VP}$
$\text{VP} \rightarrow \text{VBD NP PP}$	$\Rightarrow \text{Bush VBD NP PP}$
$\text{VBD} \rightarrow \text{held}$	$\Rightarrow \text{Bush held NP PP}$
$\text{NP} \rightarrow \text{DT NN}$	$\Rightarrow \text{Bush held DT NN PP}$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$$
\begin{array}{ll}
S \rightarrow NP\ VP & S \Rightarrow NP\ VP \\
NP \rightarrow NNP & \Rightarrow NNP\ VP \\
NNP \rightarrow \text{Bush} & \Rightarrow \text{Bush}\ VP \\
VP \rightarrow VBD\ NP\ PP & \Rightarrow \text{Bush}\ VBD\ NP\ PP \\
VBD \rightarrow \text{held} & \Rightarrow \text{Bush held}\ NP\ PP \\
NP \rightarrow DT\ NN & \Rightarrow \text{Bush held}\ DT\ NN\ PP \\
DT \rightarrow \text{a} &
\end{array}
$$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$\mathrm{S} \rightarrow \mathrm{NP\ VP}$	$\mathrm{S} \Rightarrow \mathrm{NP\ VP}$
$\mathrm{NP} \rightarrow \mathrm{NNP}$	$\Rightarrow \mathrm{NNP\ VP}$
$\mathrm{NNP} \rightarrow \mathrm{Bush}$	$\Rightarrow \mathrm{Bush\ VP}$
$\mathrm{VP} \rightarrow \mathrm{VBD\ NP\ PP}$	$\Rightarrow \mathrm{Bush\ VBD\ NP\ PP}$
$\mathrm{VBD} \rightarrow \mathrm{held}$	$\Rightarrow \mathrm{Bush\ held\ NP\ PP}$
$\mathrm{NP} \rightarrow \mathrm{DT\ NN}$	$\Rightarrow \mathrm{Bush\ held\ DT\ NN\ PP}$
$\mathrm{DT} \rightarrow \mathrm{a}$	$\Rightarrow \mathrm{Bush\ held\ a\ NN\ PP}$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$$
\begin{array}{ll}
\mathrm{S} \rightarrow \mathrm{NP}\ \mathrm{VP} & \mathrm{S} \Rightarrow \mathrm{NP}\ \mathrm{VP} \\
\mathrm{NP} \rightarrow \mathrm{NNP} & \Rightarrow \mathrm{NNP}\ \mathrm{VP} \\
\mathrm{NNP} \rightarrow \mathrm{Bush} & \Rightarrow \mathrm{Bush}\ \mathrm{VP} \\
\mathrm{VP} \rightarrow \mathrm{VBD}\ \mathrm{NP}\ \mathrm{PP} & \Rightarrow \mathrm{Bush}\ \mathrm{VBD}\ \mathrm{NP}\ \mathrm{PP} \\
\mathrm{VBD} \rightarrow \mathrm{held} & \Rightarrow \mathrm{Bush}\ \mathrm{held}\ \mathrm{NP}\ \mathrm{PP} \\
\mathrm{NP} \rightarrow \mathrm{DT}\ \mathrm{NN} & \Rightarrow \mathrm{Bush}\ \mathrm{held}\ \mathrm{DT}\ \mathrm{NN}\ \mathrm{PP} \\
\mathrm{DT} \rightarrow \mathrm{a} & \Rightarrow \mathrm{Bush}\ \mathrm{held}\ \mathrm{a}\ \mathrm{NN}\ \mathrm{PP} \\
\ldots &
\end{array}
$$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$$
\begin{array}{ll}
S \rightarrow NP\ VP & S \Rightarrow NP\ VP \\
NP \rightarrow NNP & \quad \Rightarrow NNP\ VP \\
NNP \rightarrow \text{Bush} & \quad \Rightarrow \text{Bush}\ VP \\
VP \rightarrow VBD\ NP\ PP & \quad \Rightarrow \text{Bush}\ VBD\ NP\ PP \\
VBD \rightarrow \text{held} & \quad \Rightarrow \text{Bush held}\ NP\ PP \\
NP \rightarrow DT\ NN & \quad \Rightarrow \text{Bush held}\ DT\ NN\ PP \\
DT \rightarrow \text{a} & \quad \Rightarrow \text{Bush held a}\ NN\ PP \\
\ldots & \quad \ldots
\end{array}
$$

Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

$\mathrm{S} \rightarrow \mathrm{NP}\ \mathrm{VP}$	$\mathrm{S} \Rightarrow \mathrm{NP}\ \mathrm{VP}$
$\mathrm{NP} \rightarrow \mathrm{NNP}$	$\Rightarrow \mathrm{NNP}\ \mathrm{VP}$
$\mathrm{NNP} \rightarrow \mathrm{Bush}$	$\Rightarrow \mathrm{Bush}\ \mathrm{VP}$
$\mathrm{VP} \rightarrow \mathrm{VBD}\ \mathrm{NP}\ \mathrm{PP}$	$\Rightarrow \mathrm{Bush}\ \mathrm{VBD}\ \mathrm{NP}\ \mathrm{PP}$
$\mathrm{VBD} \rightarrow \mathrm{held}$	$\Rightarrow \mathrm{Bush}\ \mathrm{held}\ \mathrm{NP}\ \mathrm{PP}$
$\mathrm{NP} \rightarrow \mathrm{DT}\ \mathrm{NN}$	$\Rightarrow \mathrm{Bush}\ \mathrm{held}\ \mathrm{DT}\ \mathrm{NN}\ \mathrm{PP}$
$\mathrm{DT} \rightarrow \mathrm{a}$	$\Rightarrow \mathrm{Bush}\ \mathrm{held}\ \mathrm{a}\ \mathrm{NN}\ \mathrm{PP}$
$\ldots$	$\ldots$
$\mathrm{NNP} \rightarrow \mathrm{Sharon}$	

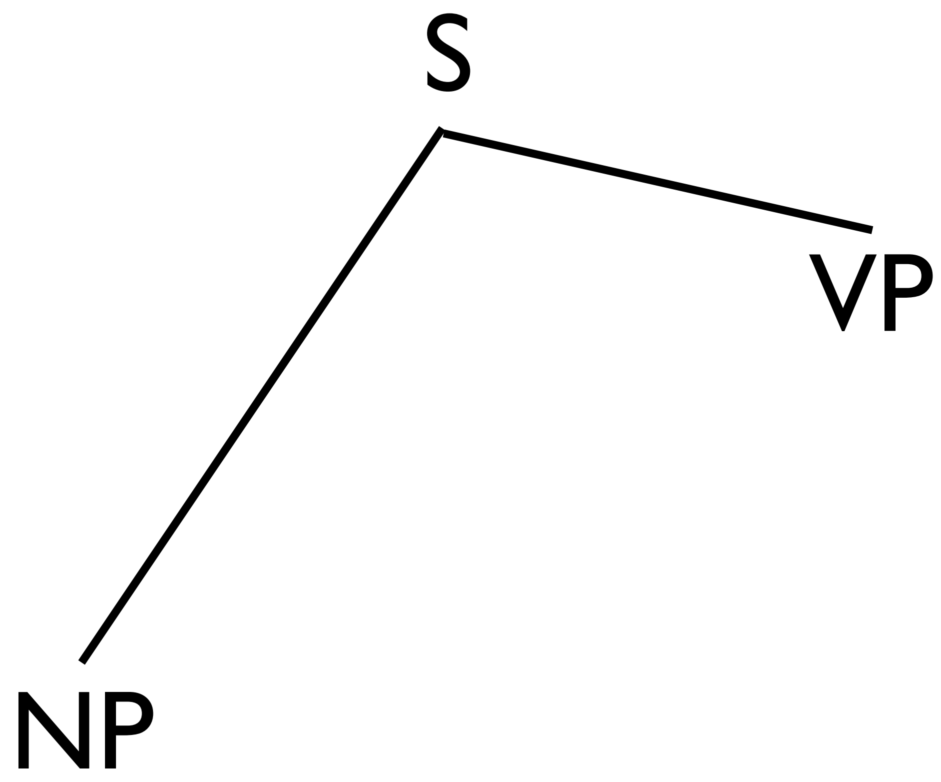
Derivation

- A derivation explains how a sentence can be generated by applying CFG rules

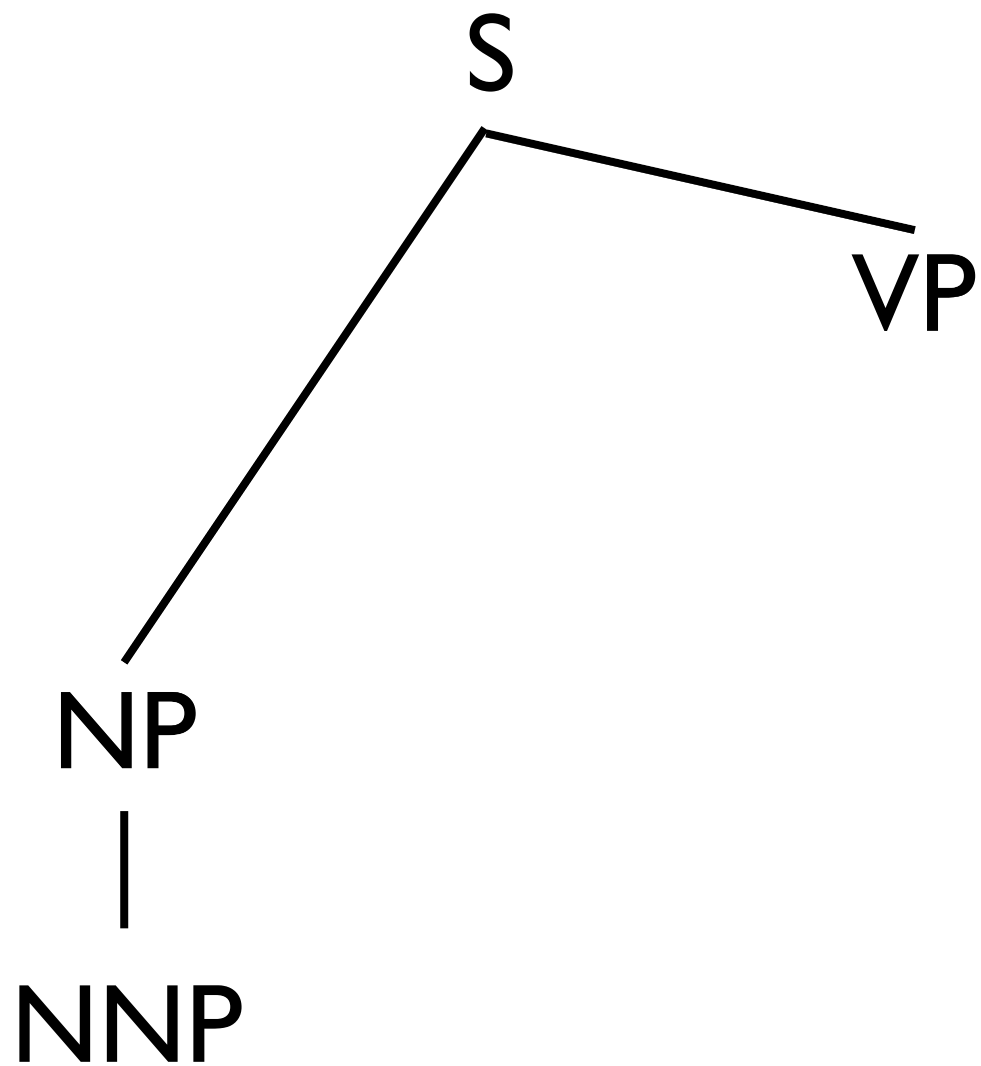
$S \rightarrow NP\ VP$	$S \Rightarrow NP\ VP$
$NP \rightarrow NNP$	$\Rightarrow NNP\ VP$
$NNP \rightarrow \text{Bush}$	$\Rightarrow \text{Bush}\ VP$
$VP \rightarrow VBD\ NP\ PP$	$\Rightarrow \text{Bush}\ VBD\ NP\ PP$
$VBD \rightarrow \text{held}$	$\Rightarrow \text{Bush held}\ NP\ PP$
$NP \rightarrow DT\ NN$	$\Rightarrow \text{Bush held}\ DT\ NN\ PP$
$DT \rightarrow \text{a}$	$\Rightarrow \text{Bush held a}\ NN\ PP$
$\ldots$	$\ldots$
$NNP \rightarrow \text{Sharon}$	$\Rightarrow \text{Bush held a talk with Sharon}$

Graphical Representation

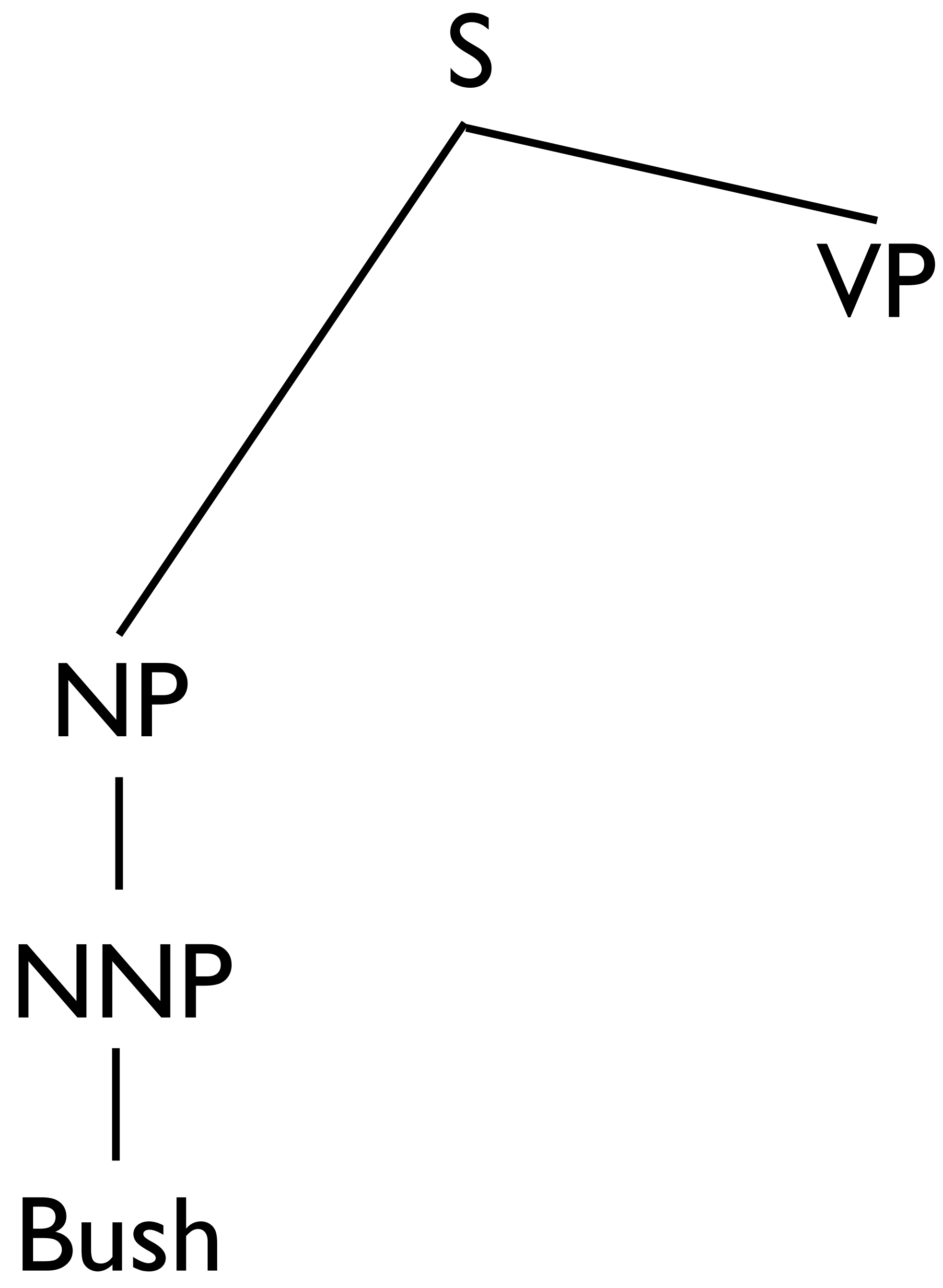
Graphical Representation



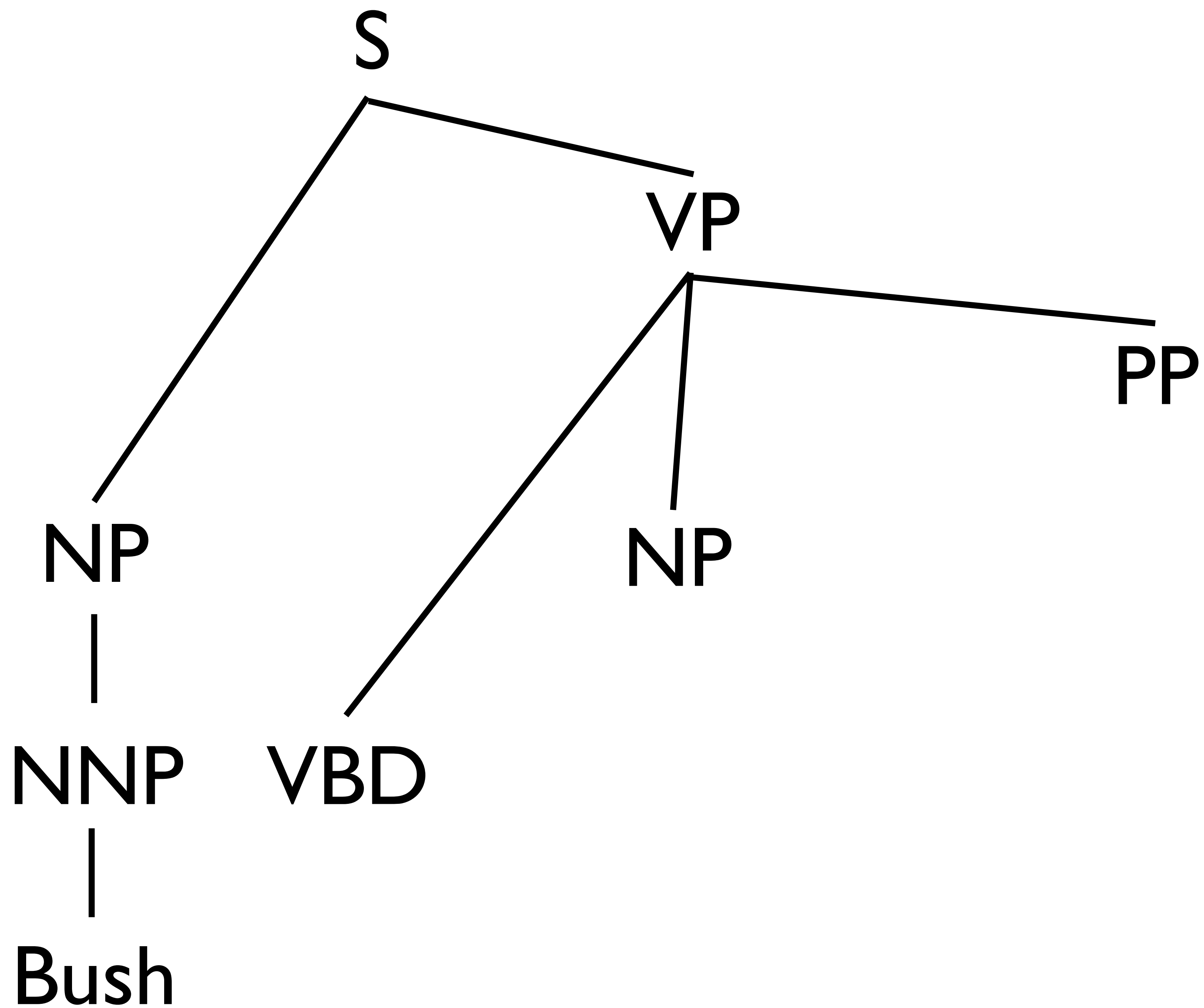
Graphical Representation



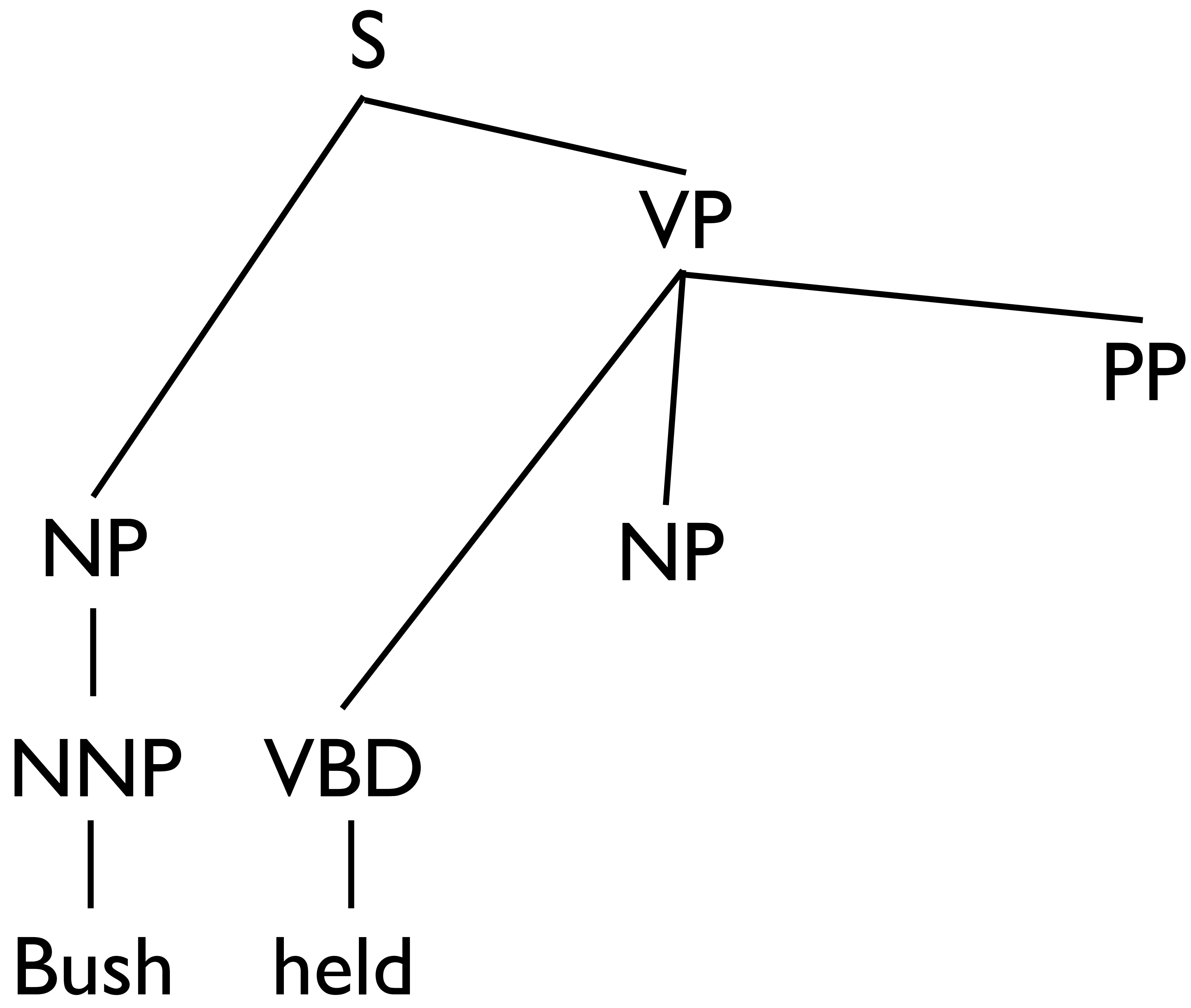
Graphical Representation



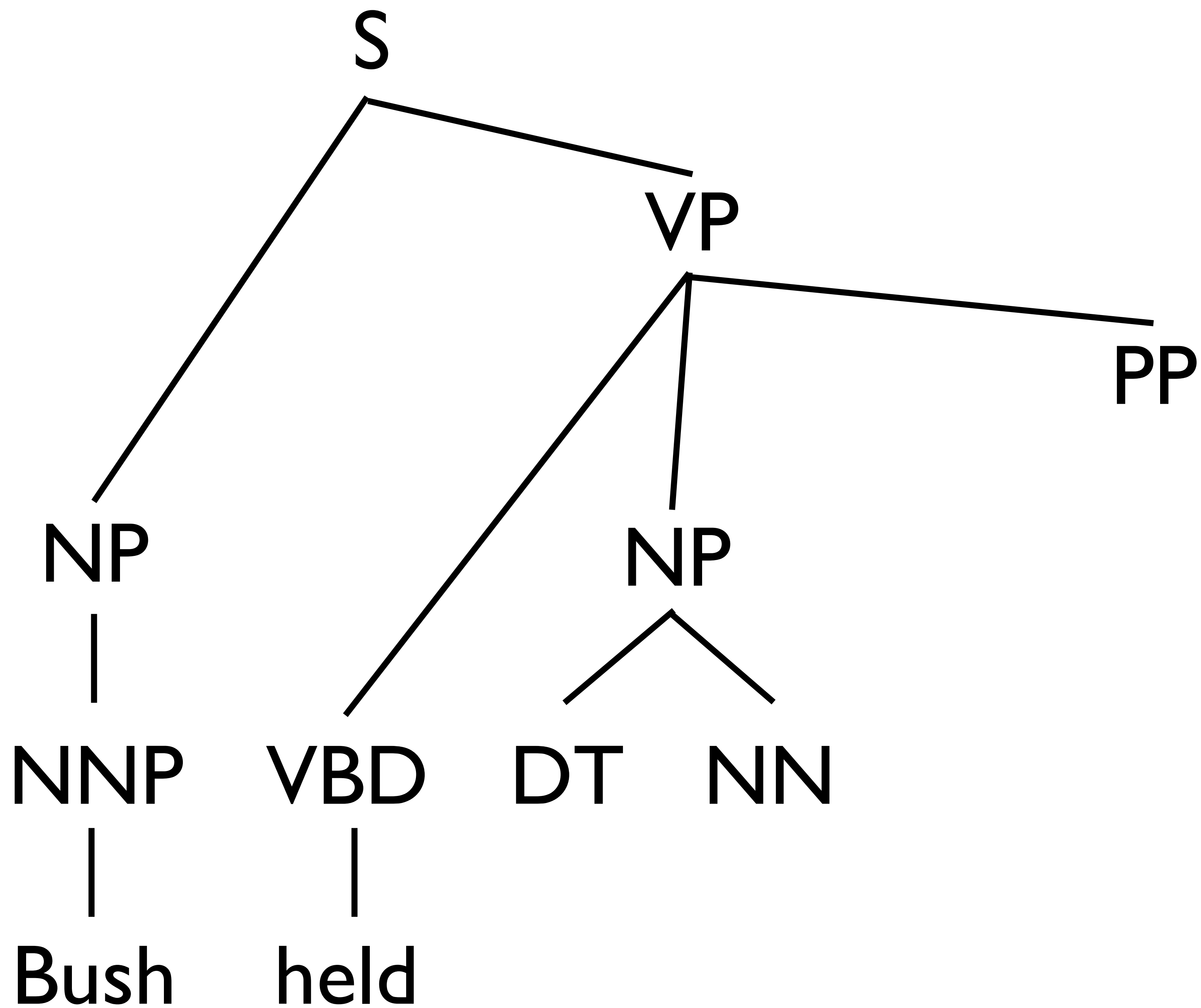
Graphical Representation



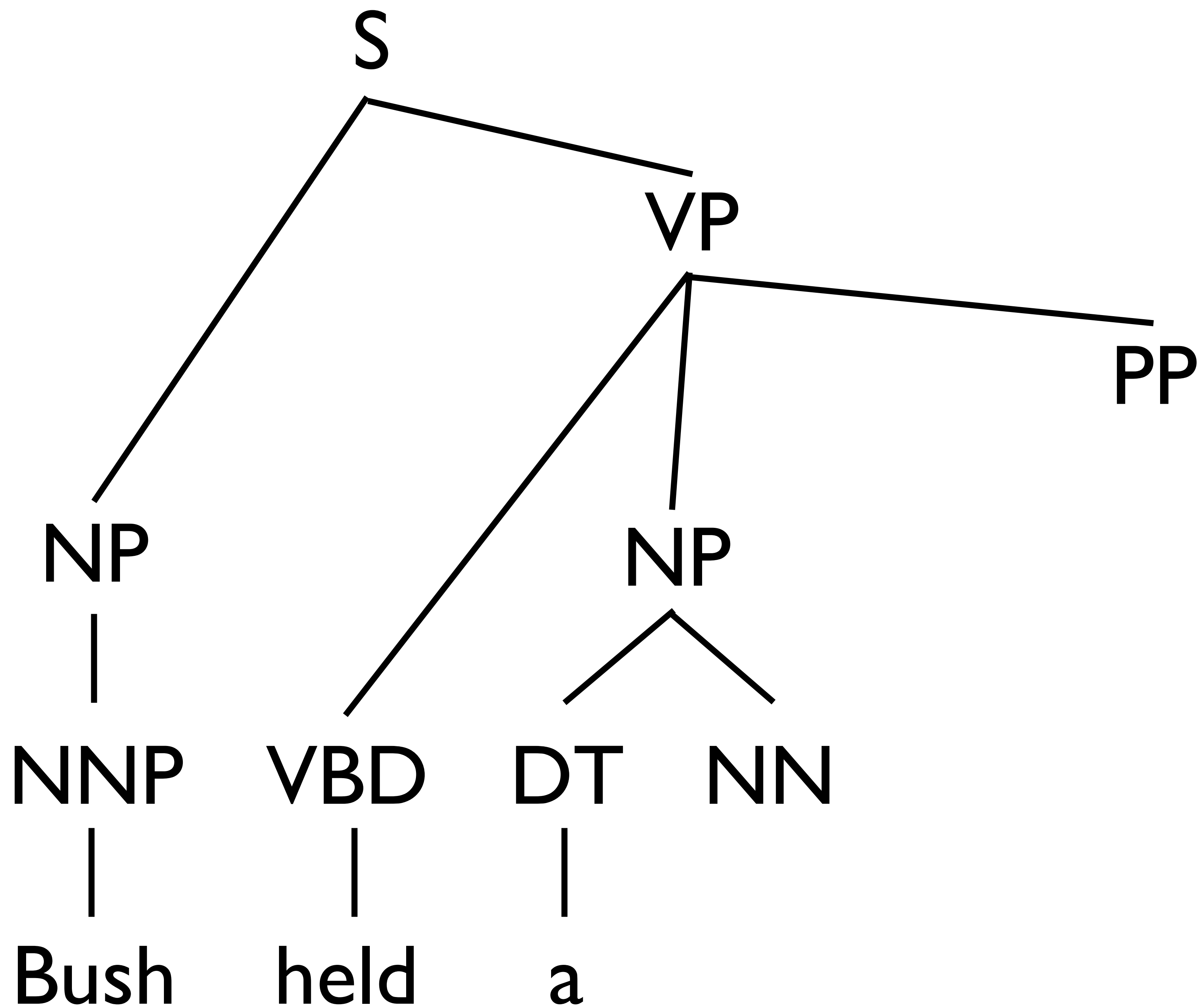
Graphical Representation



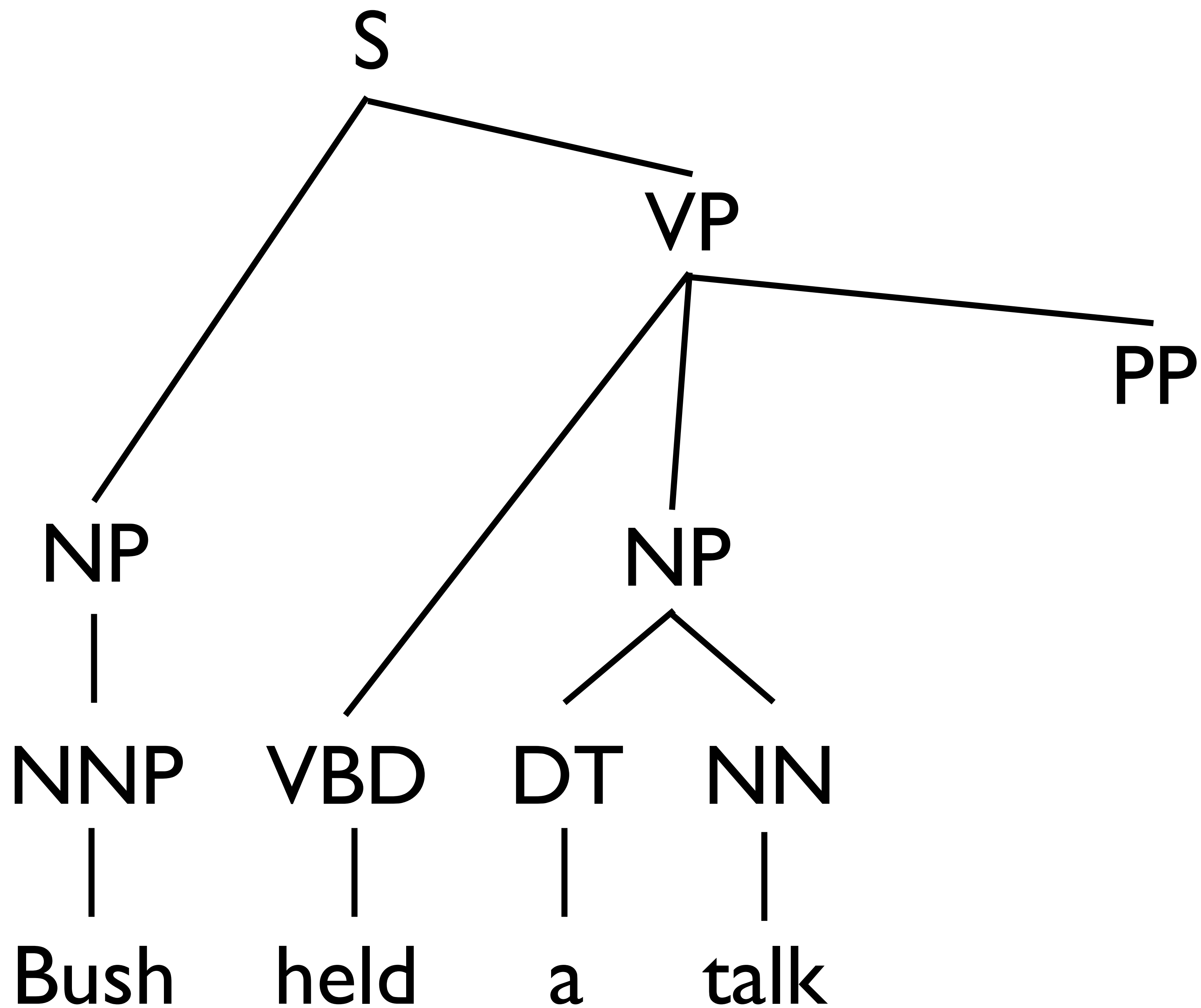
Graphical Representation



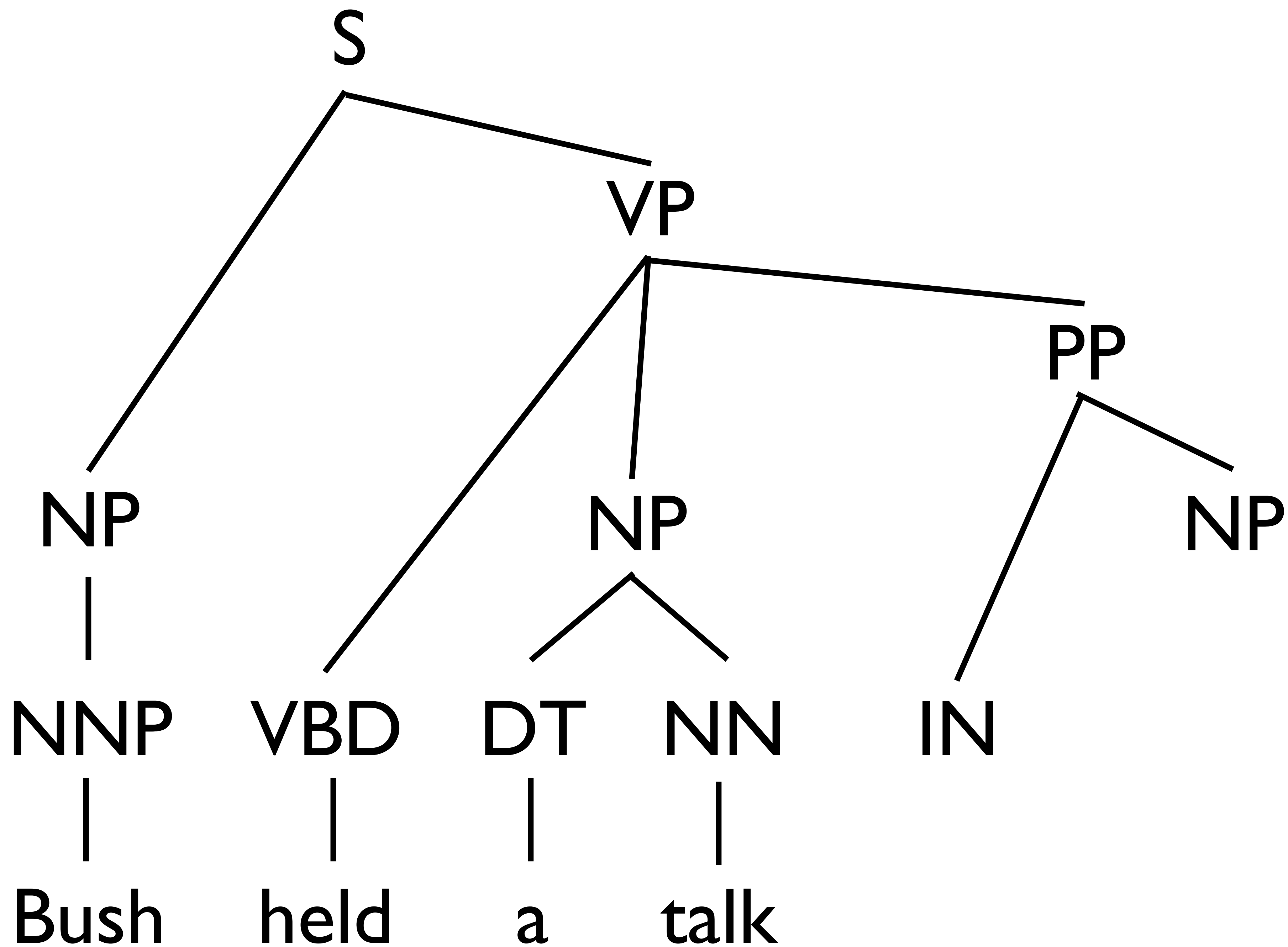
Graphical Representation



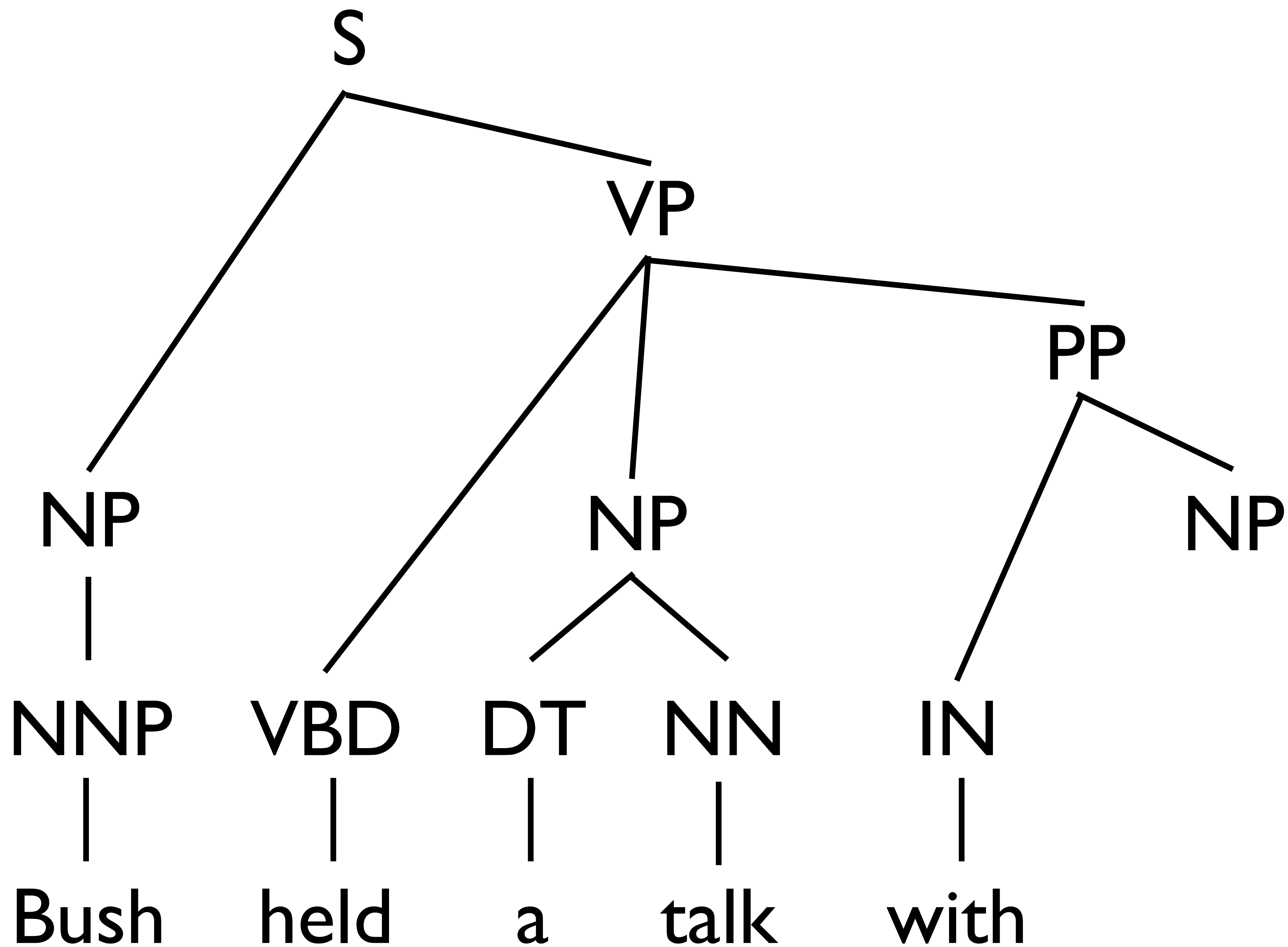
Graphical Representation



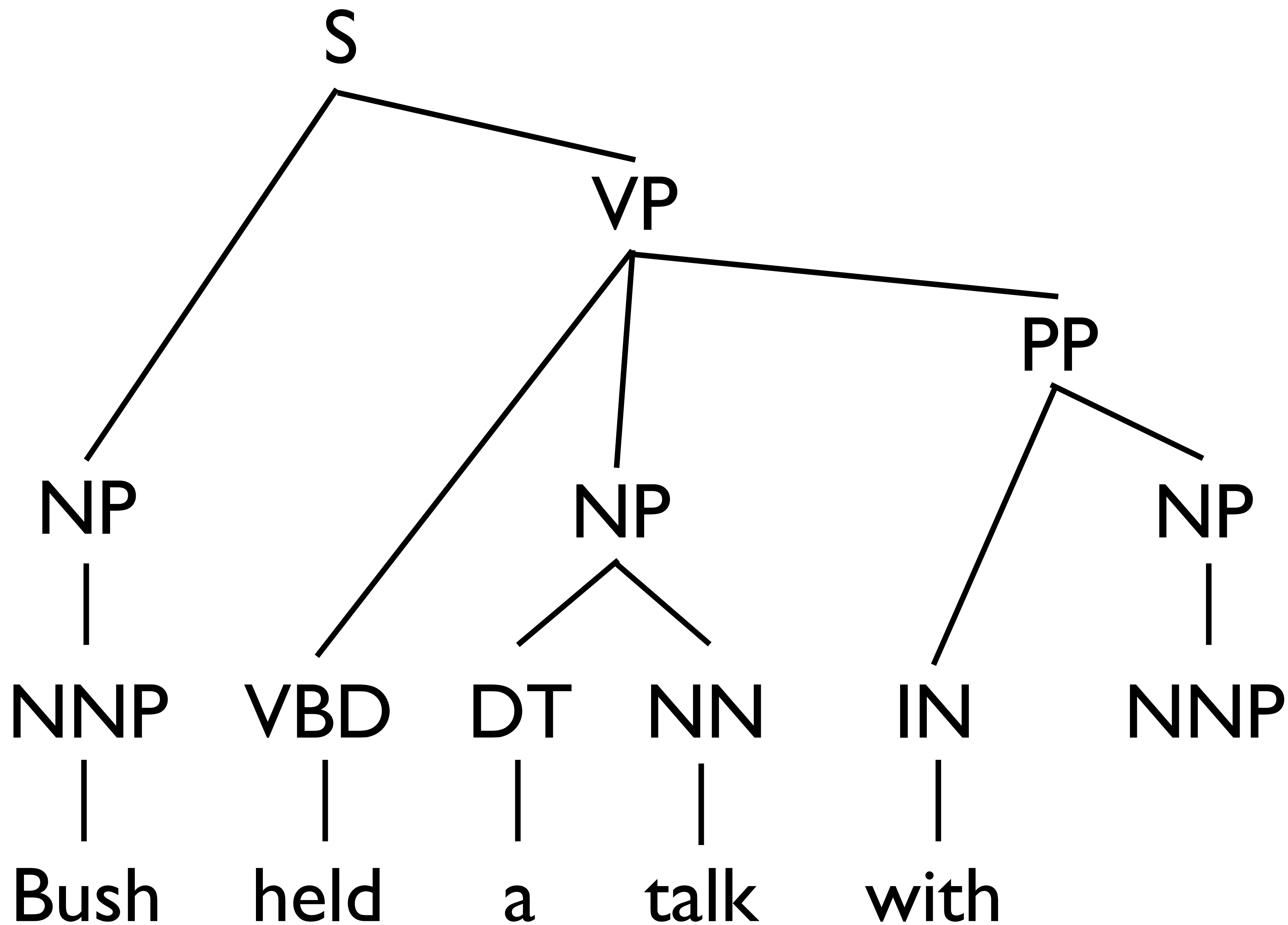
Graphical Representation



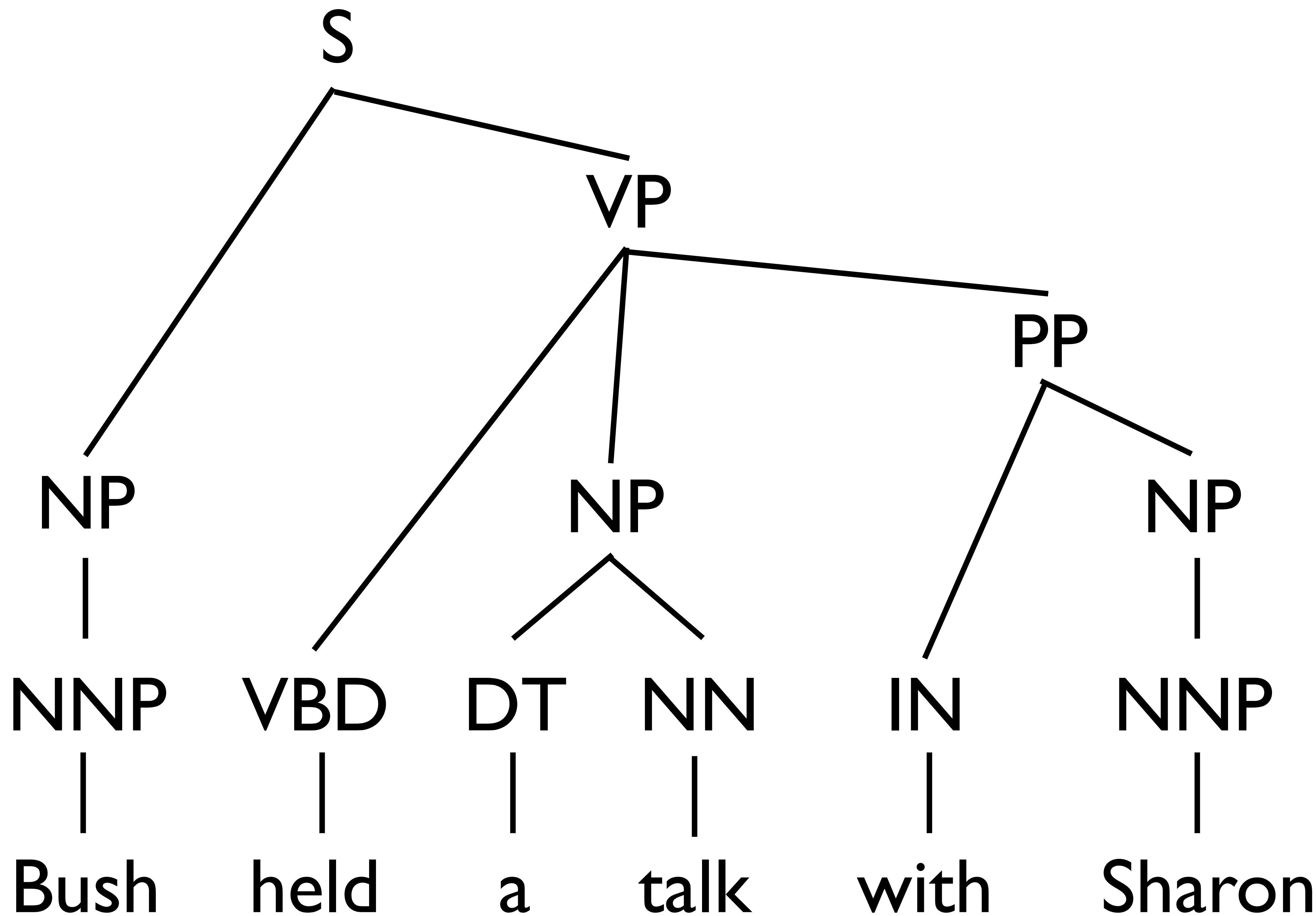
Graphical Representation



Graphical Representation



Graphical Representation



Synchronous Context-Free Grammar

- Synchronous Context-free grammar describes how two natural language sentences are generated simultaneously

(Chiang, 2005)

Synchronous Context-Free Grammar

- Synchronous Context-free grammar describes how two natural language sentences are generated simultaneously

lexical rules $\mathrm{NN} \rightarrow \langle \mathrm{bushi}, \mathrm{Bush} \rangle$

(Chiang, 2005)

Synchronous Context-Free Grammar

- Synchronous Context-free grammar describes how two natural language sentences are generated simultaneously

lexical rules $\mathrm{NN} \rightarrow \langle \mathrm{bushi}, \mathrm{Bush} \rangle$

syntactic rules $\mathrm{S} \rightarrow \langle \mathrm{NP}_1 \ \mathrm{VP}_2, \mathrm{NP}_1 \ \mathrm{VP}_2 \rangle$

(Chiang, 2005)

Synchronous Context-Free Grammar

- Synchronous Context-free grammar describes how two natural language sentences are generated simultaneously

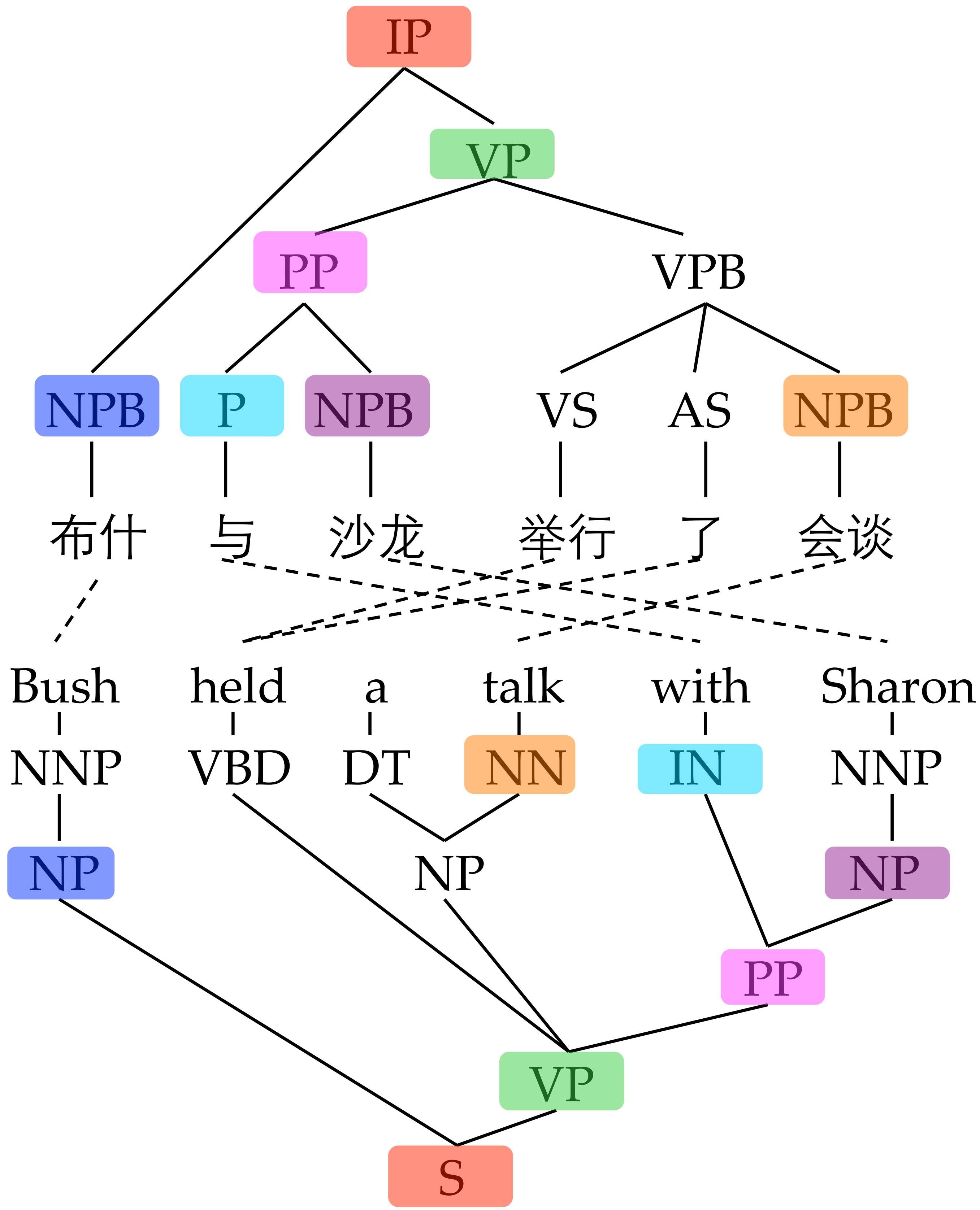
lexical rules $\mathrm{NN} \rightarrow \langle \mathrm{bushi}, \ \mathrm{Bush} \rangle$

syntactic rules $\mathrm{S} \rightarrow \langle \mathrm{NP}_1 \ \mathrm{VP}_2, \mathrm{NP}_1 \ \mathrm{VP}_2 \rangle$

Unfortunately, SCFG suffers from the non-isomorphism problem

(Chiang, 2005)

Non-Isomorphism



Syntax-based MT

SCFGs without linguistic syntax

inverted transduction grammar

hierarchical phrase-based model

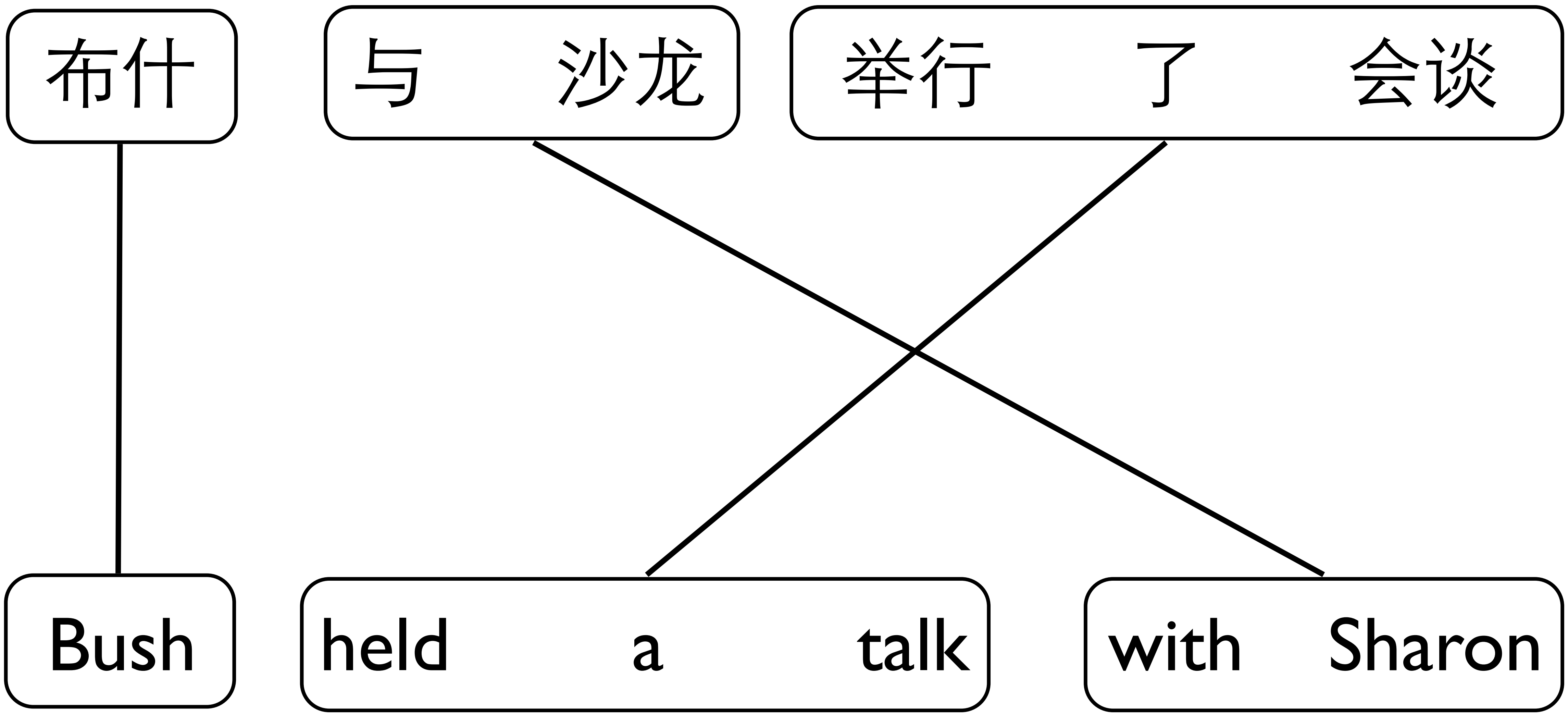
STSGs with linguistic syntax

string-to-tree

tree-to-string

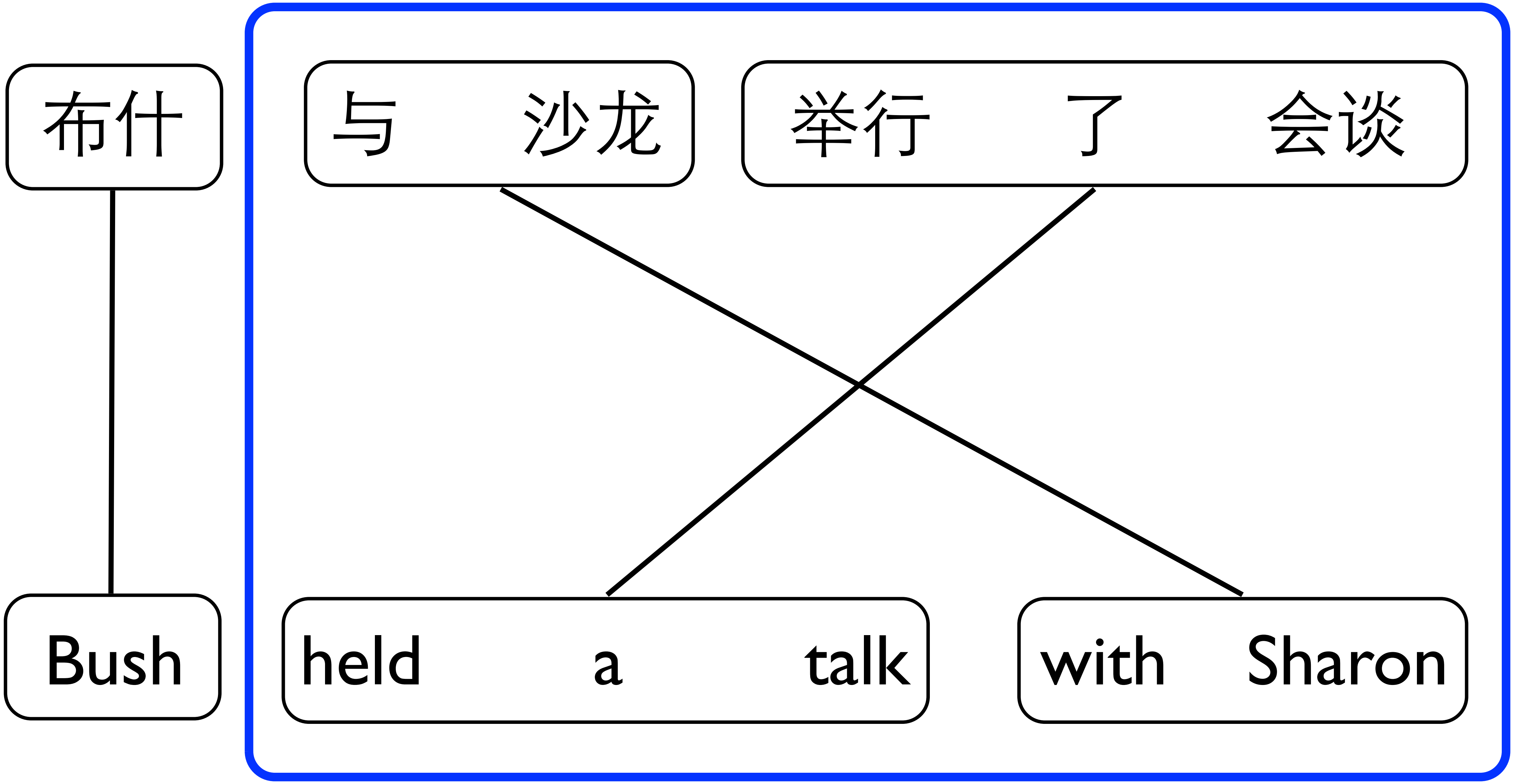
tree-to-tree

Block Merging



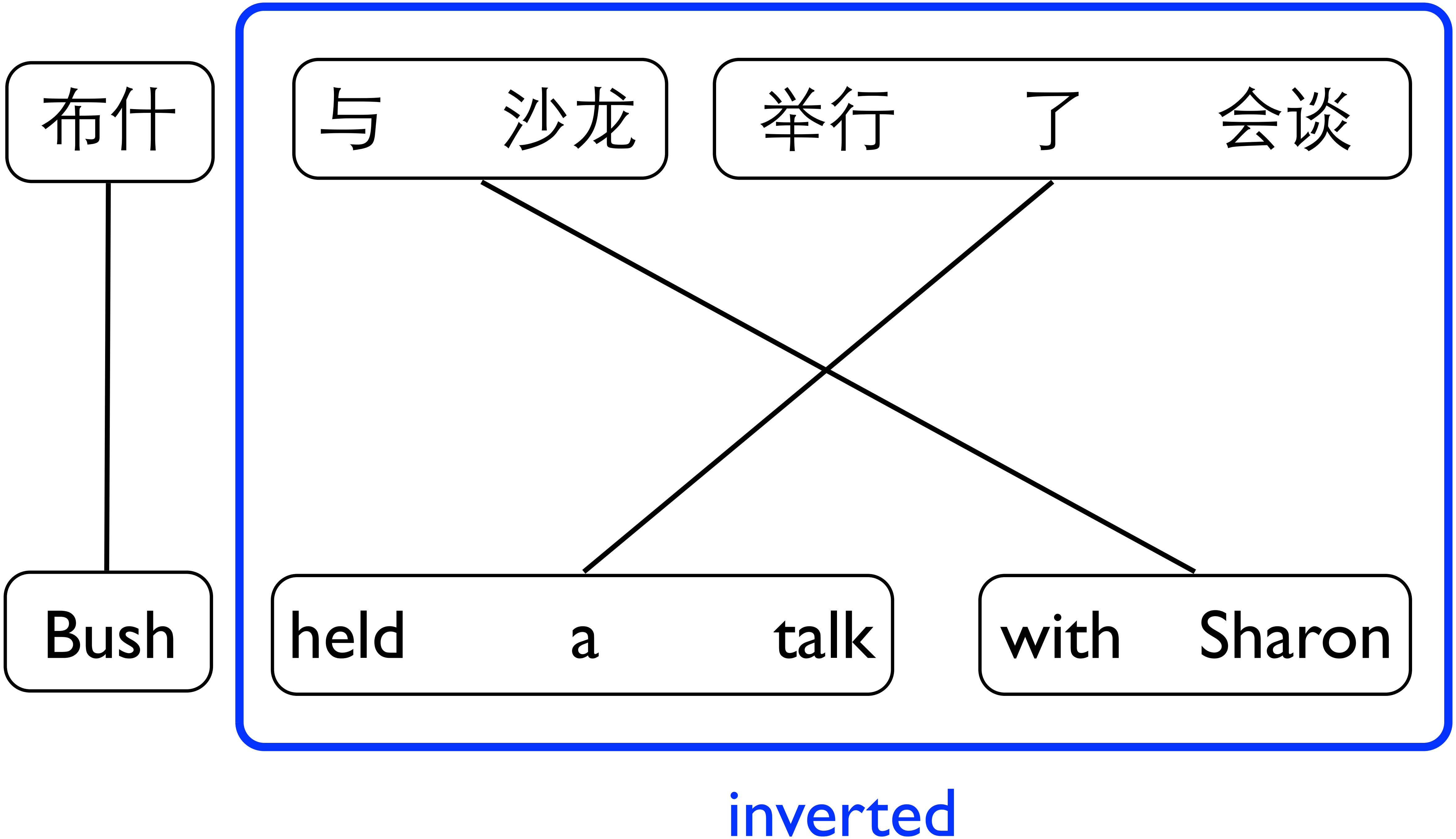
(Wu, 1997; Xiong et al., 2006)

Block Merging



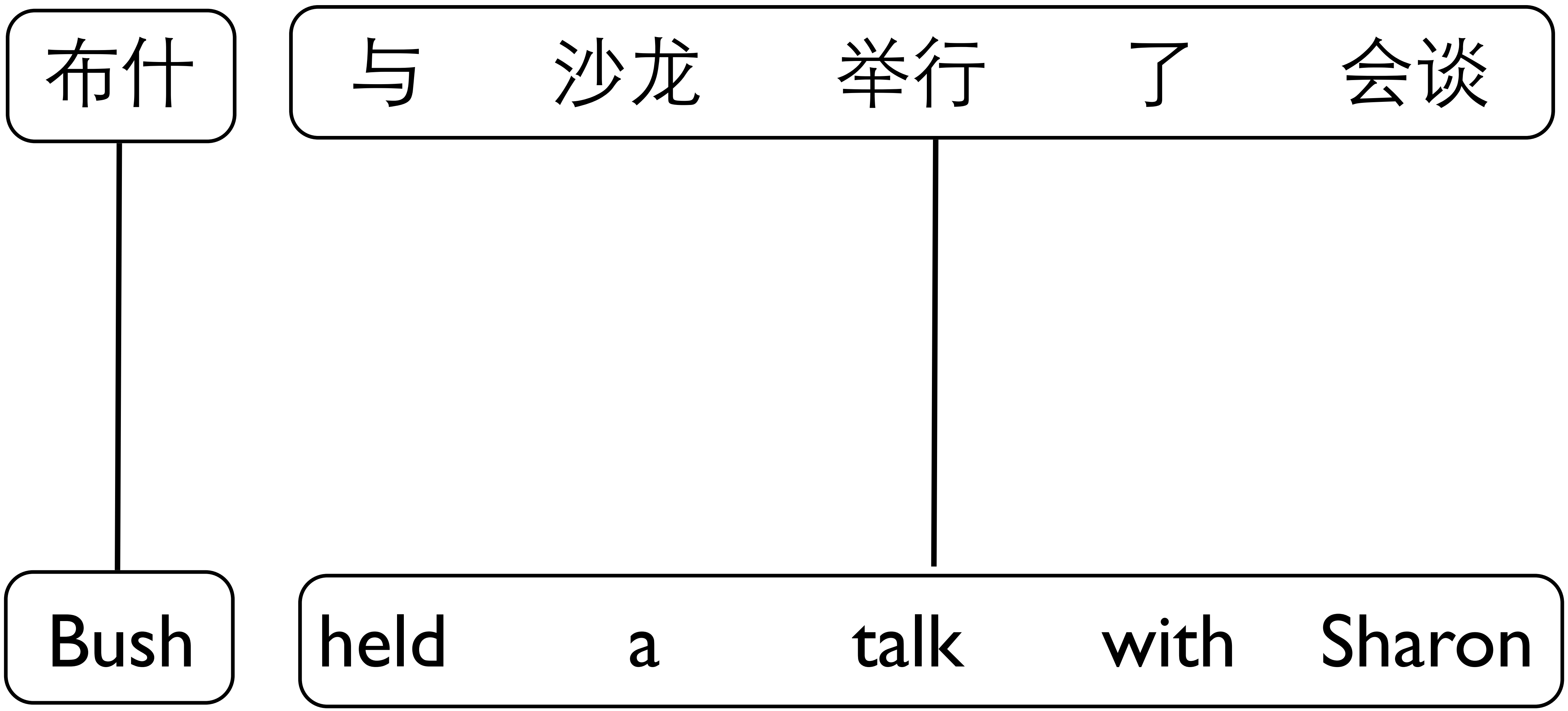
(Wu, 1997; Xiong et al., 2006)

Block Merging



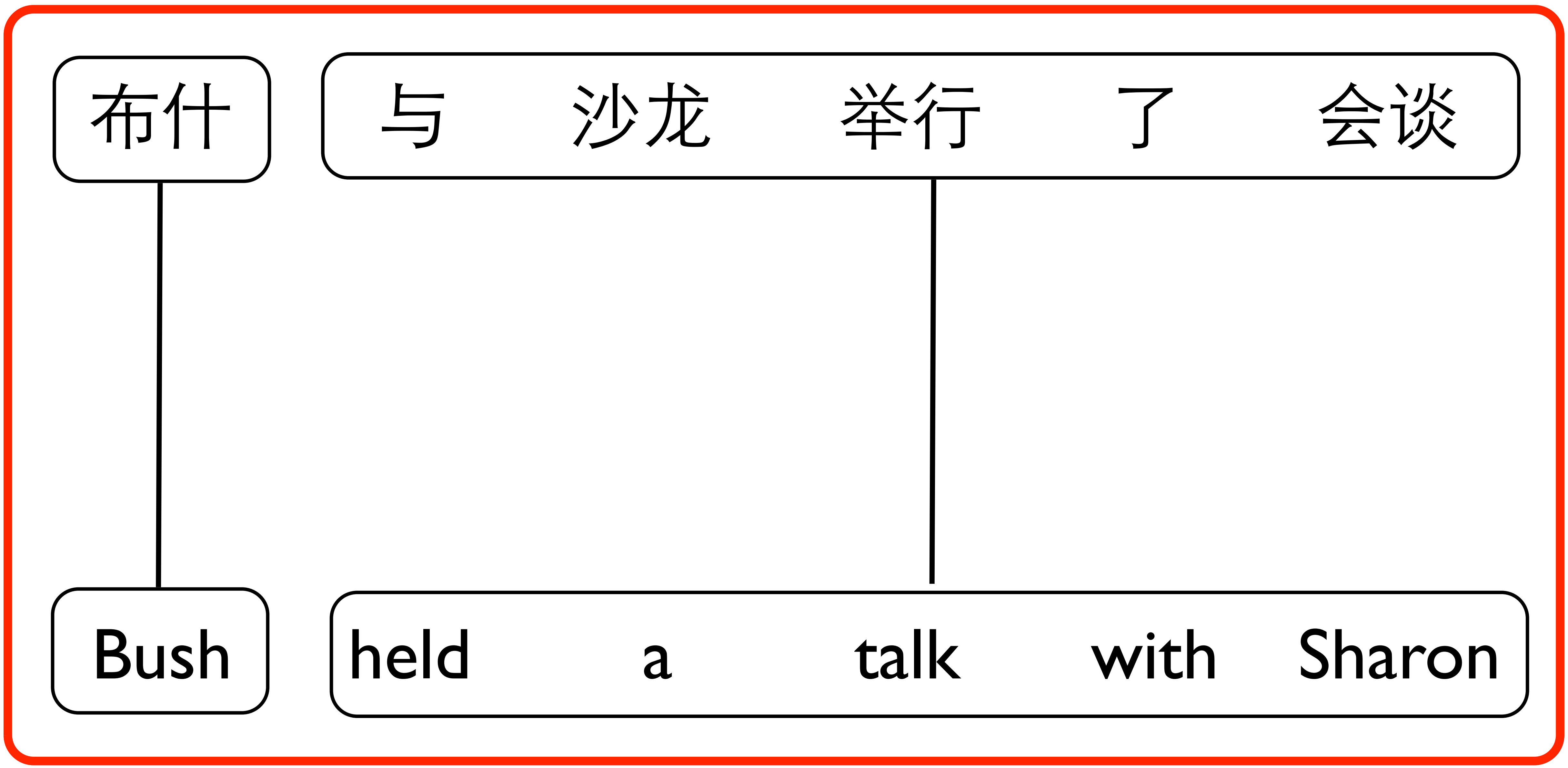
(Wu, 1997; Xiong et al., 2006)

Block Merging



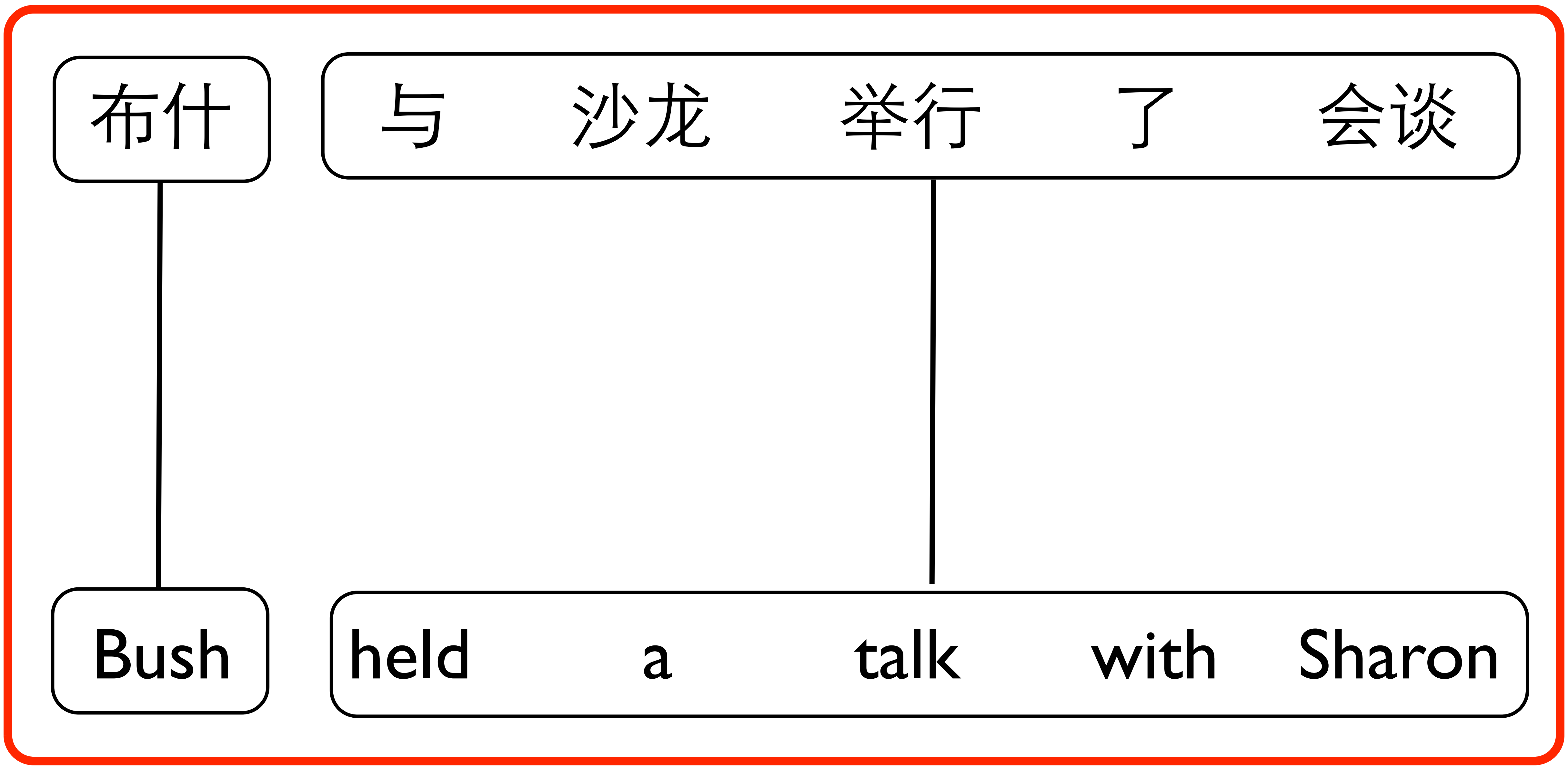
(Wu, 1997; Xiong et al., 2006)

Block Merging



(Wu, 1997; Xiong et al., 2006)

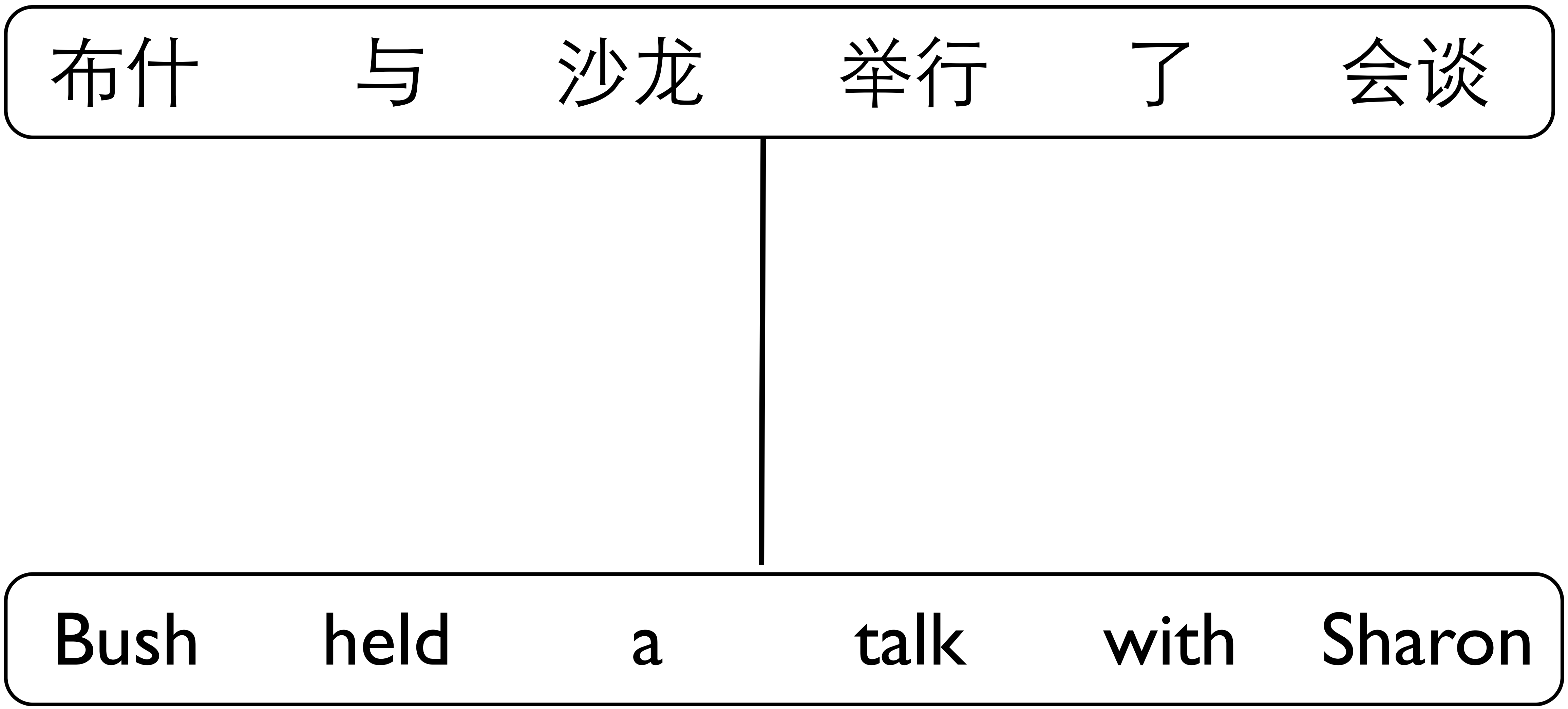
Block Merging



straight

(Wu, 1997; Xiong et al., 2006)

Block Merging



(Wu, 1997; Xiong et al., 2006)

Block Merging

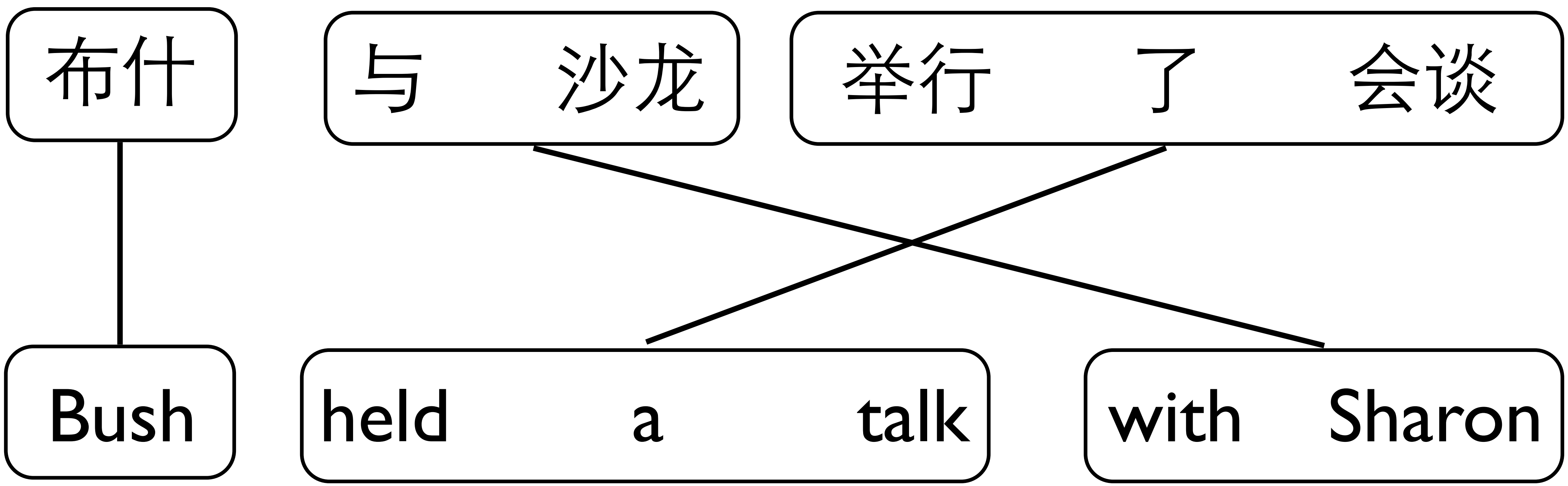
布什 与 沙龙 举行 了 会谈

Bush held a talk with Sharon

describe reordering using two operators

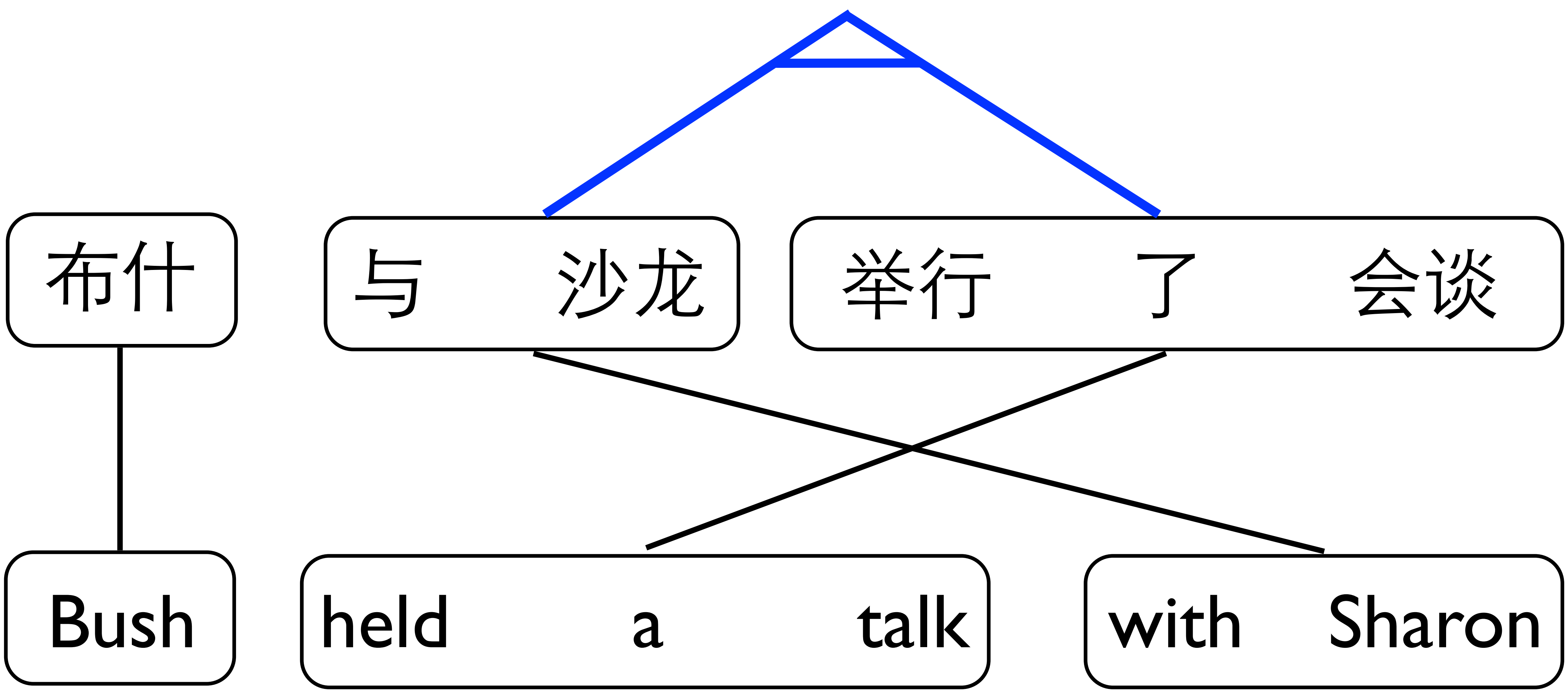
(Wu, 1997; Xiong et al., 2006)

Block Merging



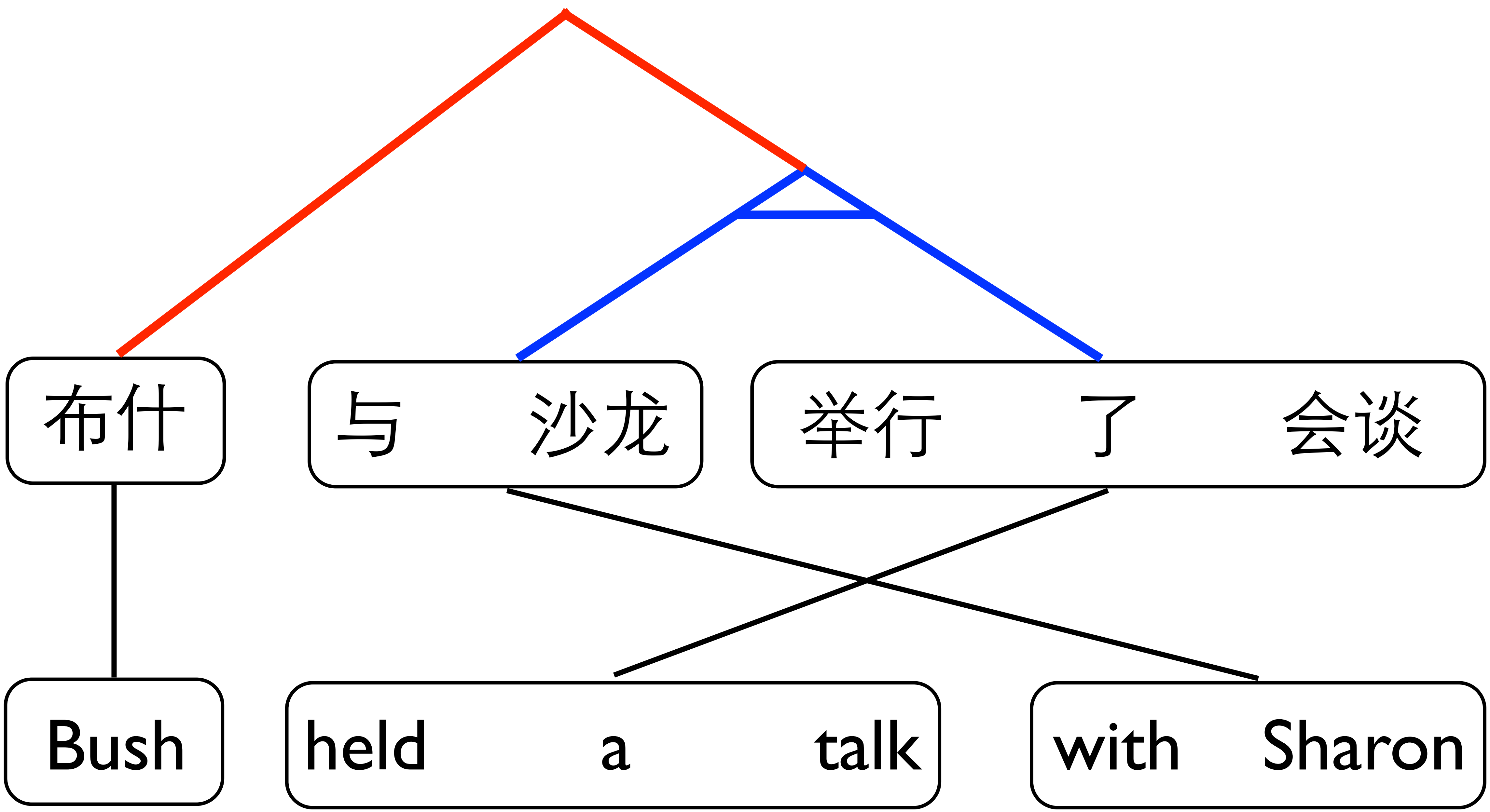
(Wu, 1997; Xiong et al., 2006)

Block Merging



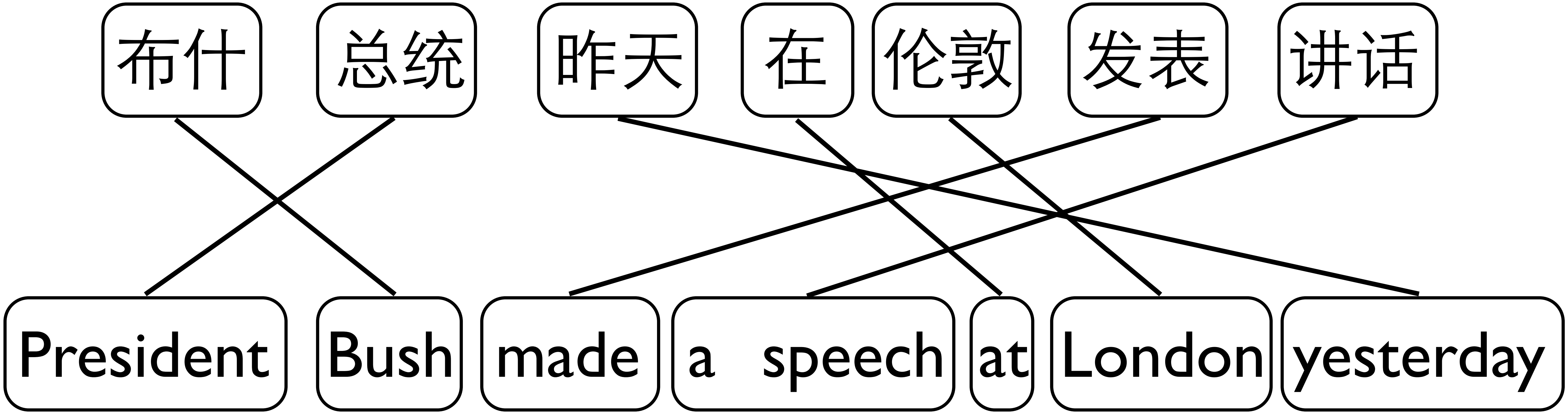
(Wu, 1997; Xiong et al., 2006)

Block Merging



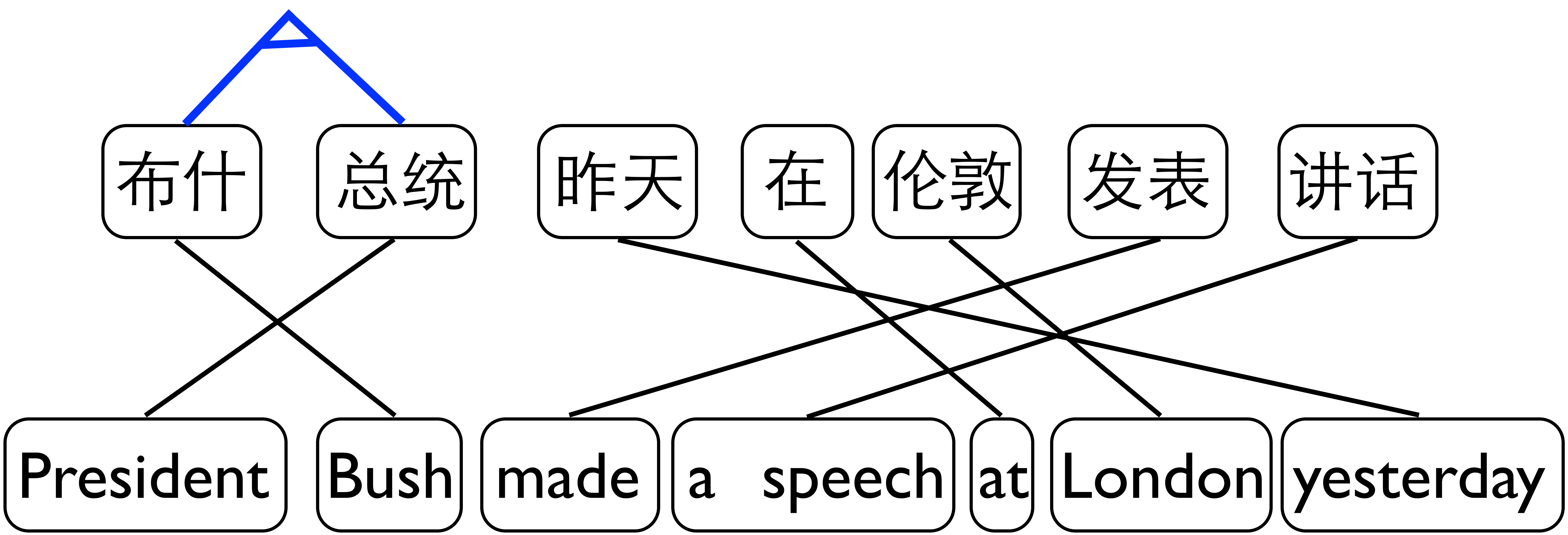
(Wu, 1997; Xiong et al., 2006)

Block Merging



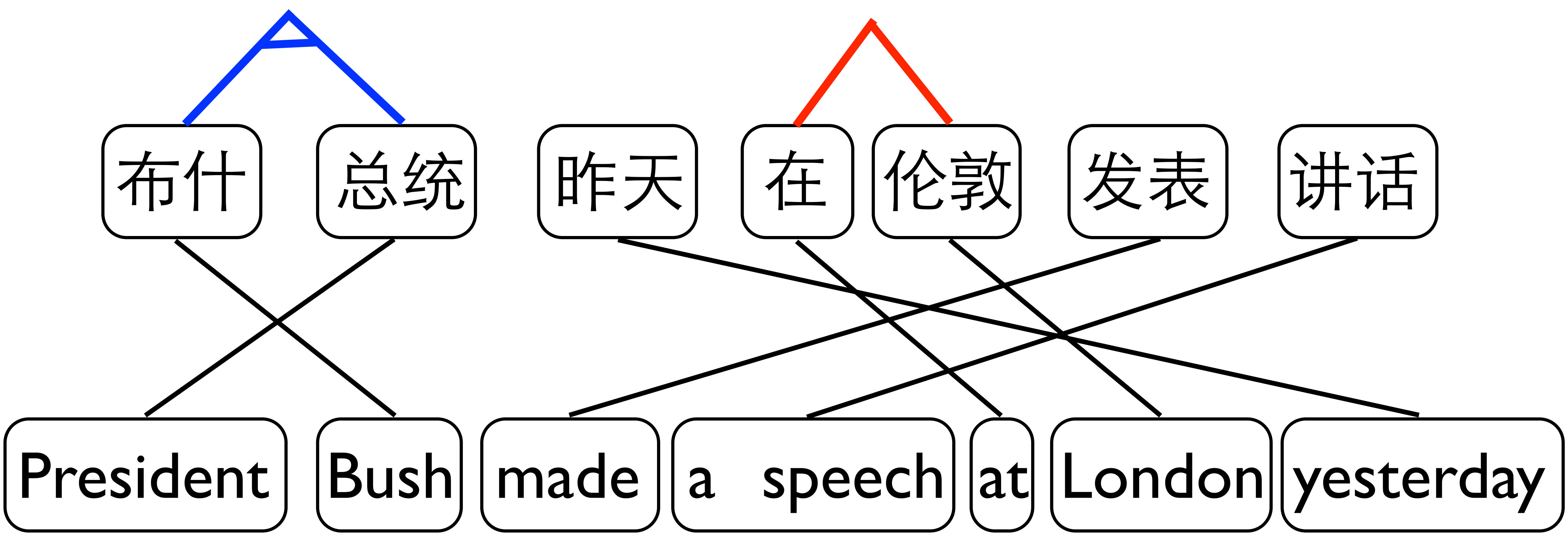
(Wu, 1997; Xiong et al., 2006)

Block Merging



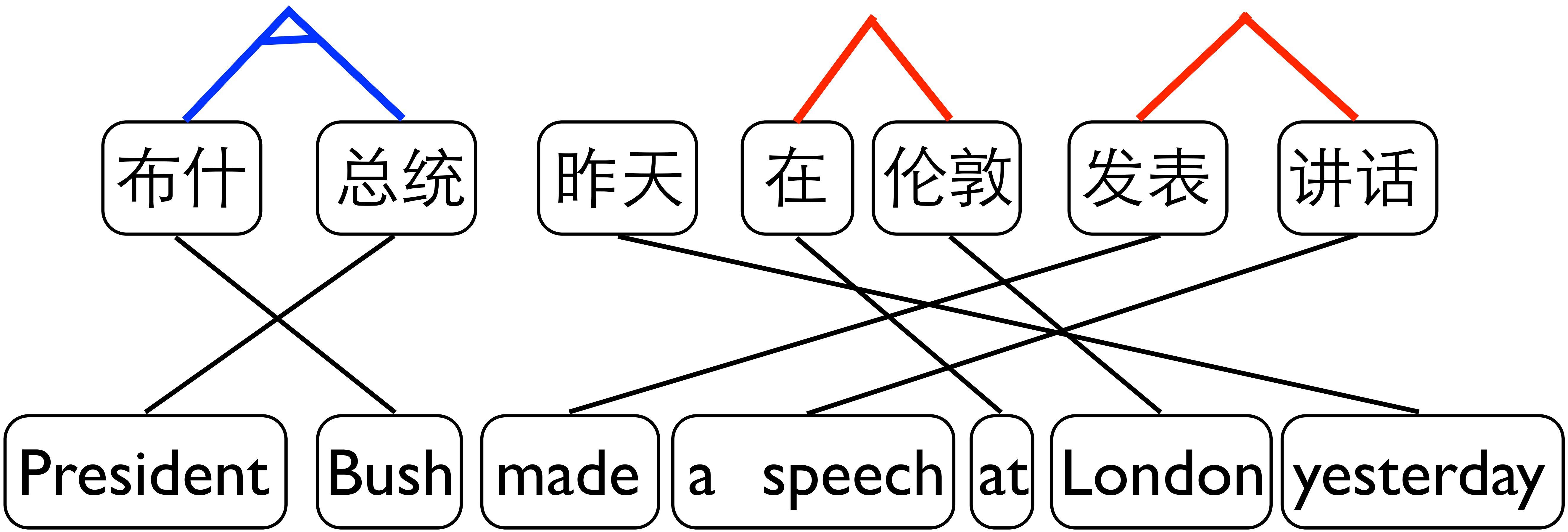
(Wu, 1997; Xiong et al., 2006)

Block Merging



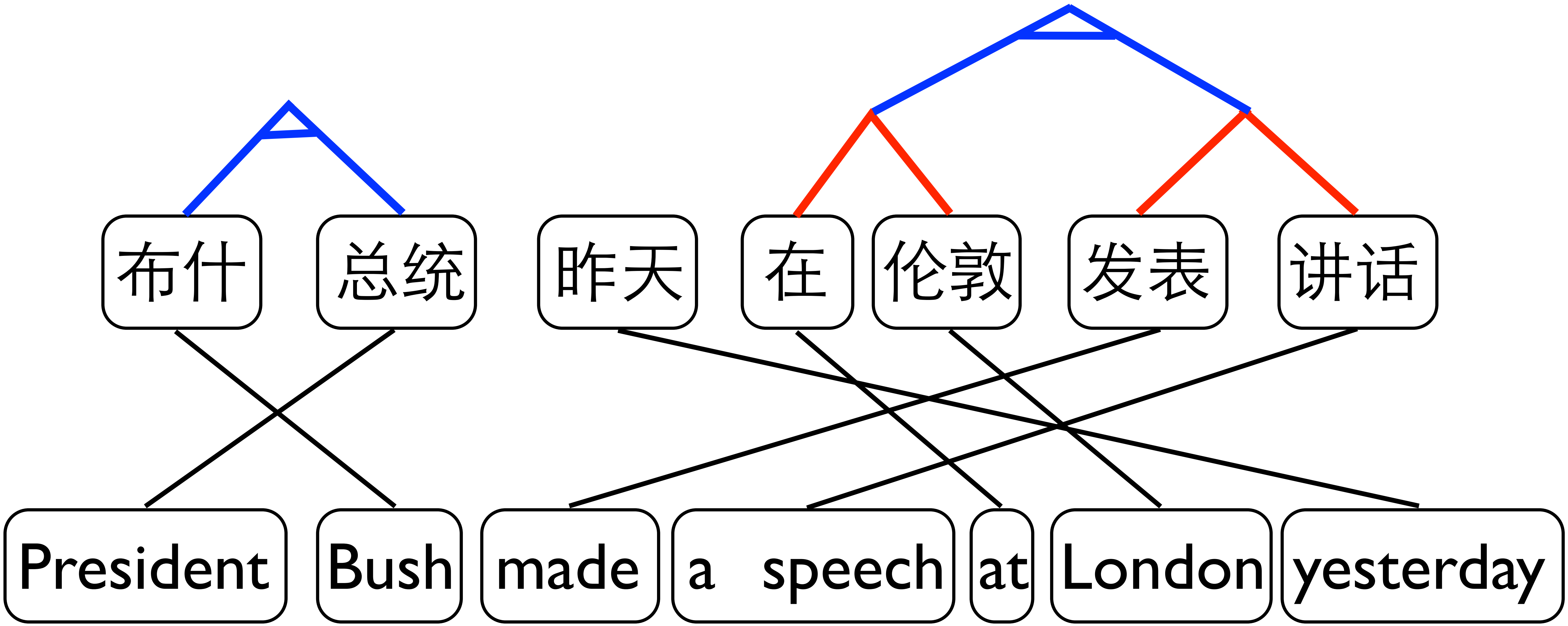
(Wu, 1997; Xiong et al., 2006)

Block Merging



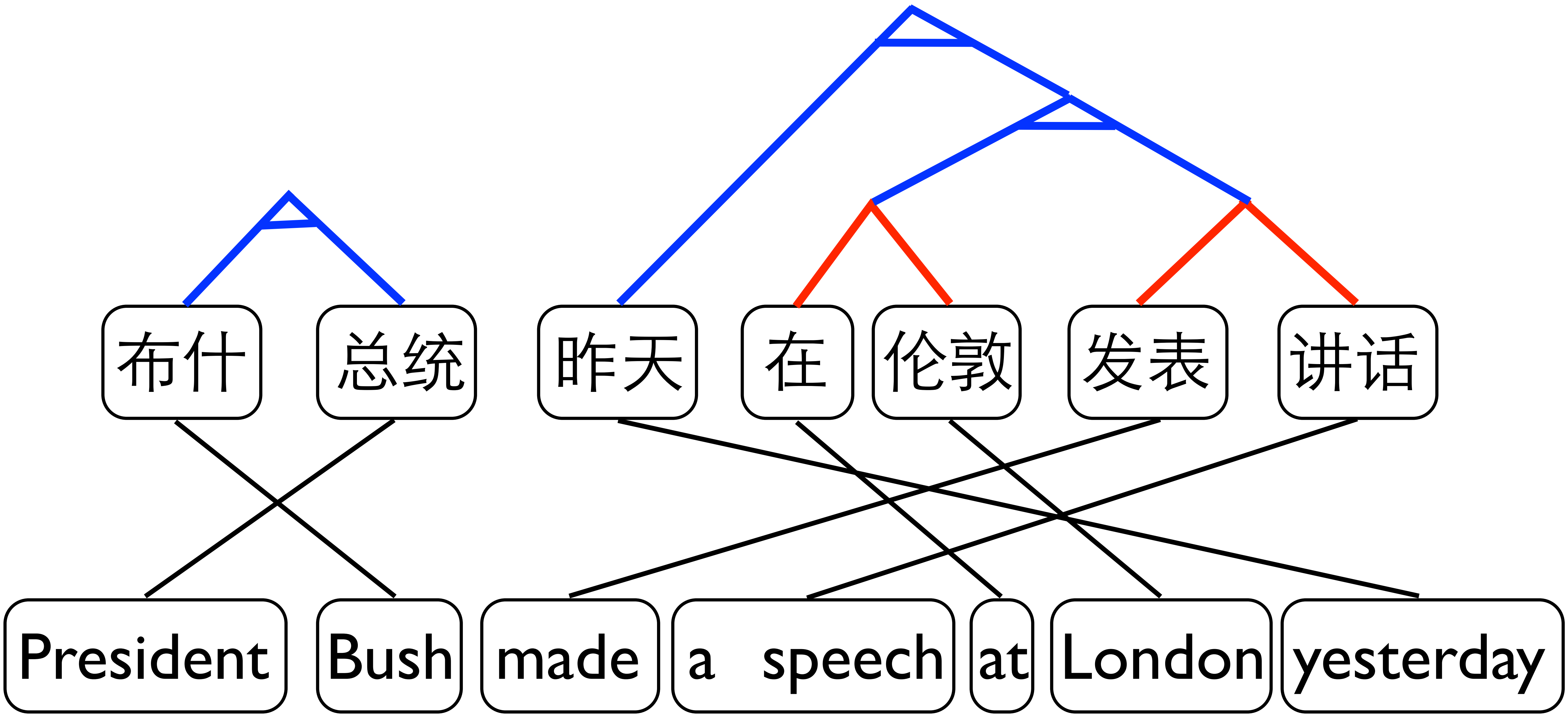
(Wu, 1997; Xiong et al., 2006)

Block Merging



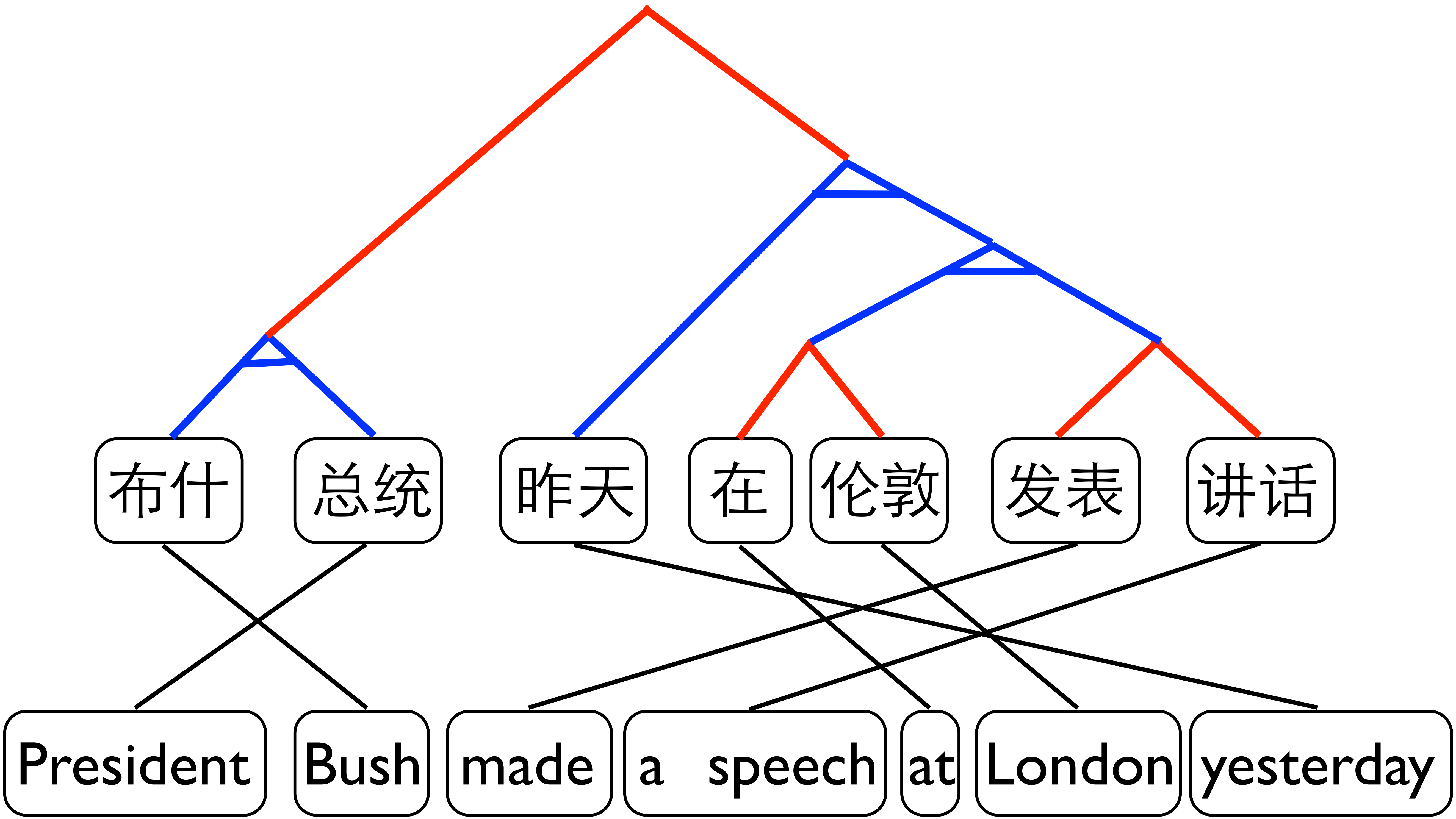
(Wu, 1997; Xiong et al., 2006)

Block Merging



(Wu, 1997; Xiong et al., 2006)

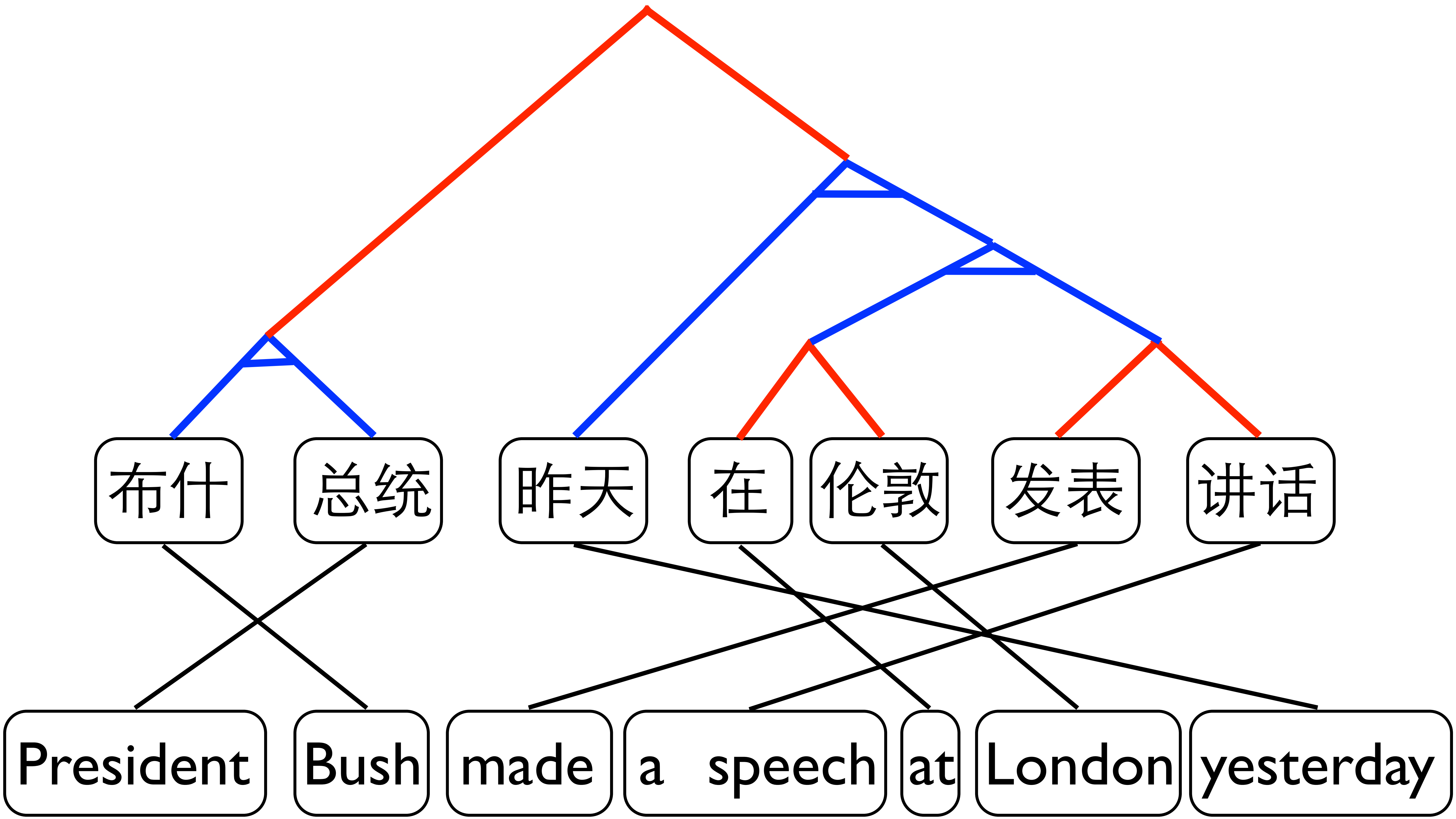
Block Merging



(Wu, 1997; Xiong et al., 2006)

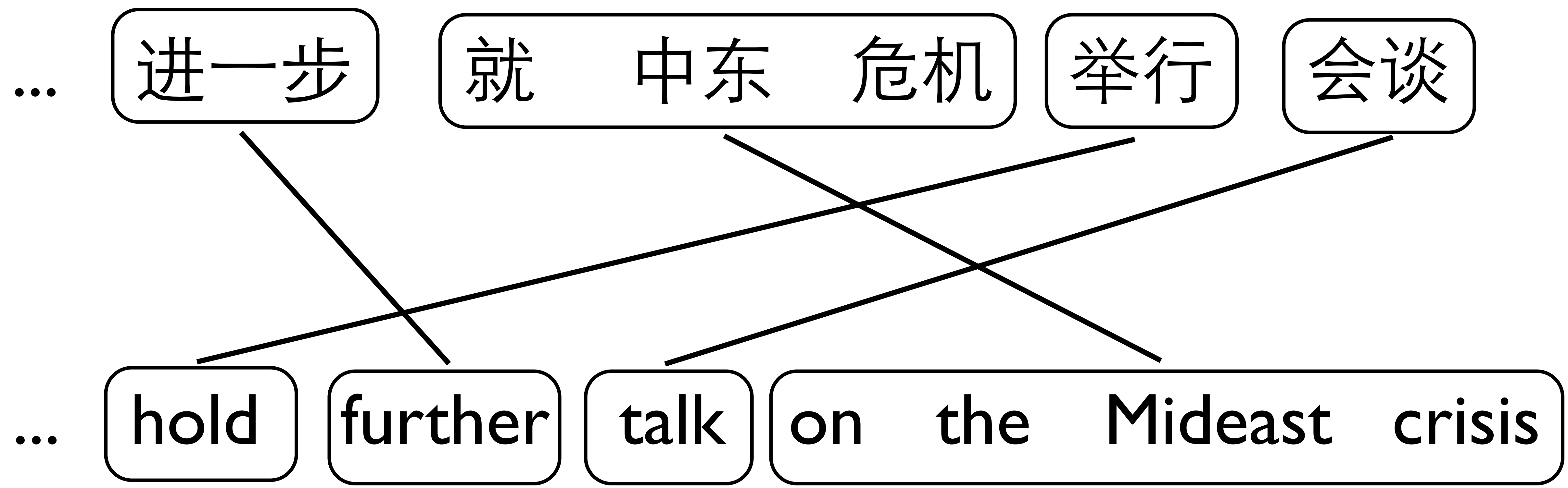
Block Merging

Q: can you find a counter example?



(Wu, 1997; Xiong et al., 2006)

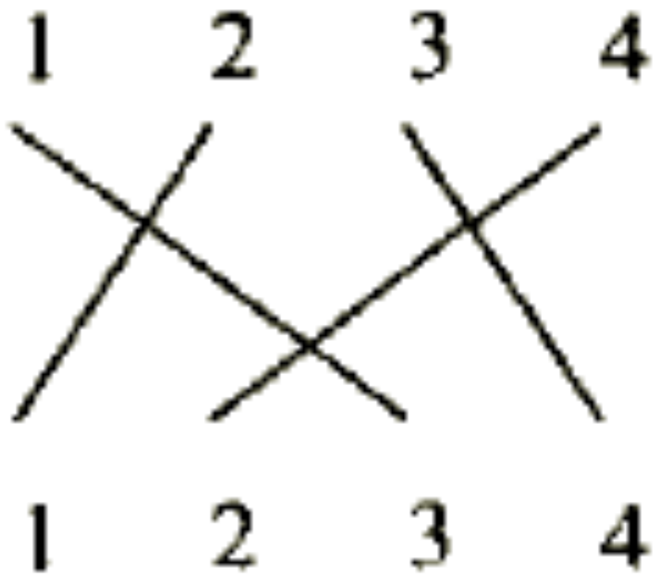
Counter Example

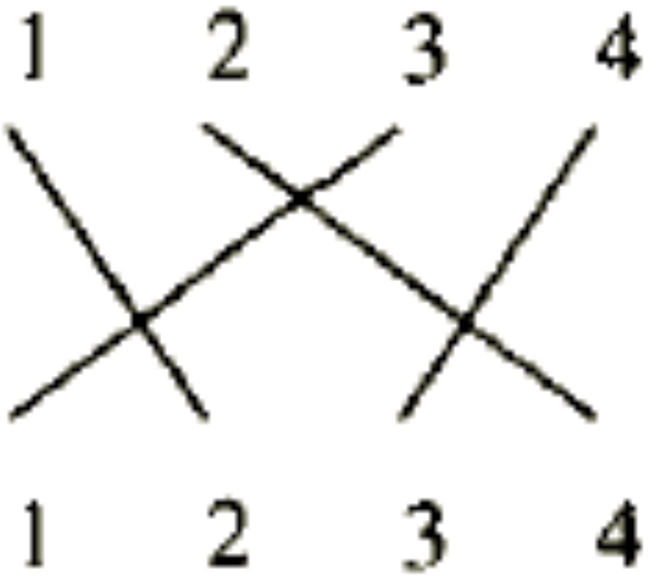


(Huang et al., 2009)

Expressiveness of ITG

“inside-out”





(Wu, 1997)

Inverted Transduction Grammar

- Inverted Transduction Grammar explains how two natural language sentences are generated synchronously using two block-merging operators

(Wu, 1997; Xiong et al., 2006)

Inverted Transduction Grammar

- Inverted Transduction Grammar explains how two natural language sentences are generated synchronously using two block-merging operators

$$X \rightarrow f/e$$

(Wu, 1997; Xiong et al., 2006)

Inverted Transduction Grammar

- Inverted Transduction Grammar explains how two natural language sentences are generated synchronously using two block-merging operators

lexical rules

$$X \rightarrow f/e$$

(Wu, 1997; Xiong et al., 2006)

Inverted Transduction Grammar

- Inverted Transduction Grammar explains how two natural language sentences are generated synchronously using two block-merging operators

lexical rules	syntactic rules
$X \rightarrow f/e$	$X \rightarrow [X^1, X^2]$
	$X \rightarrow \langle X^1, X^2 \rangle$

(Wu, 1997; Xiong et al., 2006)

Inverted Transduction Grammar

- Inverted Transduction Grammar explains how two natural language sentences are generated synchronously using two block-merging operators

lexical rules	syntactic rules	
$X \rightarrow f/e$	$X \rightarrow [X^1, X^2]$	straight
	$X \rightarrow \langle X^1, X^2 \rangle$	

(Wu, 1997; Xiong et al., 2006)

Inverted Transduction Grammar

- Inverted Transduction Grammar explains how two natural language sentences are generated synchronously using two block-merging operators

lexical rules

$$X \to f/e$$

syntactic rules

$$X \to [X^1, X^2] \quad \text{straight}$$

$$X \to \langle X^1, X^2 \rangle \quad \text{inverted}$$

(Wu, 1997; Xiong et al., 2006)

CKY Parsing

布什 与 沙龙 举行 了 会谈

(Wu, 1997; Xiong et al., 2006)

CKY Parsing

布什　与　沙龙　举行　了　会谈

Bush

(Wu, 1997; Xiong et al., 2006)

CKY Parsing

X1

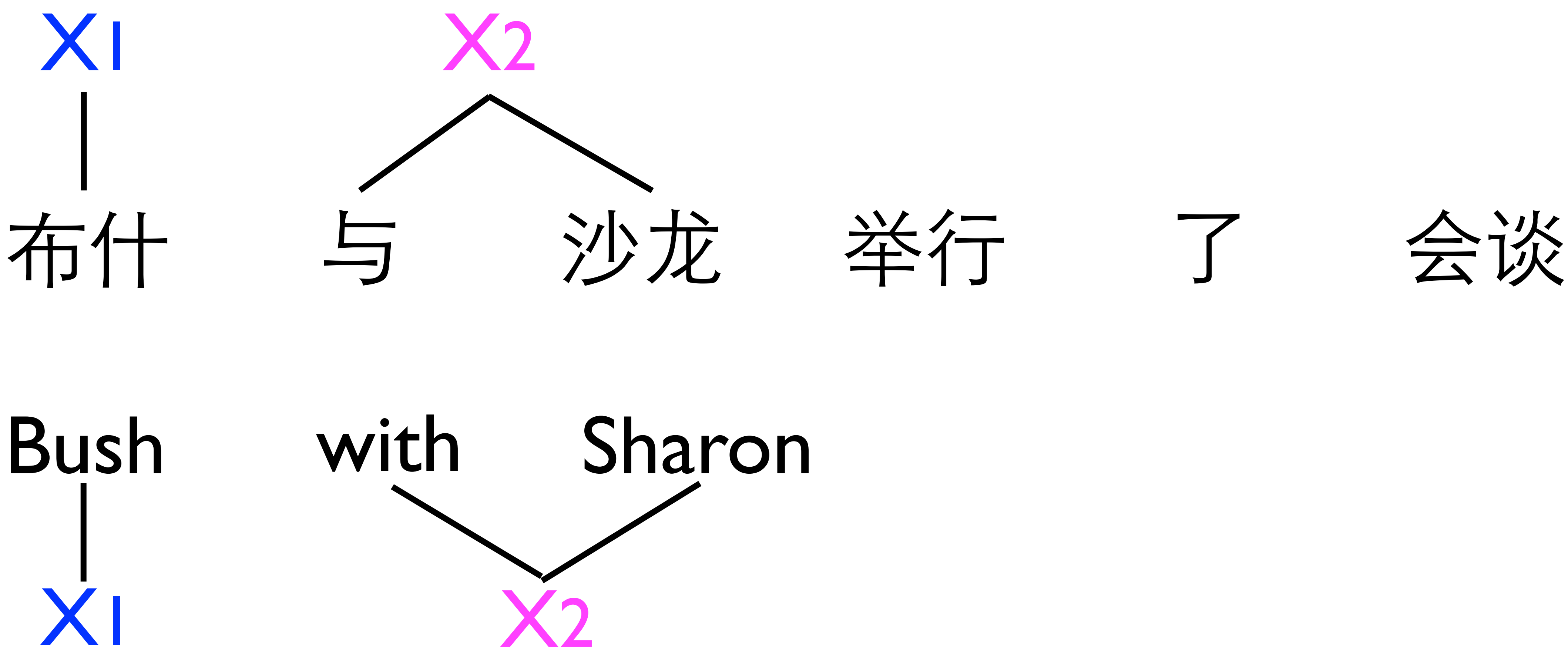
布什　与　沙龙　举行　了　会谈

Bush

X1

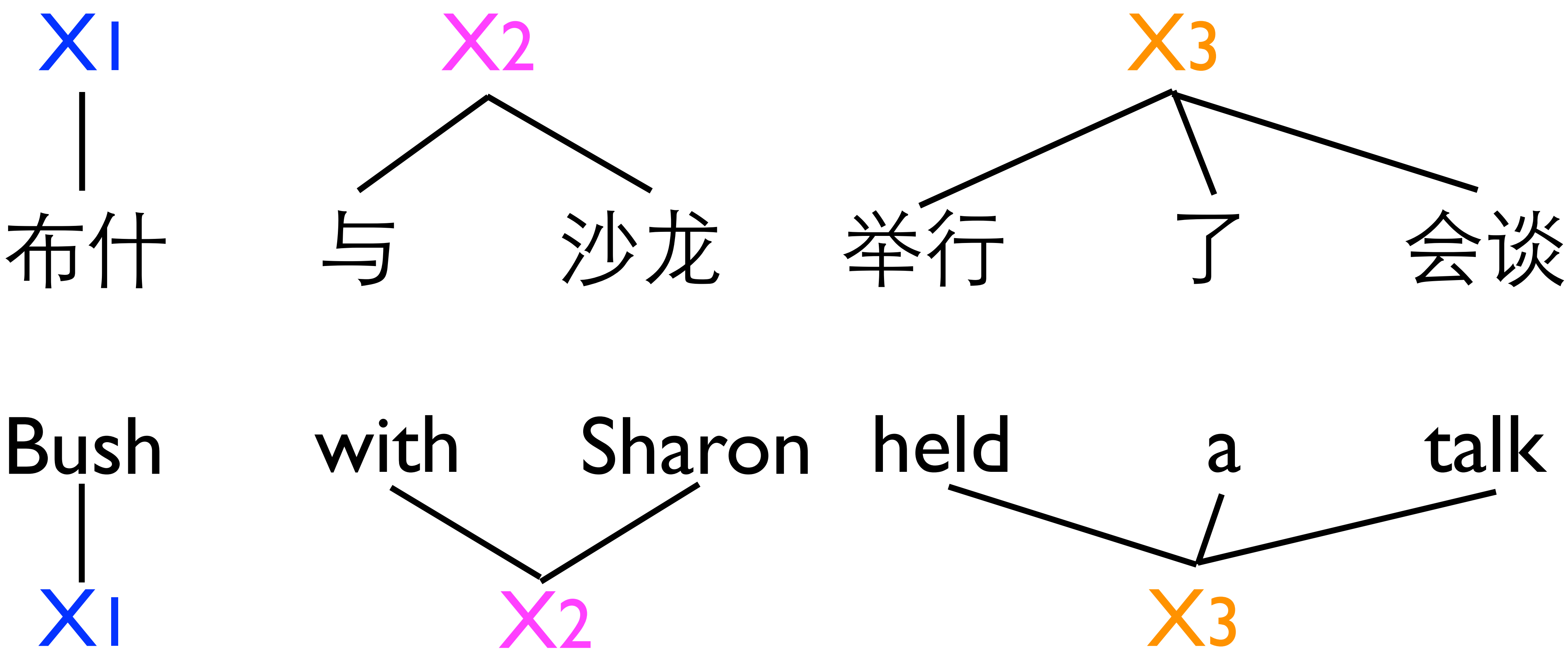
(Wu, 1997; Xiong et al., 2006)

CKY Parsing



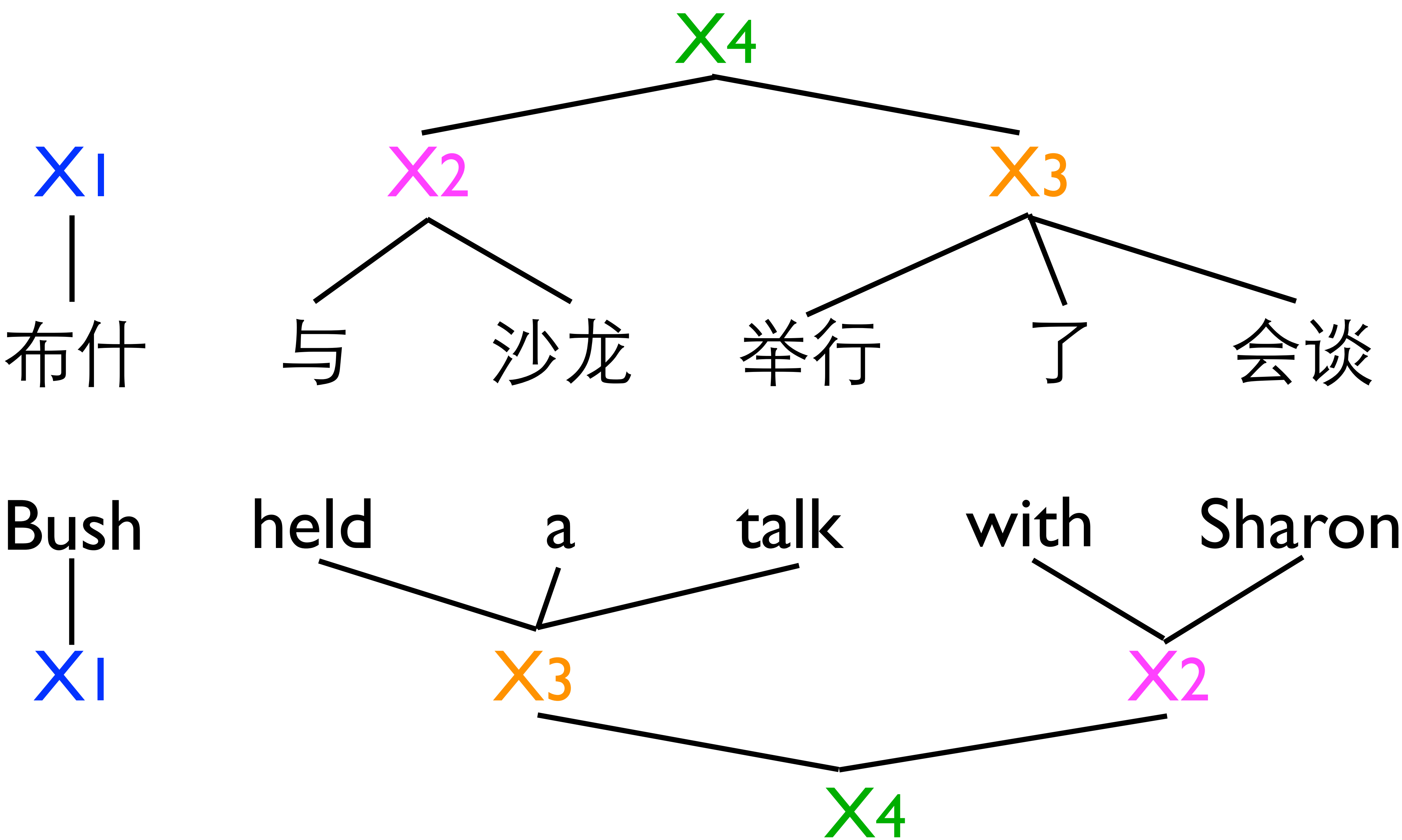
(Wu, 1997; Xiong et al., 2006)

CKY Parsing



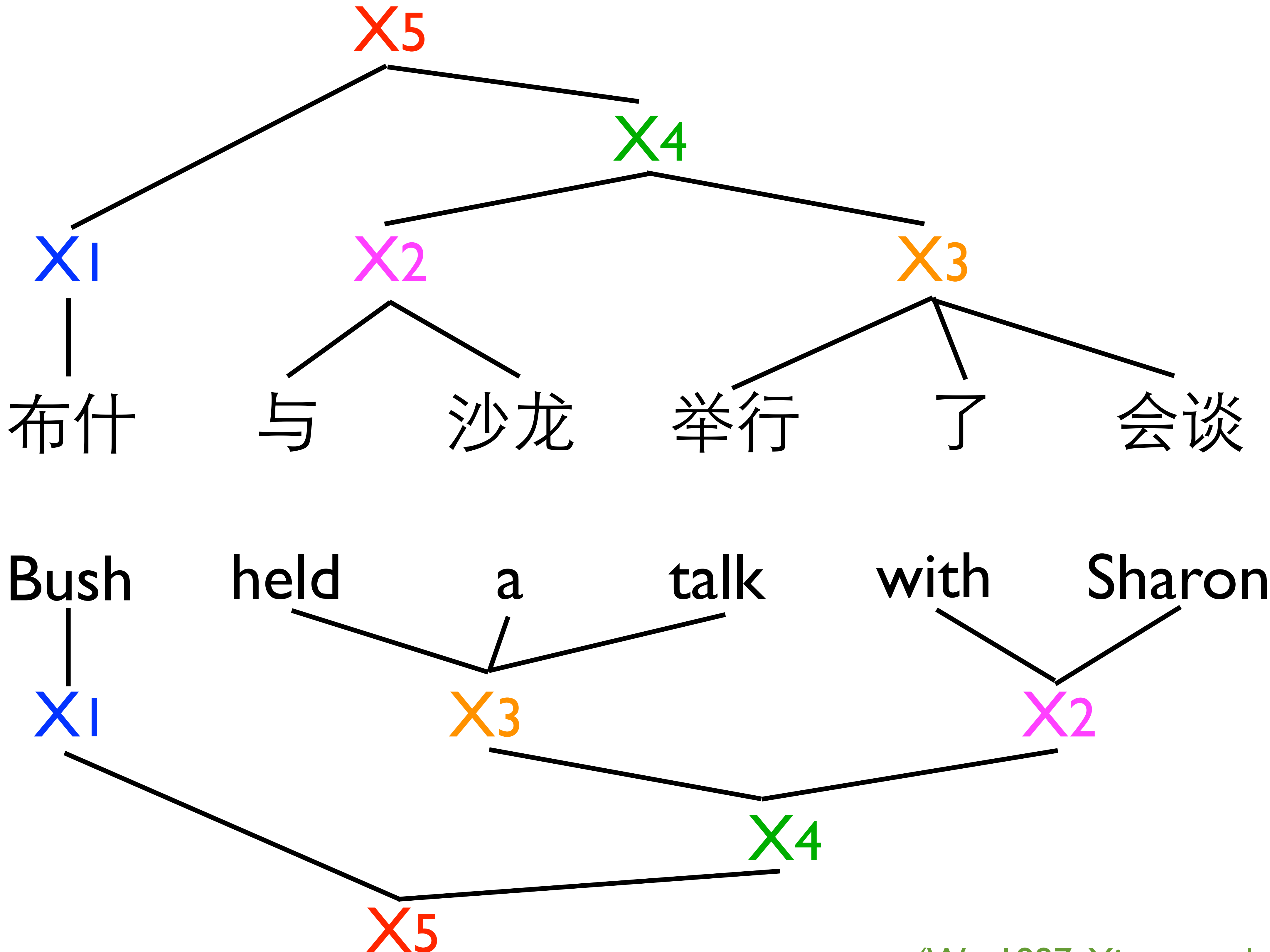
(Wu, 1997; Xiong et al., 2006)

CKY Parsing



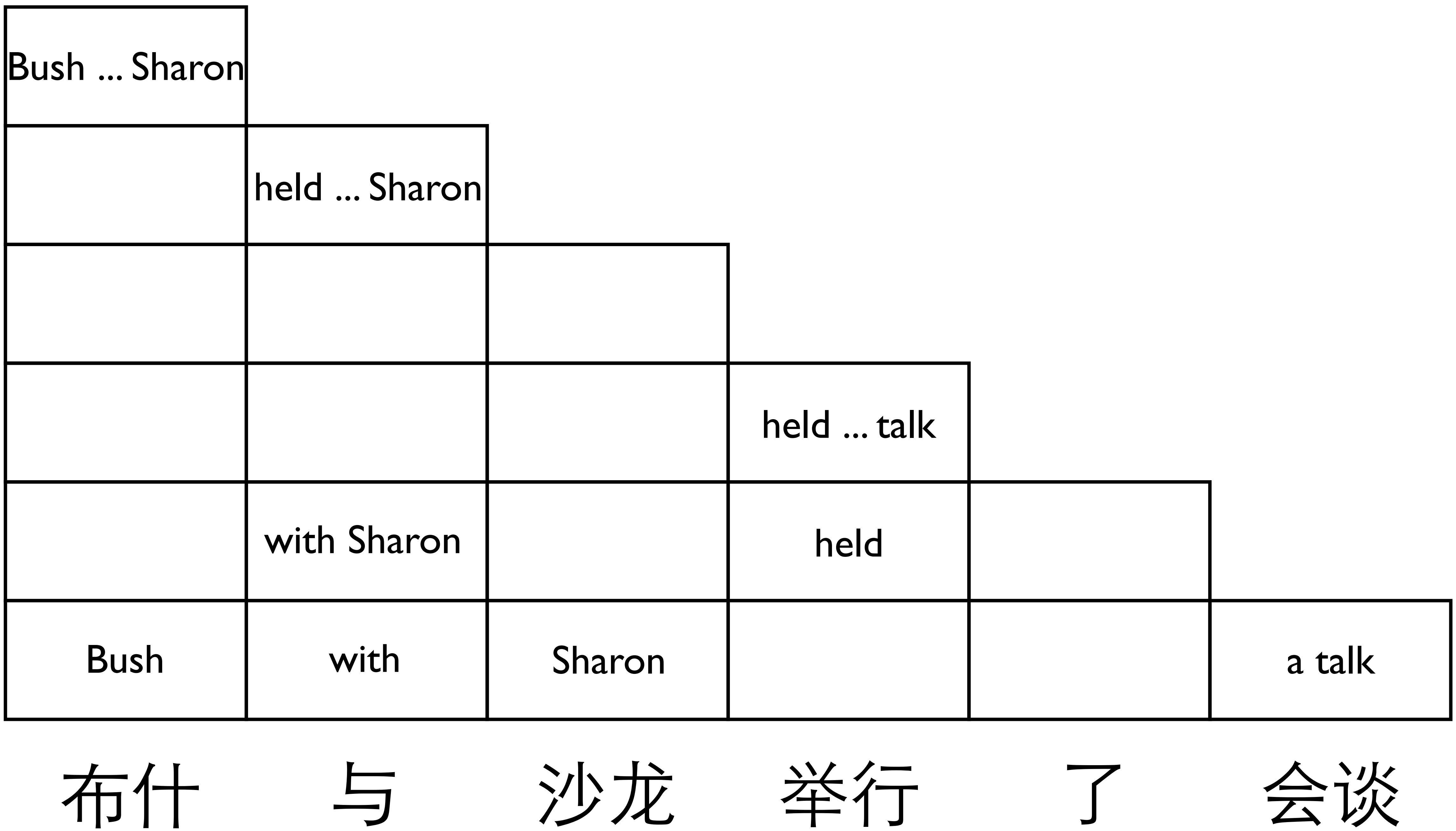
(Wu, 1997; Xiong et al., 2006)

CKY Parsing



(Wu, 1997; Xiong et al., 2006)

Chart



(Wu, 1997; Xiong et al., 2006)

Syntax-based MT

SCFGs without linguistic syntax

inverted transduction grammar

hierarchical phrase-based model

STSGs with linguistic syntax

string-to-tree

tree-to-string

tree-to-tree

Translation Templates

从北京到上海	from Beijing to Shanghai
从武汉到天津	from Wuhan to Tianjin
从广州到重庆	from Guangzhou to Chongqing

(Chiang, 2005)

Translation Templates

从北京到上海	from Beijing to Shanghai
从武汉到天津	from Wuhan to Tianjin
从广州到重庆	from Guangzhou to Chongqing

(Chiang, 2005)

Translation Templates

从北京到上海	from Beijing to Shanghai
从武汉到天津	from Wuhan to Tianjin
从广州到重庆	from Guangzhou to Chongqing

(从 X_1 到 X_2, from X_1 to X_2)

(Chiang, 2005)

Translation Templates

从北京到上海	from Beijing to Shanghai
从武汉到天津	from Wuhan to Tianjin
从广州到重庆	from Guangzhou to Chongqing

(从 X_1 到 X_2, from X_1 to X_2)

(北京, Beijing) (上海, Shanghai) (武汉, Wuhan)

(天津, Tianjin) (广州, Guangzhou) (重庆, Chongqing)

(Chiang, 2005)

Translation Templates

从北京到上海	from Beijing to Shanghai
从武汉到天津	from Wuhan to Tianjin
从广州到重庆	from Guangzhou to Chongqing

(从 X_1 到 X_2, from X_1 to X_2) *hierarchical phrase pair*

(北京, Beijing) (上海, Shanghai) (武汉, Wuhan)

(天津, Tianjin) (广州, Guangzhou) (重庆, Chongqing)

(Chiang, 2005)

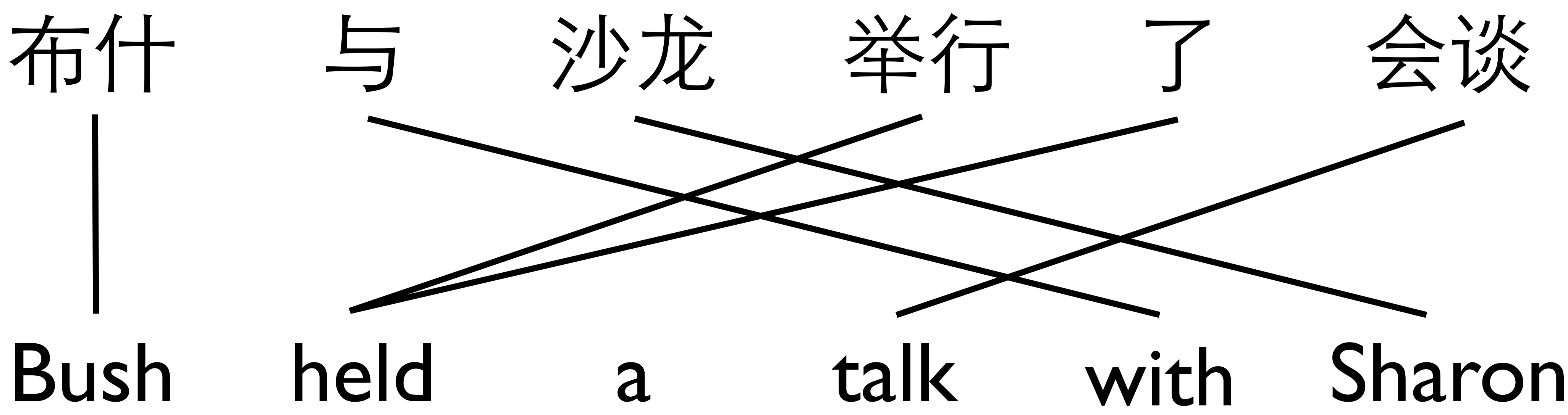
Hierarchical Phrase Extraction

布什 与 沙龙 举行 了 会谈

Bush held a talk with Sharon

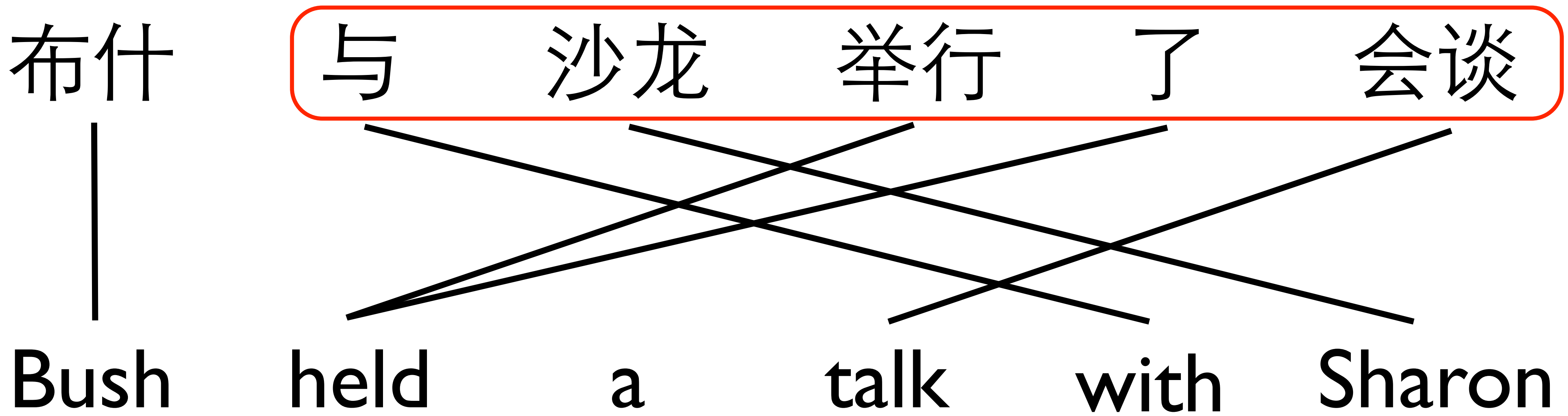
(Chiang, 2005)

Hierarchical Phrase Extraction



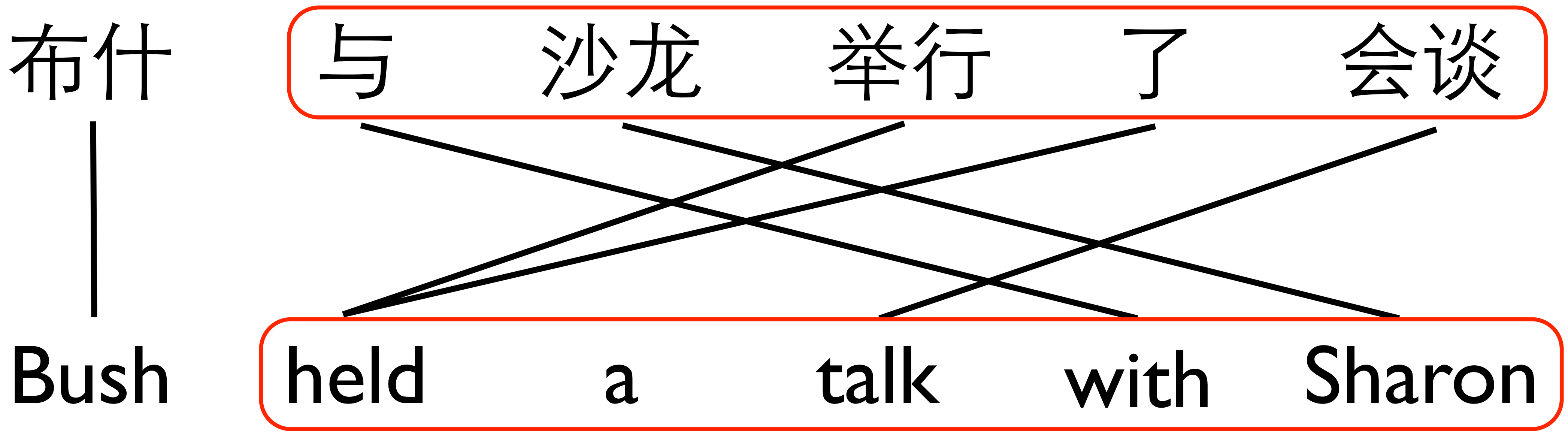
(Chiang, 2005)

Hierarchical Phrase Extraction



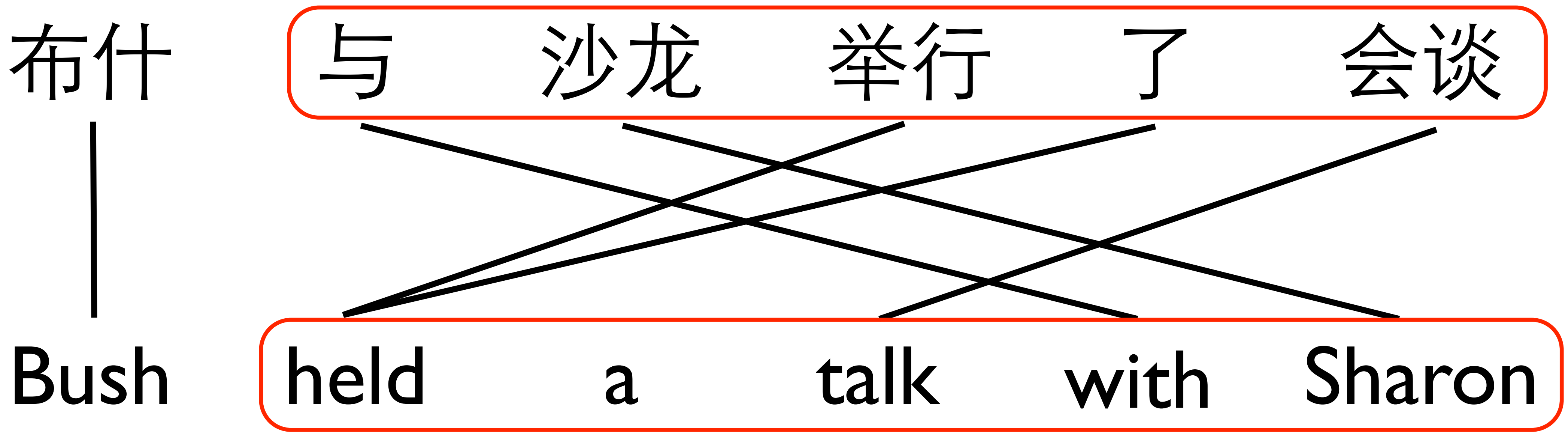
(Chiang, 2005)

Hierarchical Phrase Extraction



(Chiang, 2005)

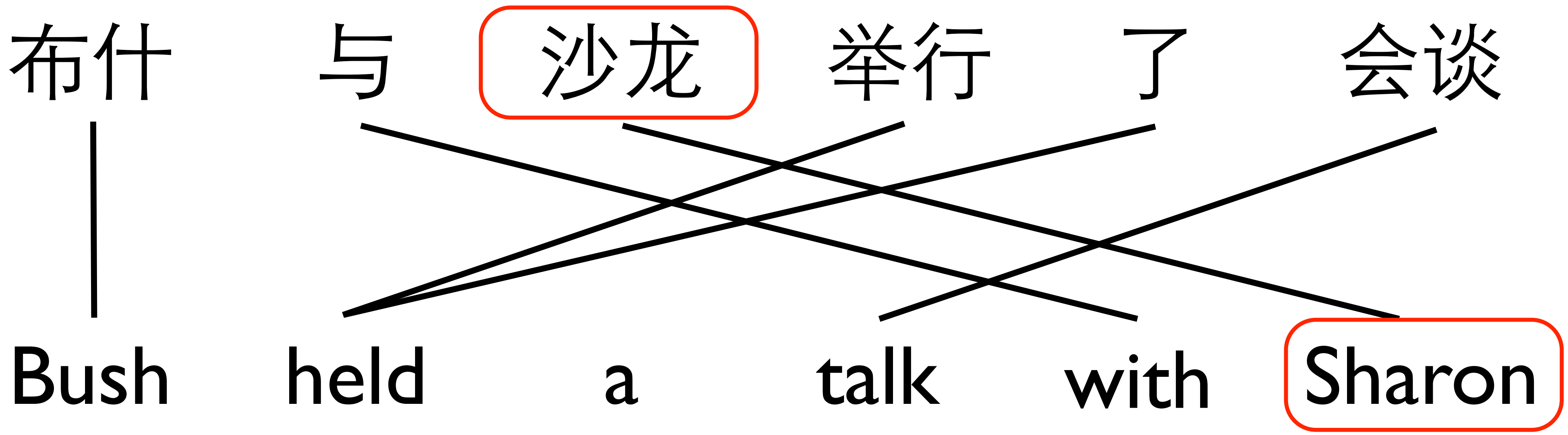
Hierarchical Phrase Extraction



(与 沙龙 举行 了 会谈, held a talk with Sharon)

(Chiang, 2005)

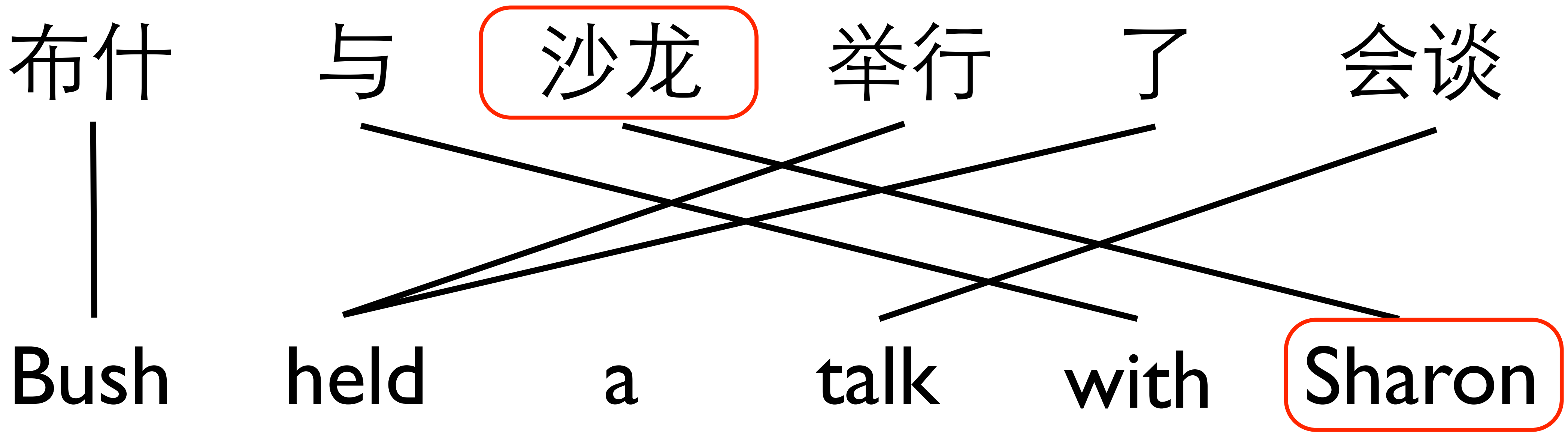
Hierarchical Phrase Extraction



(与 沙龙 举行 了 会谈, held a talk with Sharon)

(Chiang, 2005)

Hierarchical Phrase Extraction

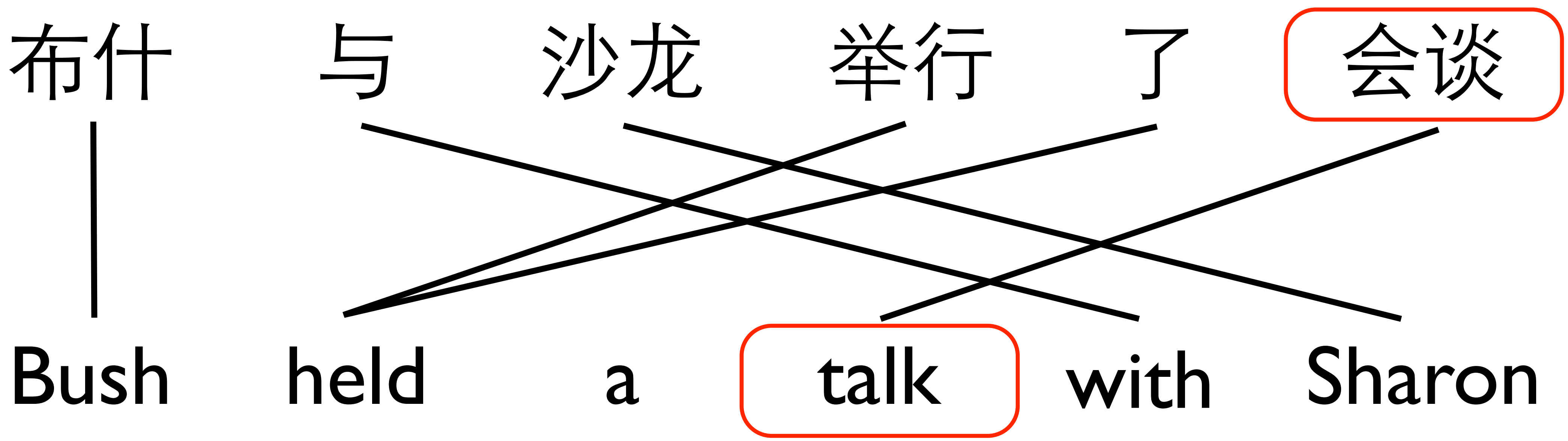


(与 沙龙 举行 了 会谈, held a talk with Sharon)

(沙龙, Sharon)

(Chiang, 2005)

Hierarchical Phrase Extraction

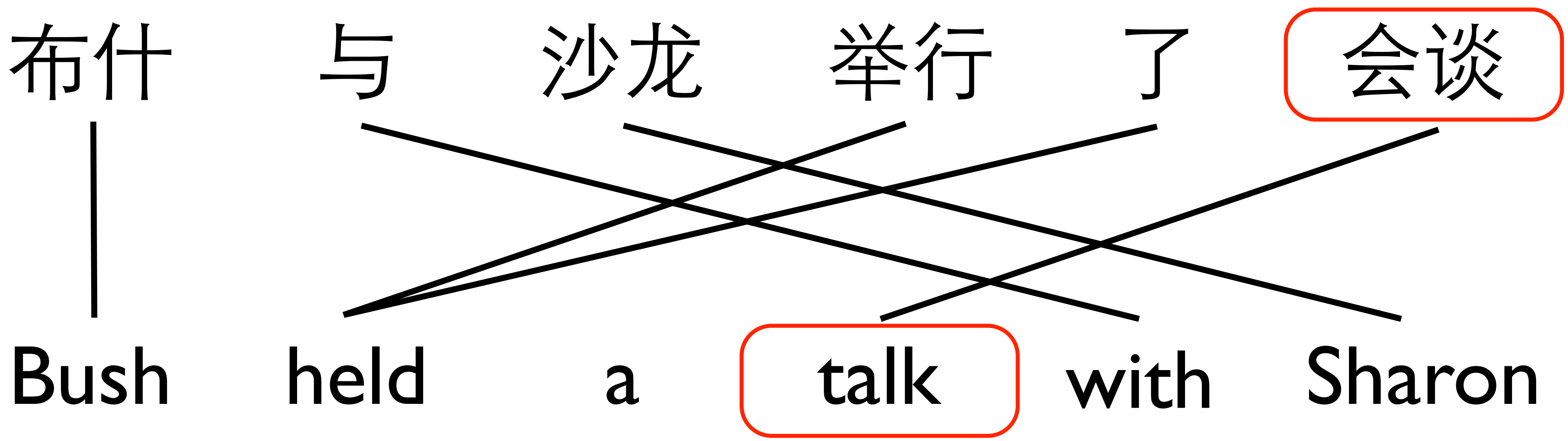


(与 沙龙 举行 了 会谈, held a talk with Sharon)

(沙龙, Sharon)

(Chiang, 2005)

Hierarchical Phrase Extraction



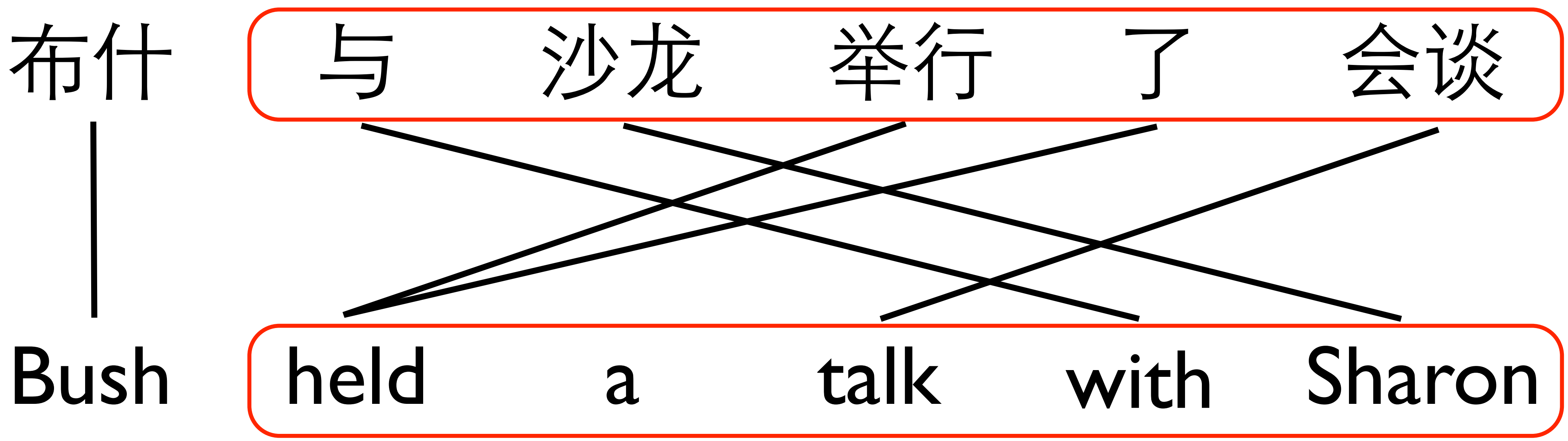
(与 沙龙 举行 了 会谈, held a talk with Sharon)

(沙龙, Sharon)

(会谈, talk)

(Chiang, 2005)

Hierarchical Phrase Extraction



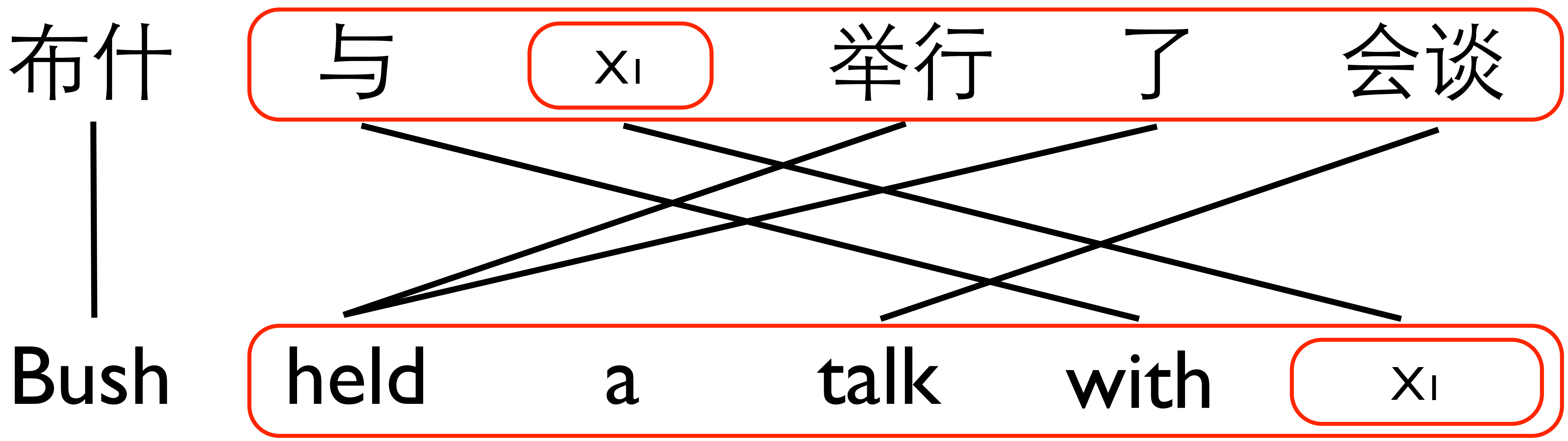
(与 沙龙 举行 了 会谈, held a talk with Sharon)

(沙龙, Sharon)

(会谈, talk)

(Chiang, 2005)

Hierarchical Phrase Extraction



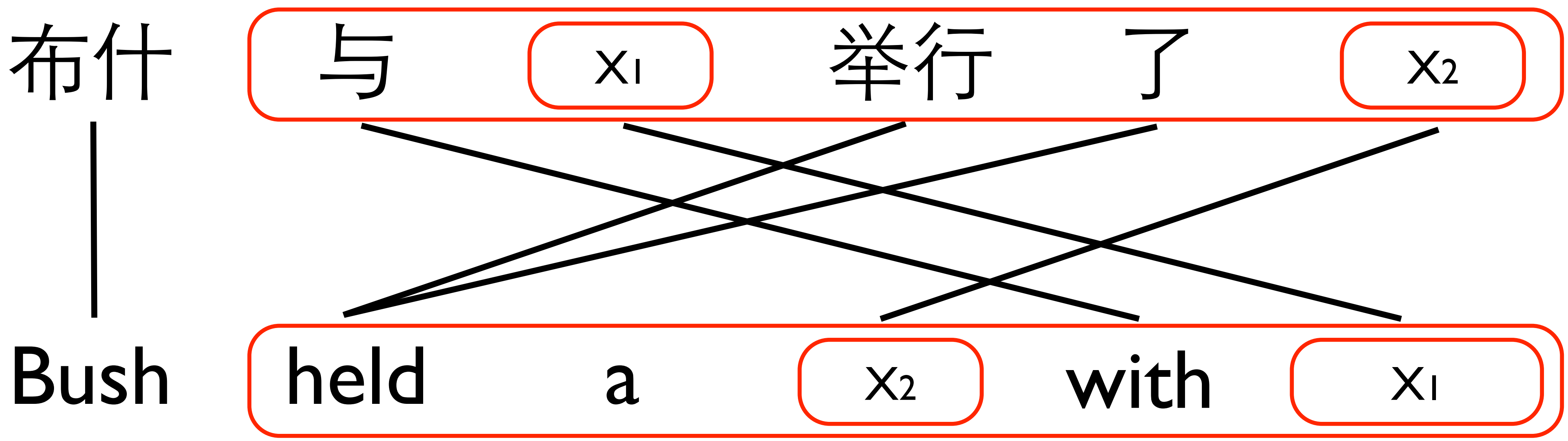
(与 沙龙 举行 了 会谈, held a talk with Sharon)

(沙龙, Sharon)

(会谈, talk)

(Chiang, 2005)

Hierarchical Phrase Extraction



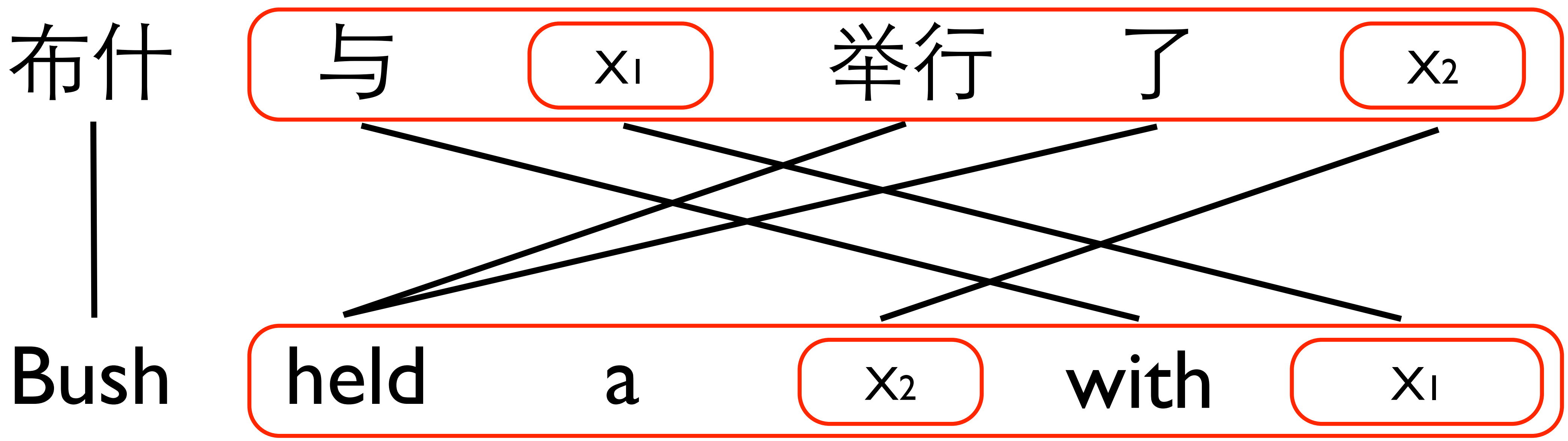
(与 沙龙 举行 了 会谈, held a talk with Sharon)

(沙龙, Sharon)

(会谈, talk)

(Chiang, 2005)

Hierarchical Phrase Extraction



(与 沙龙 举行 了 会谈, held a talk with Sharon)

(沙龙, Sharon)

(会谈, talk)

(与 X_1 举行 了 X_2, held a X_2 with X_1)

(Chiang, 2005)

Hierarchical Phrase-based Translation

- Hierarchical phrase-based model is based on SCFG without linguistic syntax

(Chiang, 2005)

Hierarchical Phrase-based Translation

- Hierarchical phrase-based model is based on SCFG without linguistic syntax

lexical rules

$$X \rightarrow (\text{沙龙}, \text{Sharon})$$

$$X \rightarrow (\text{会谈}, \text{talk})$$

(Chiang, 2005)

Hierarchical Phrase-based Translation

- Hierarchical phrase-based model is based on SCFG without linguistic syntax

lexical rules

$$X \rightarrow (\text{沙龙}, \text{Sharon})$$

$$X \rightarrow (\text{会谈}, \text{talk})$$

syntactic rules

(Chiang, 2005)

Hierarchical Phrase-based Translation

- Hierarchical phrase-based model is based on SCFG without linguistic syntax

lexical rules

$$X \longrightarrow (\text{沙龙}, \text{Sharon})$$

$$X \longrightarrow (\text{会谈}, \text{talk})$$

syntactic rules

$$X \longrightarrow (\text{与 } X_1 \text{ 举行 了 } X_2, \text{held a } X_2 \text{ with } X_1)$$

(Chiang, 2005)

Hierarchical Phrase-based Translation

- Hierarchical phrase-based model is based on SCFG without linguistic syntax

lexical rules

$$X \rightarrow (\text{沙龙}, \text{Sharon})$$

$$X \rightarrow (\text{会谈}, \text{talk})$$

syntactic rules

$$X \rightarrow (\text{与 } X_1 \text{ 举行 了 } X_2, \text{held a } X_2 \text{ with } X_1)$$

$$X \rightarrow (\text{布什 } X_1, \text{Bush } X_1)$$

(Chiang, 2005)

Hierarchical Phrase-based Translation

- Hierarchical phrase-based model is based on SCFG without linguistic syntax

lexical rules

$X \longrightarrow$ (沙龙, Sharon)

$X \longrightarrow$ (会谈, talk)

syntactic rules

$X \longrightarrow$ (与 X_1 举行 了 X_2, held a X_2 with X_1)

$X \longrightarrow$ (布什 X_1, Bush X_1)

ITG is a special case of SCFG

(Chiang, 2005)

CKY Parsing

布什 与 沙龙 举行 了 会谈

(Chiang, 2005)

CKY Parsing

X → (沙龙, Sharon)

布什　与　沙龙　举行　了　会谈

(Chiang, 2005)

CKY Parsing

X → (沙龙, Sharon)

布什　　与　　沙龙　　举行　　了　　会谈

Sharon

(Chiang, 2005)

CKY Parsing

X → (沙龙, Sharon)

X_1

布什 与 沙龙 举行 了 会谈

Sharon

X_1

(Chiang, 2005)

CKY Parsing

X → (沙龙, Sharon)

X → (会谈, talk)

X1

布什 与 沙龙 举行 了 会谈

Sharon

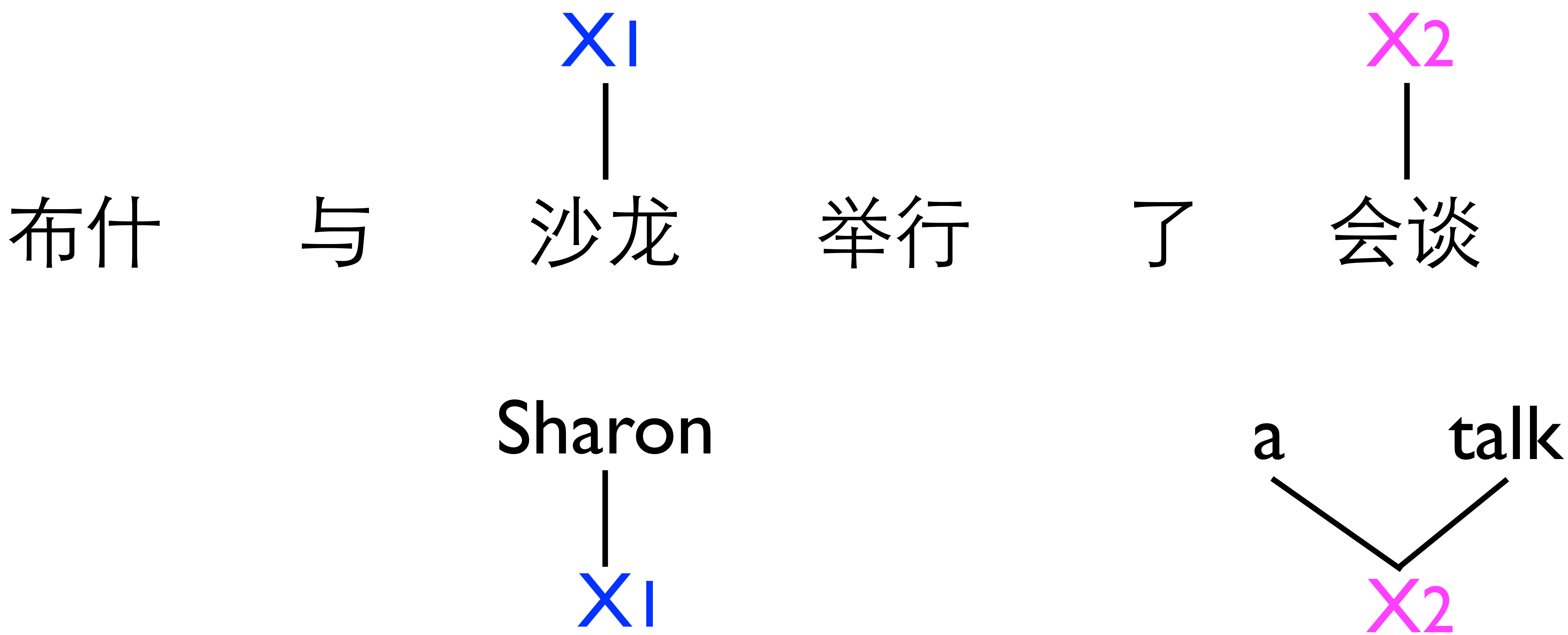
X1

(Chiang, 2005)

CKY Parsing

X → (沙龙, Sharon)

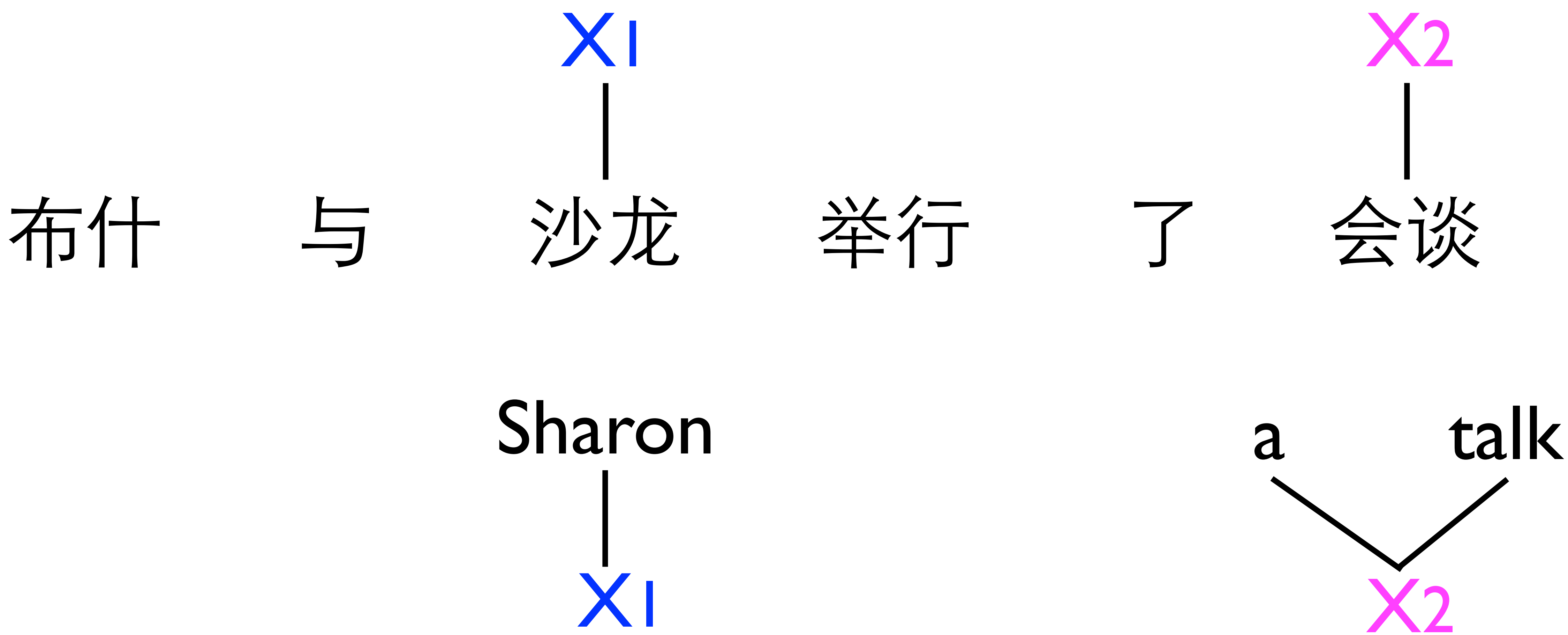
X → (会谈, talk)



(Chiang, 2005)

CKY Parsing

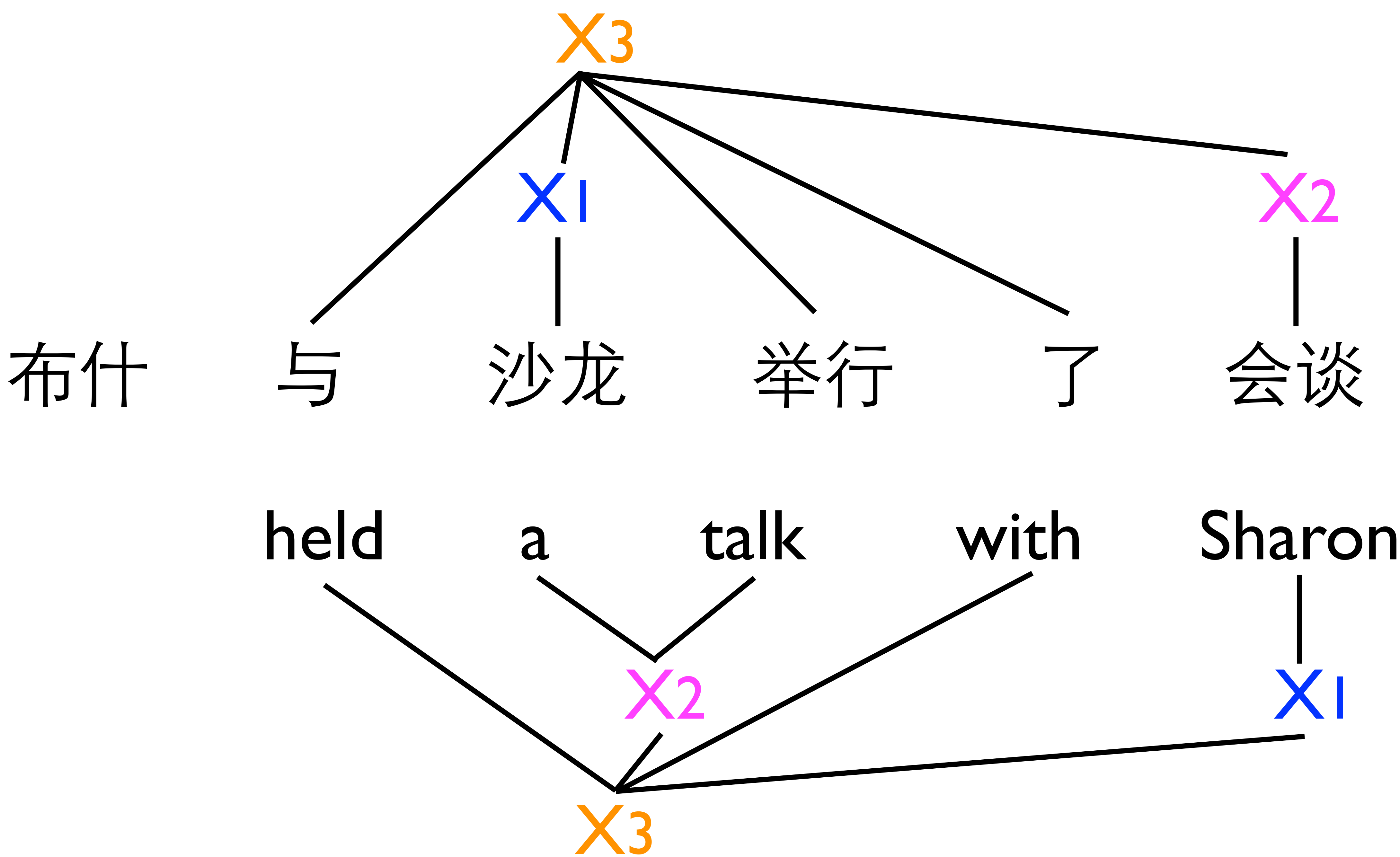
$X \rightarrow$ (与 X_1 举行 了 X_2, held a X_2 with X_1)



(Chiang, 2005)

CKY Parsing

X → (与 X_1 举行 了 X_2, held a X_2 with X_1)

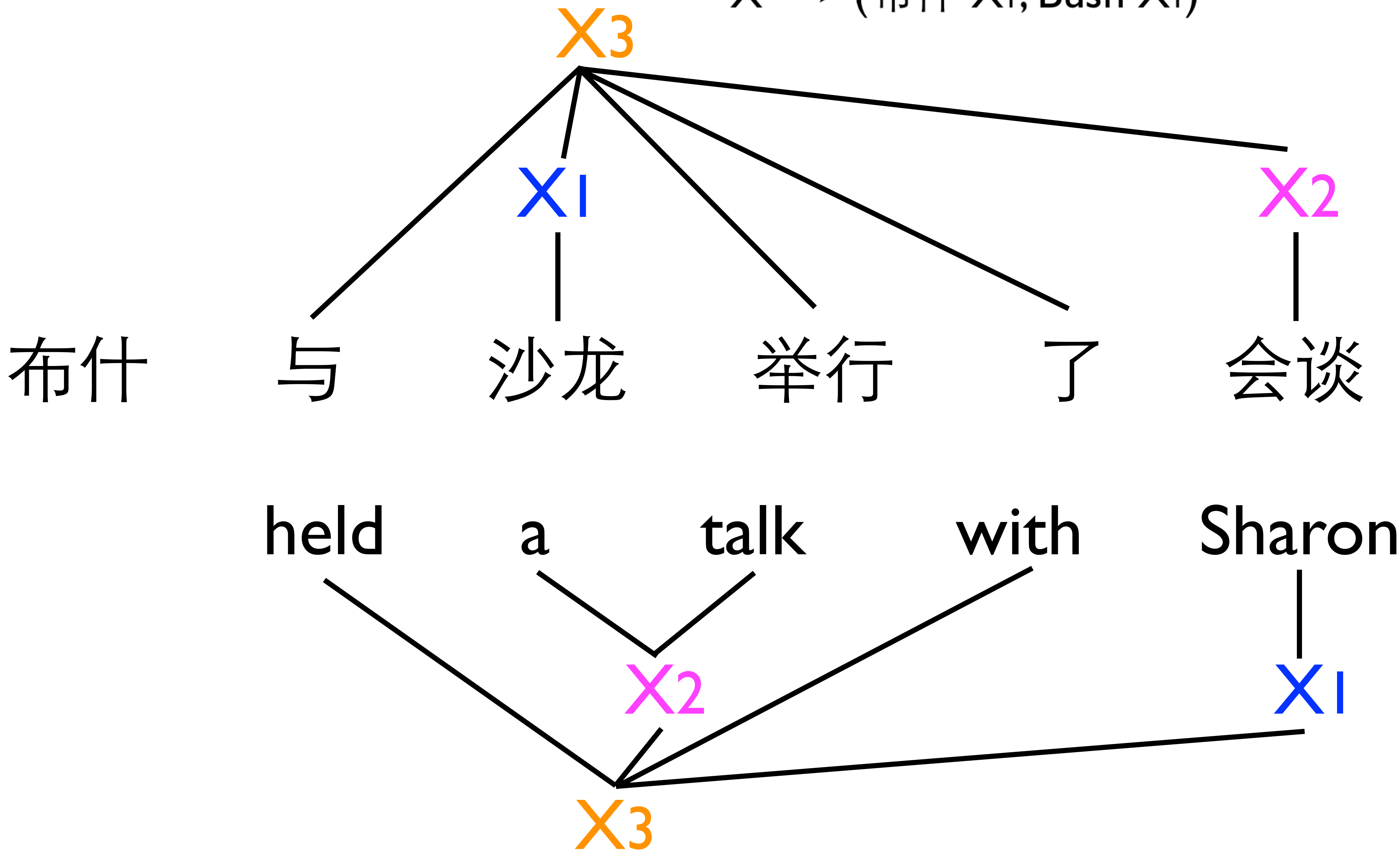


(Chiang, 2005)

CKY Parsing

$X \rightarrow$ (与 X_1 举行 了 X_2, held a X_2 with X_1)

$X \rightarrow$ (布什 X_1, Bush X_1)

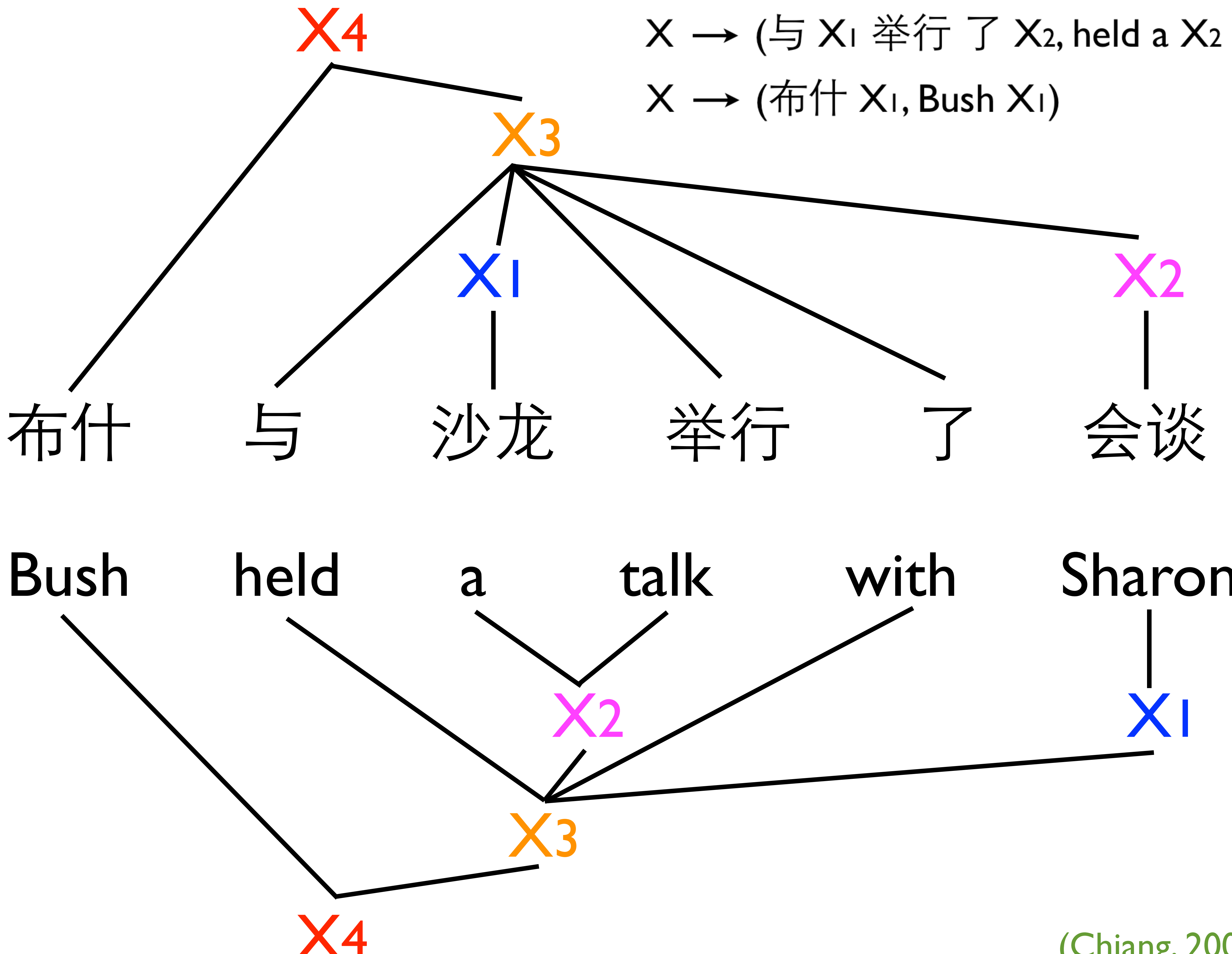


(Chiang, 2005)

CKY Parsing

X → (与 X_1 举行 了 X_2, held a X_2 with X_1)

X → (布什 X_1, Bush X_1)



(Chiang, 2005)

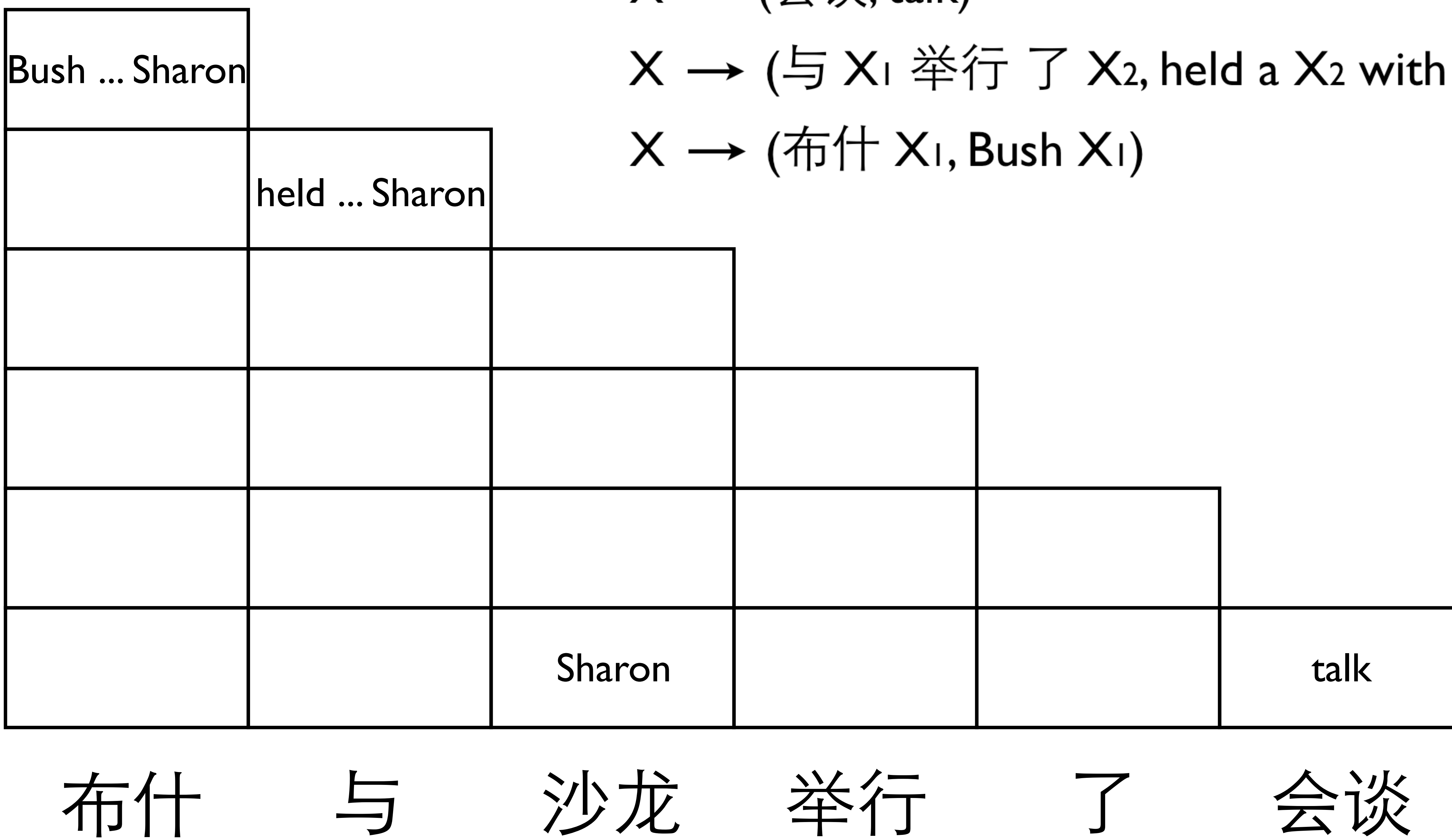
Chart

$X \rightarrow$ (沙龙, Sharon)

$X \rightarrow$ (会谈, talk)

$X \rightarrow$ (与 X_1 举行 了 X_2, held a X_2 with X_1)

$X \rightarrow$ (布什 X_1, Bush X_1)



(Chiang, 2005)

Syntax-based MT

SCFGs without linguistic syntax

inverted transduction grammar

hierarchical phrase-based model

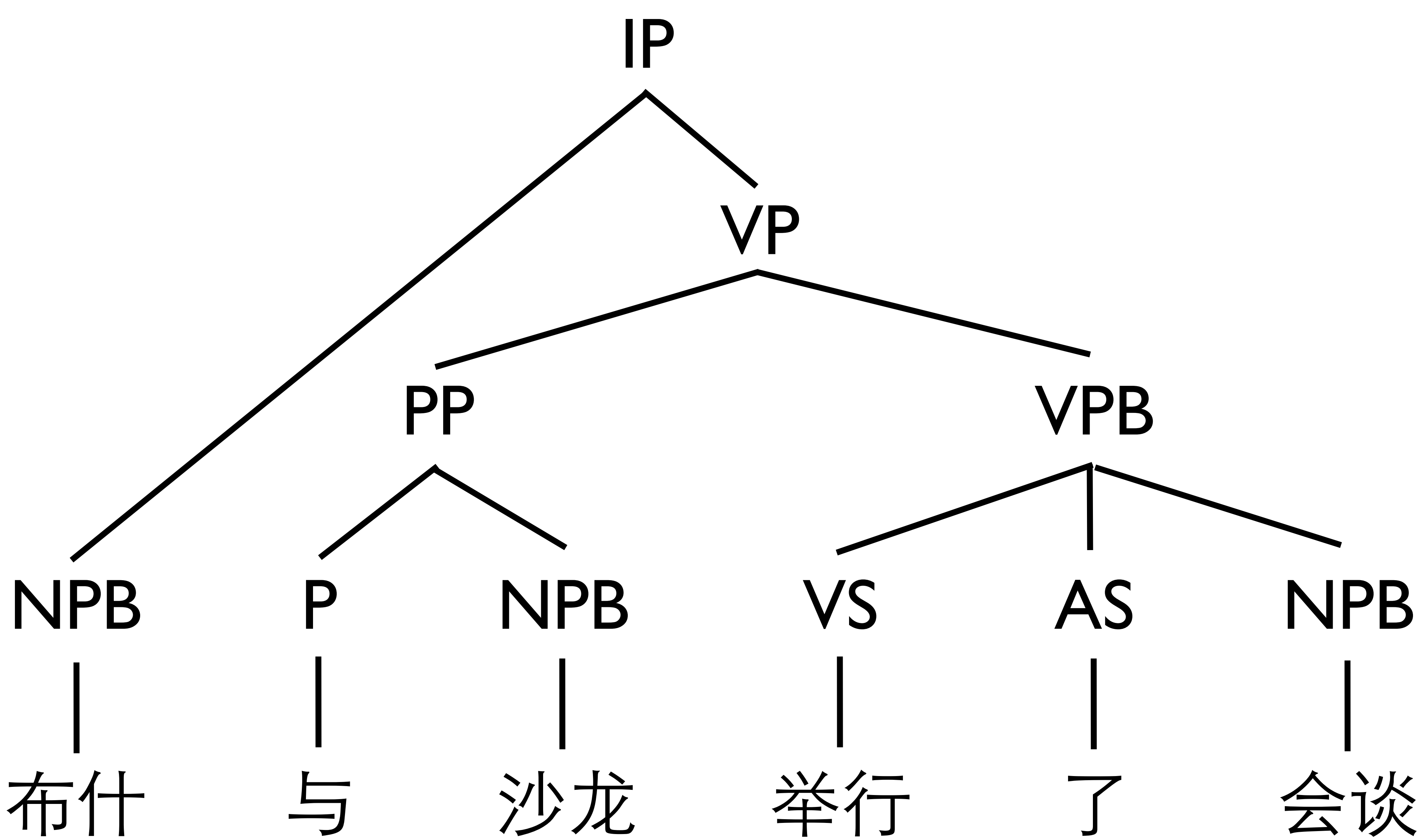
STSGs with linguistic syntax

string-to-tree

tree-to-string

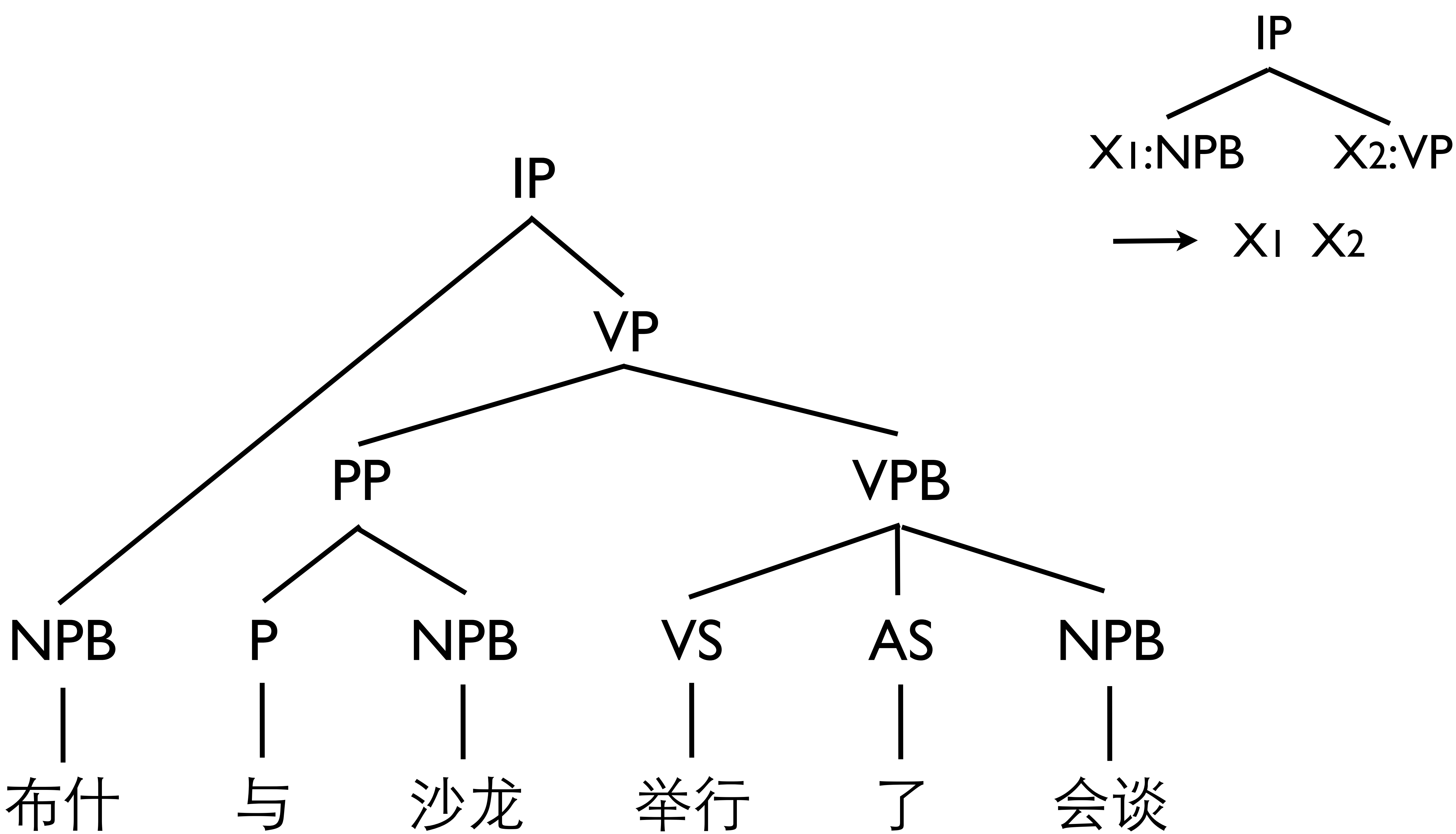
tree-to-tree

Tree-to-String Translation



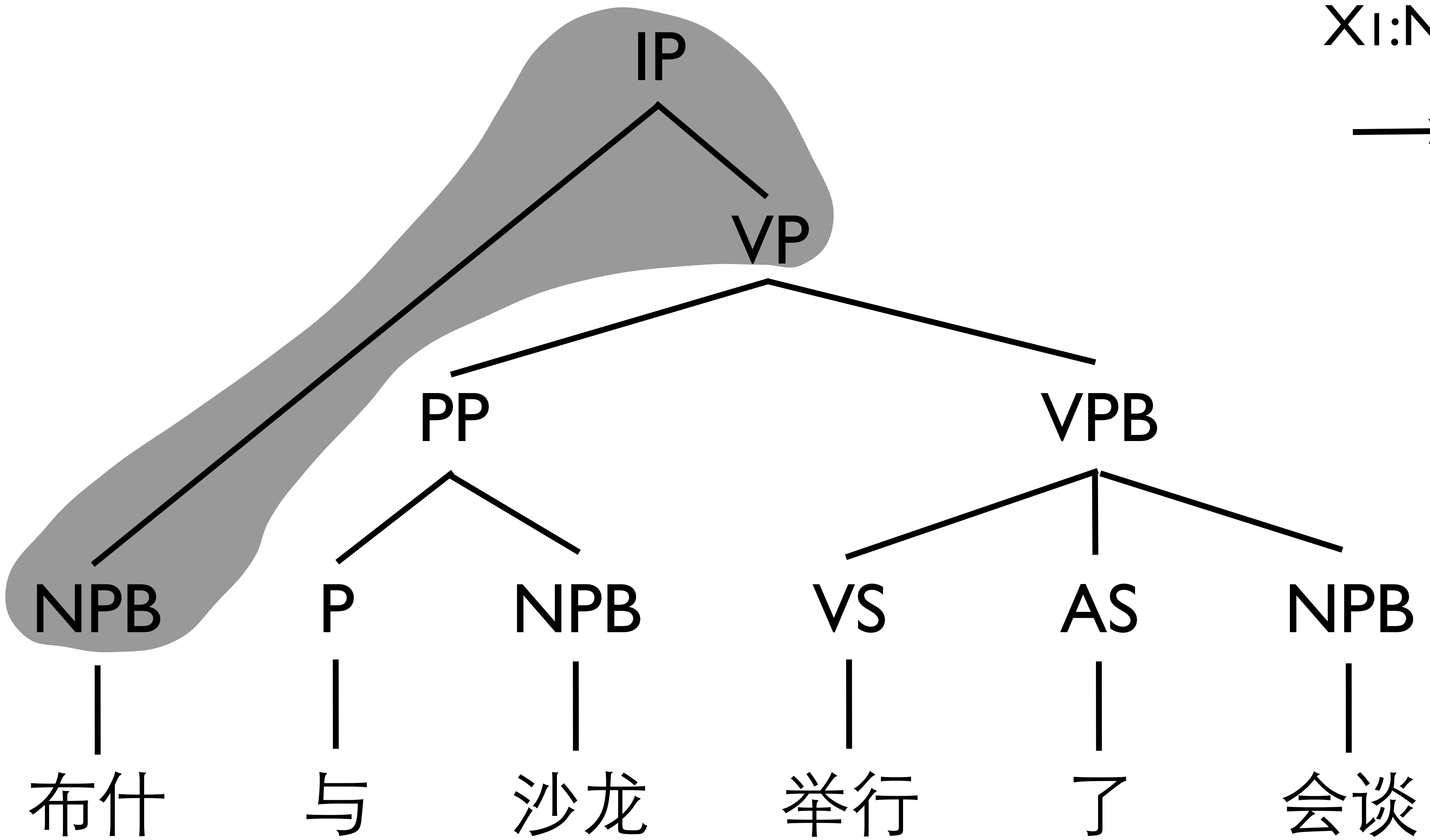
(Liu et al., 2006; Huang et al., 2006)

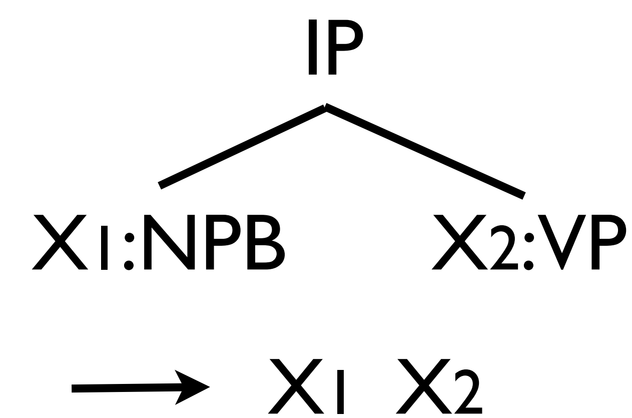
Tree-to-String Translation



(Liu et al., 2006; Huang et al., 2006)

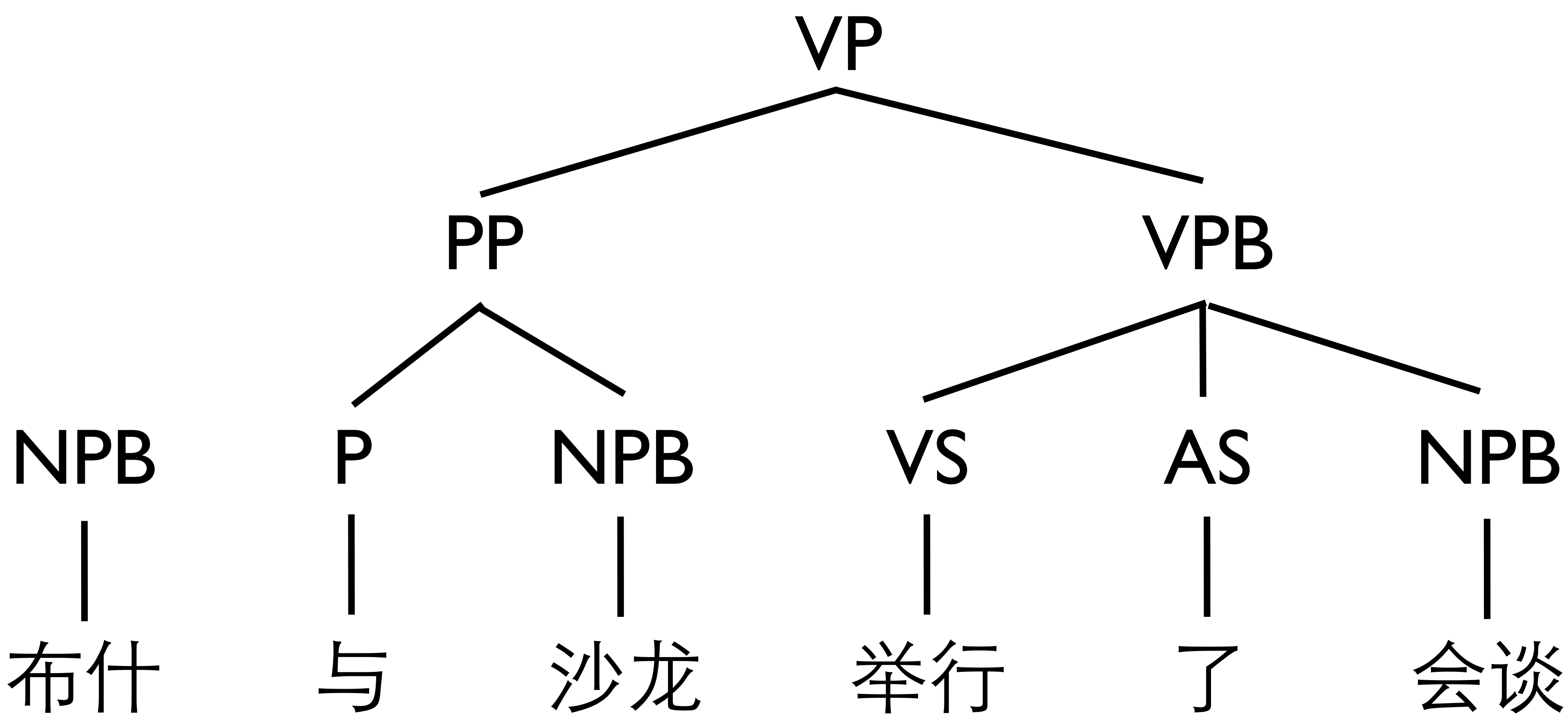
Tree-to-String Translation





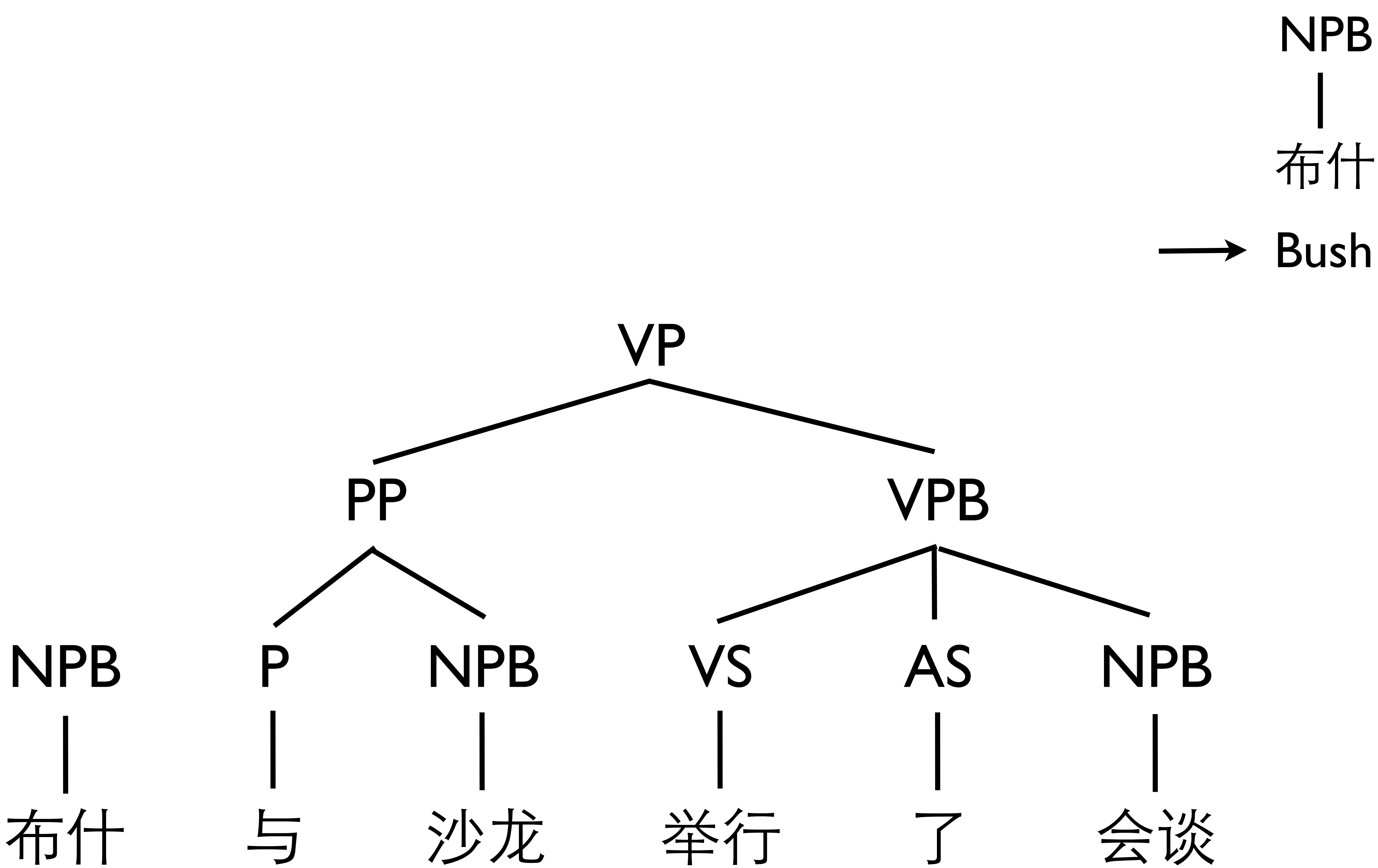
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



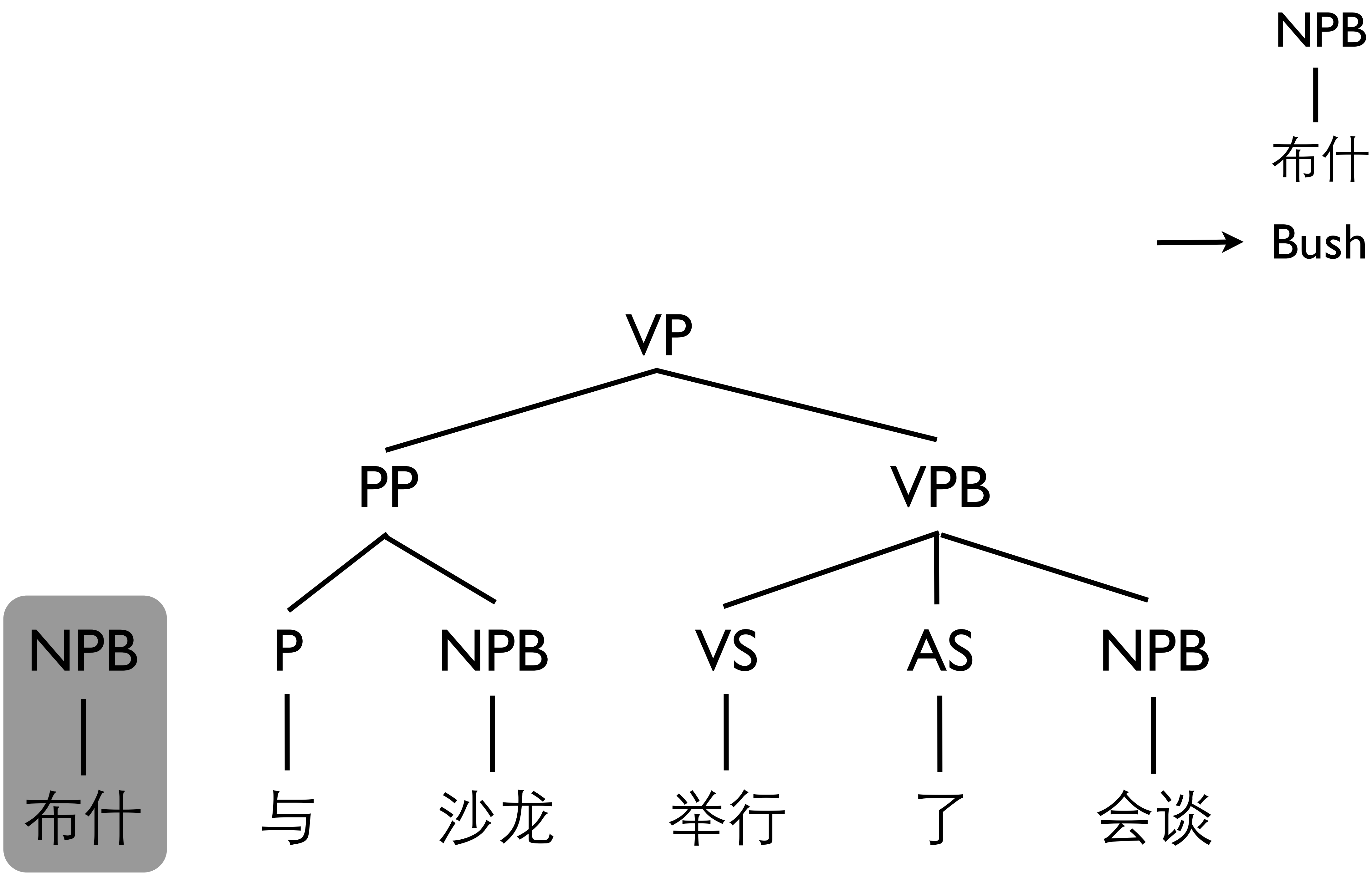
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



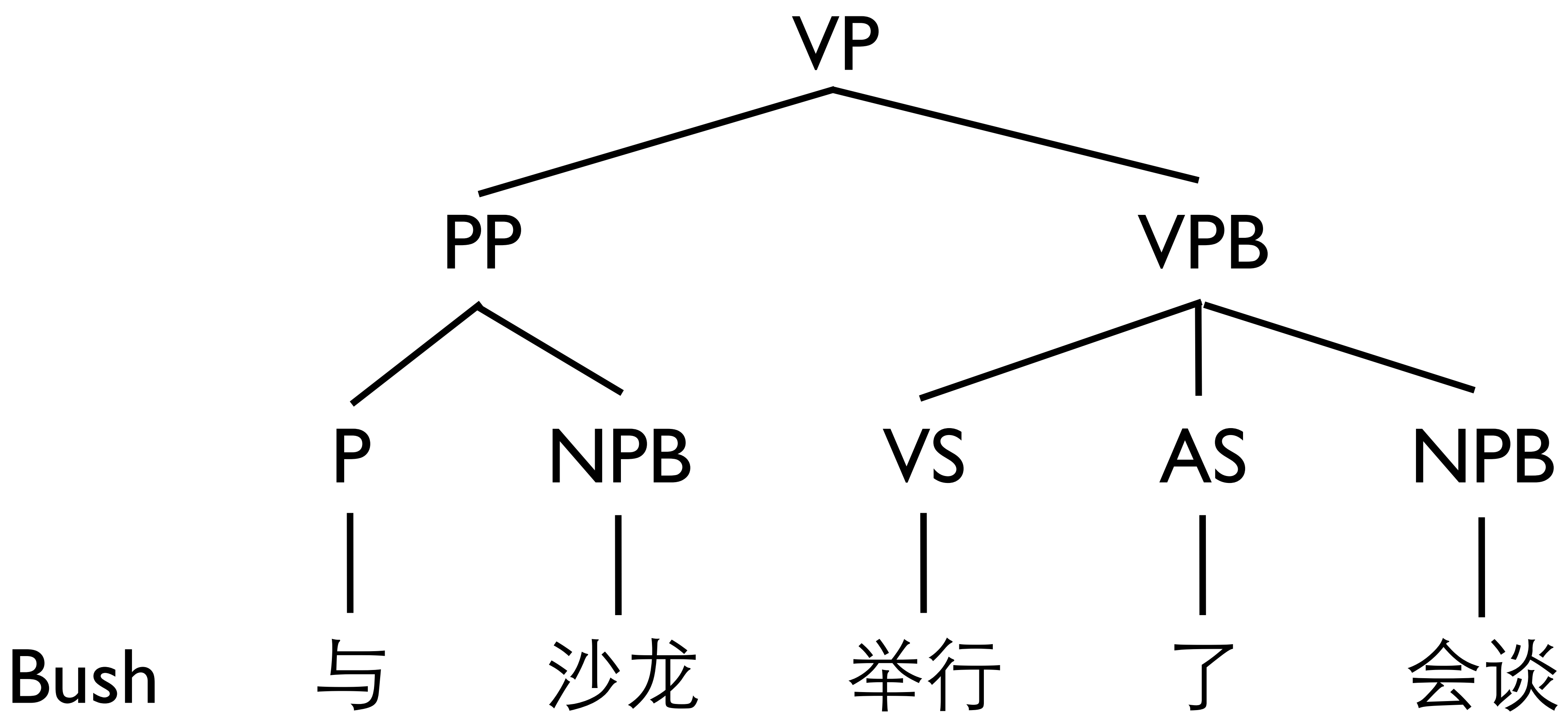
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



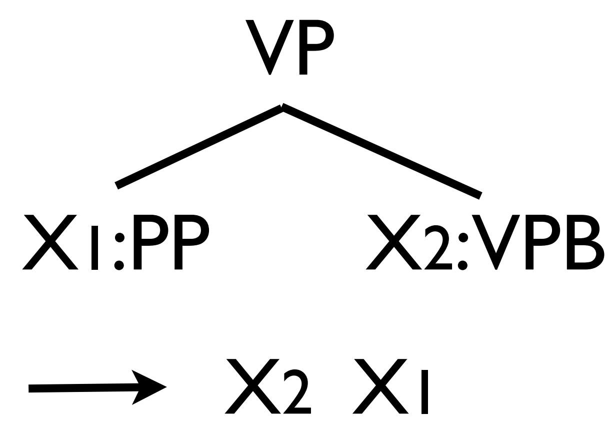
(Liu et al., 2006; Huang et al., 2006)

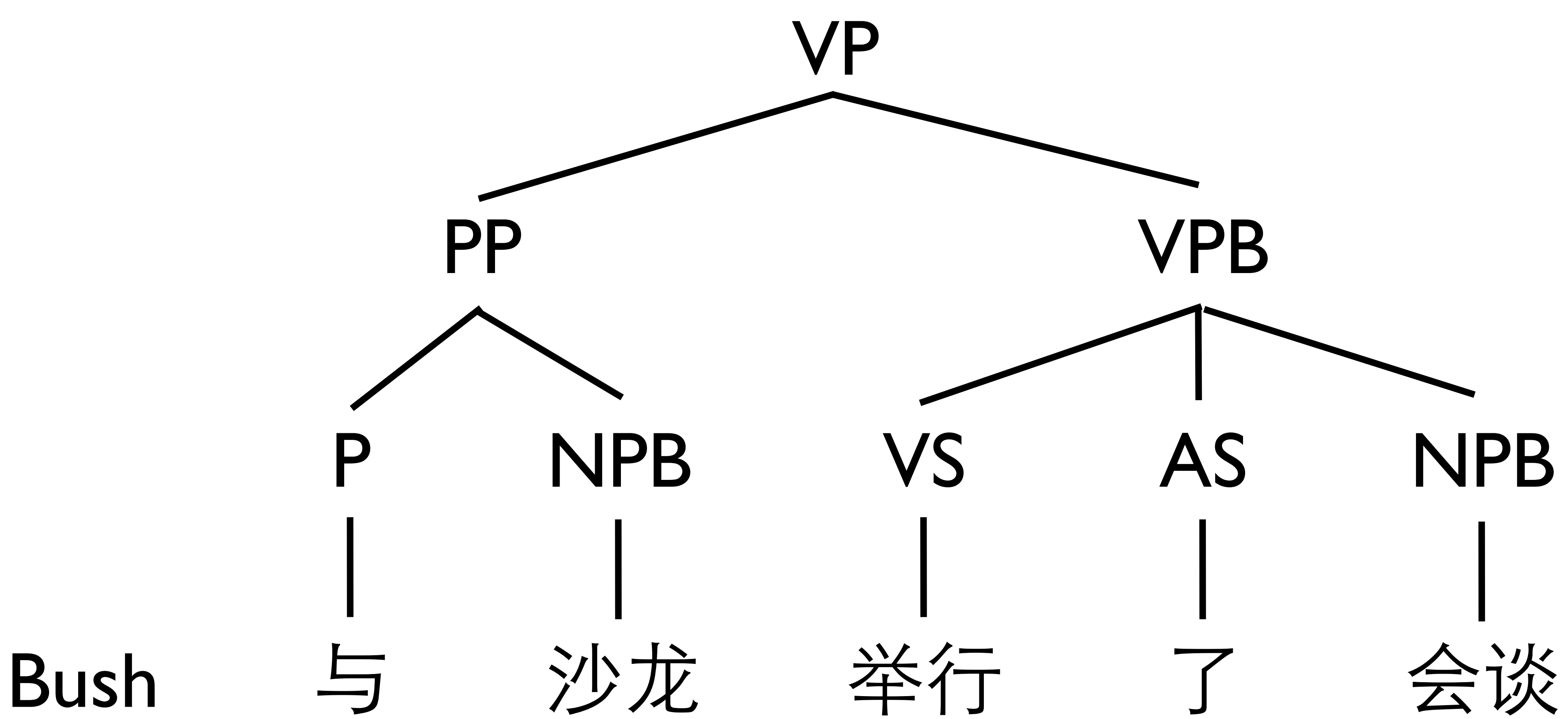
Tree-to-String Translation



(Liu et al., 2006; Huang et al., 2006)

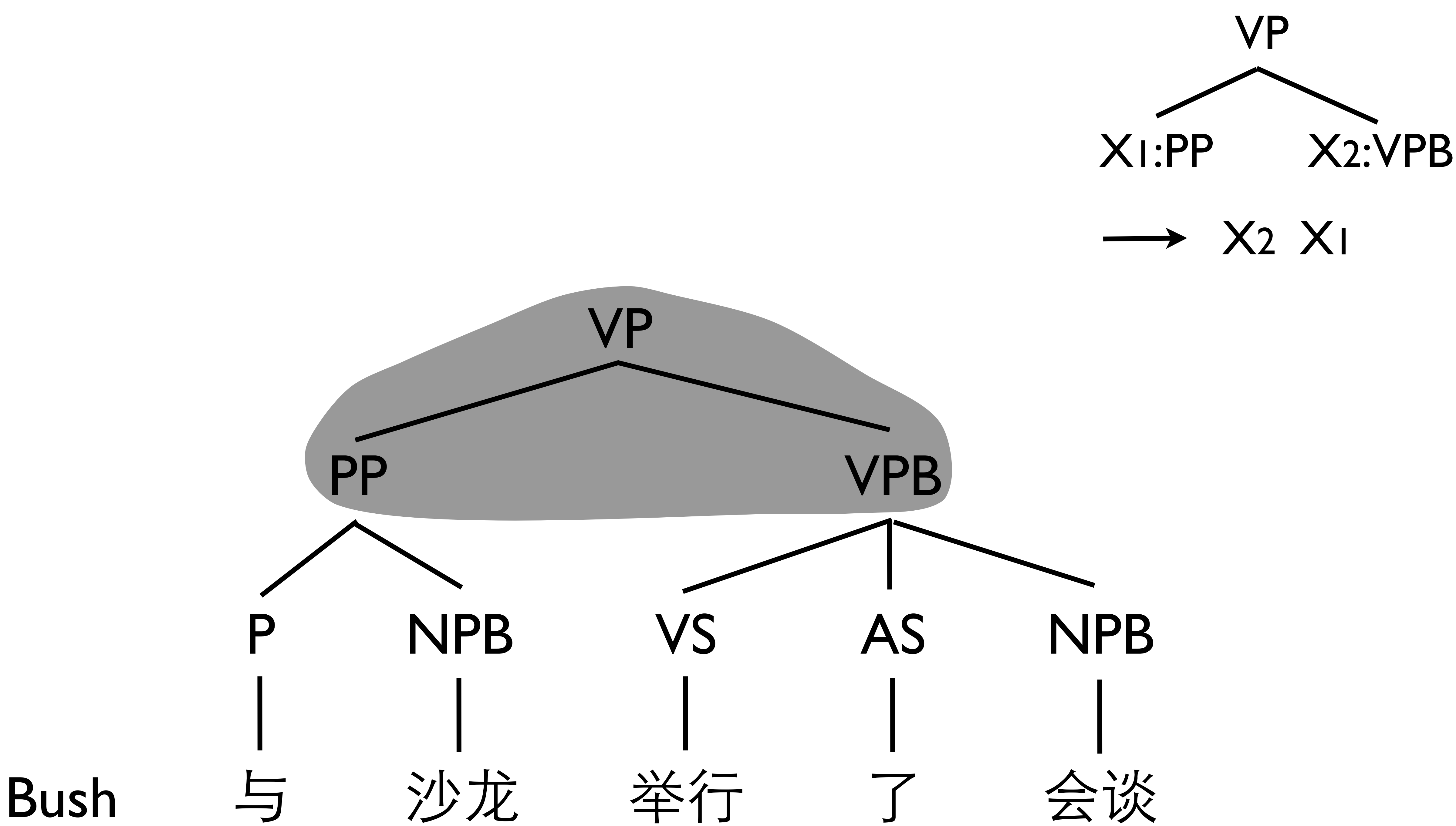
Tree-to-String Translation





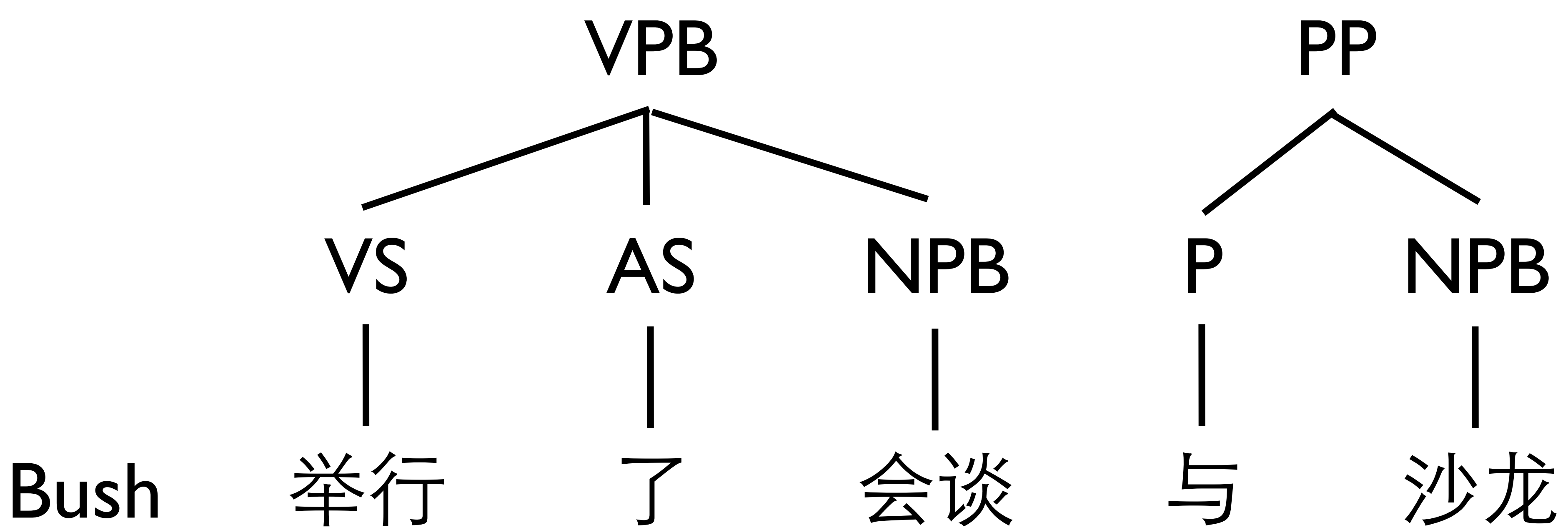
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



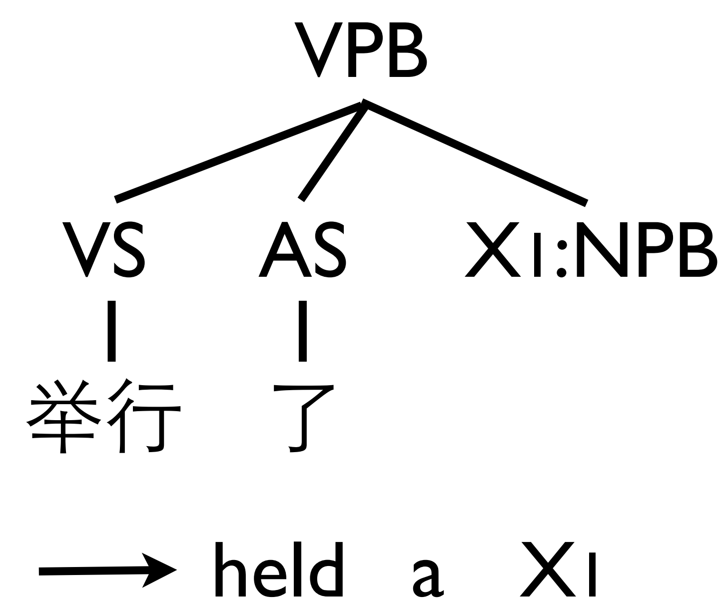
(Liu et al., 2006; Huang et al., 2006)

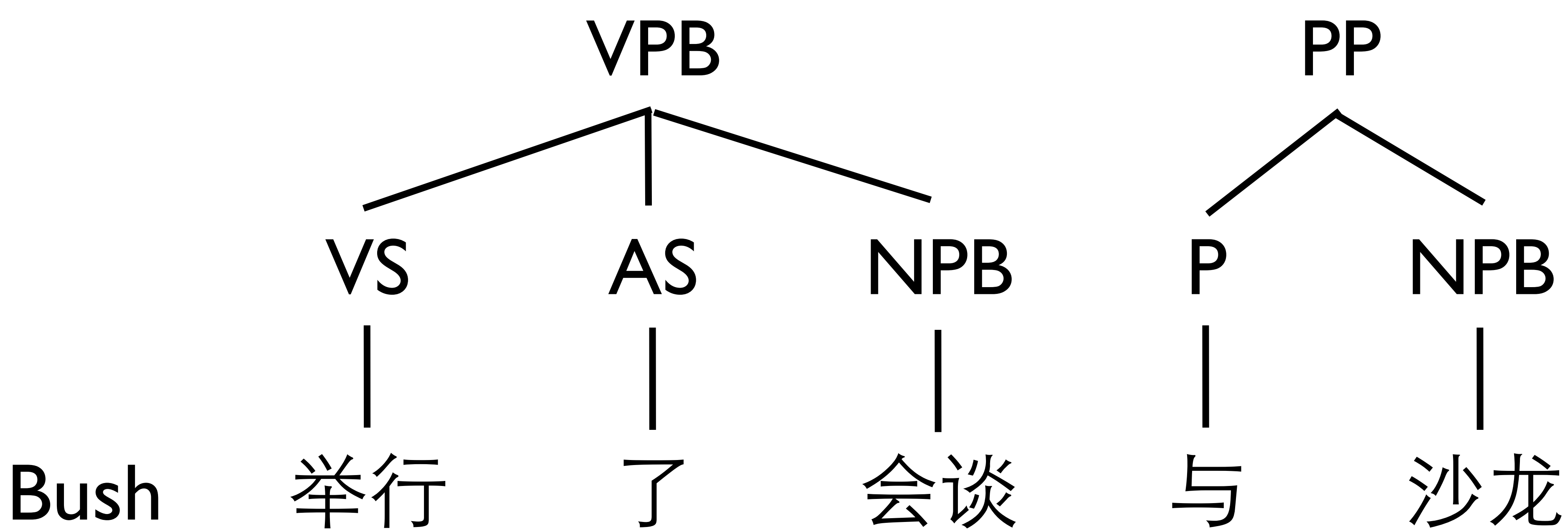
Tree-to-String Translation



(Liu et al., 2006; Huang et al., 2006)

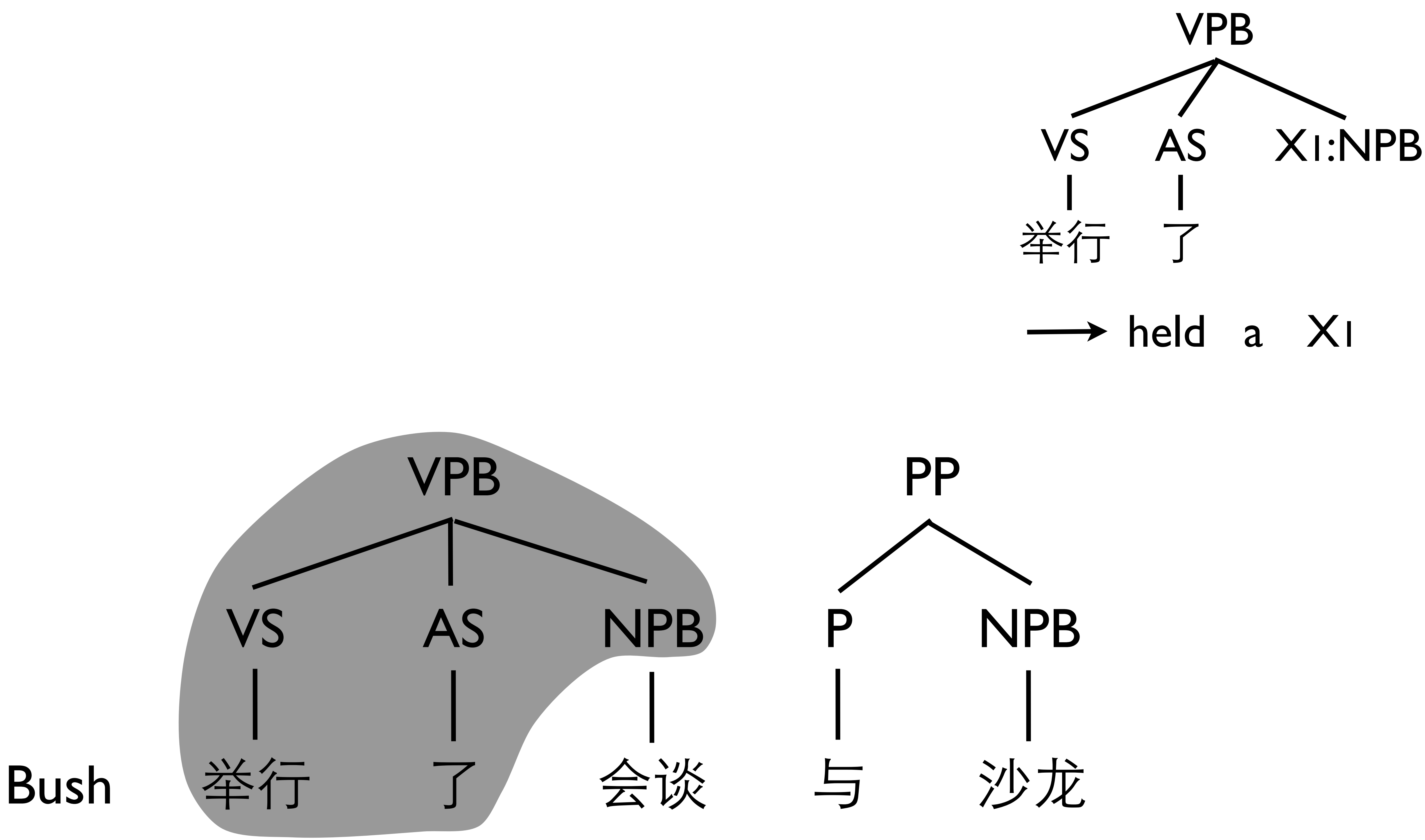
Tree-to-String Translation





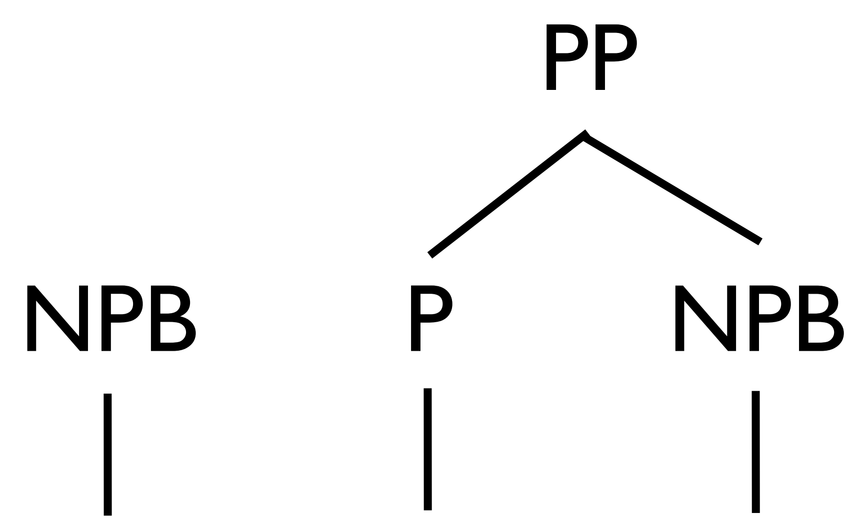
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



(Liu et al., 2006; Huang et al., 2006)

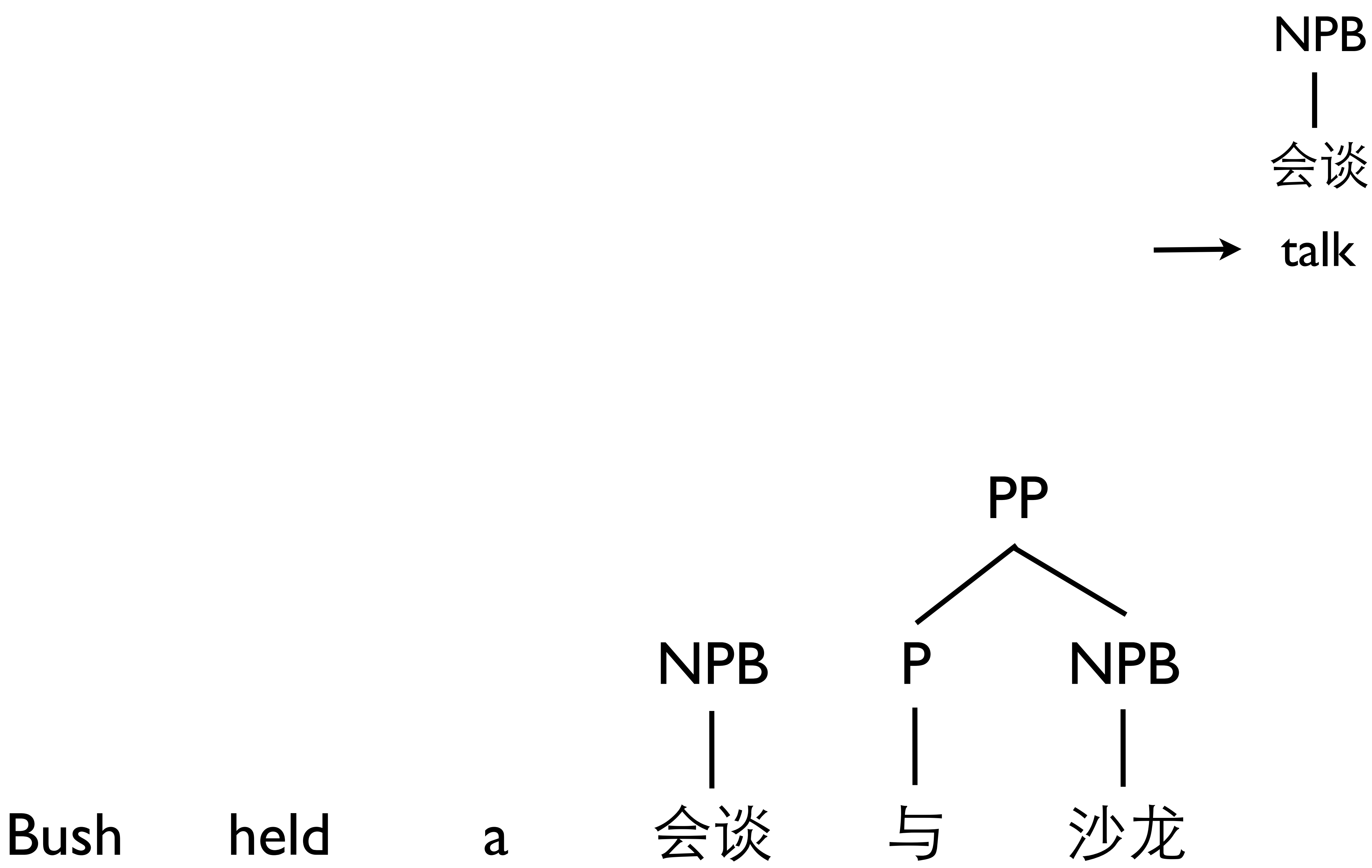
Tree-to-String Translation



Bush held a 会谈 与 沙龙

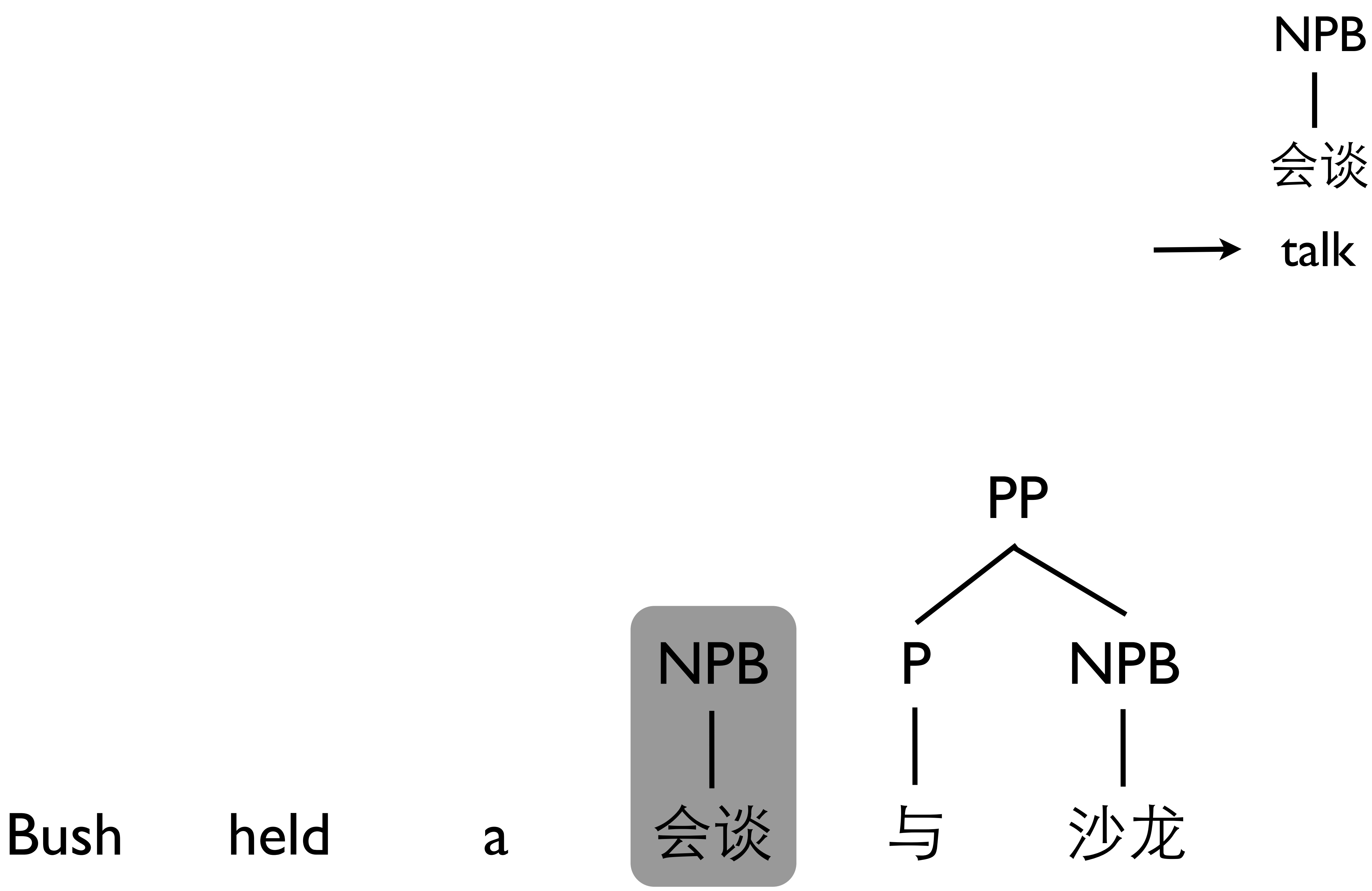
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



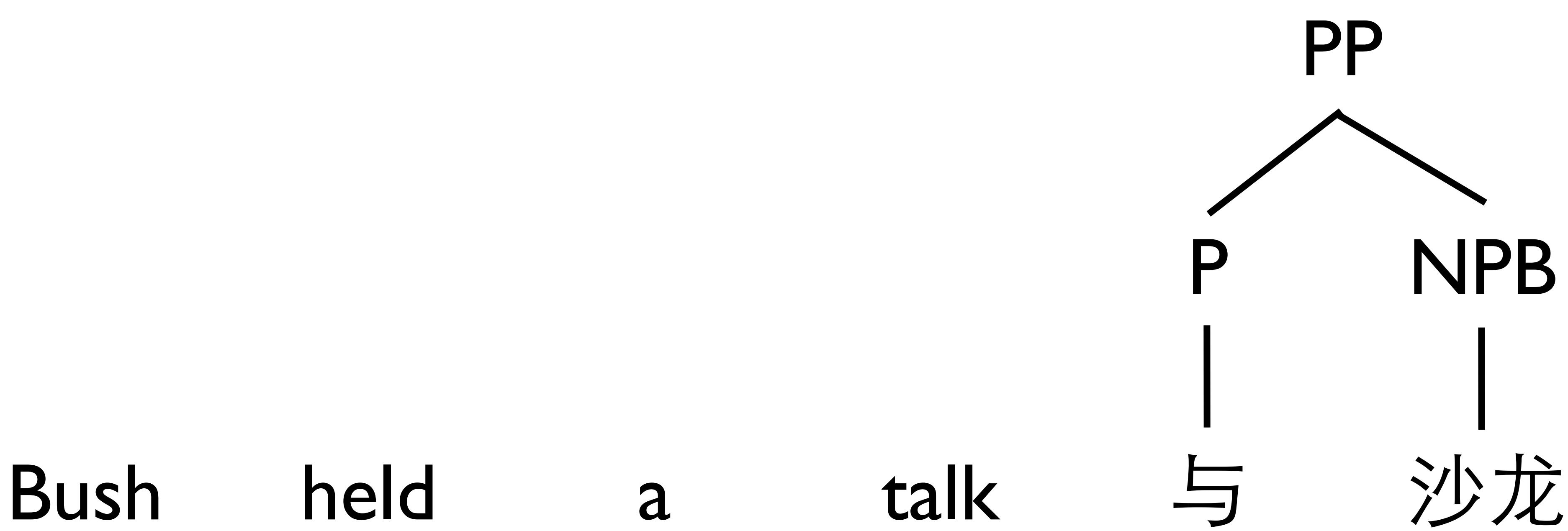
(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



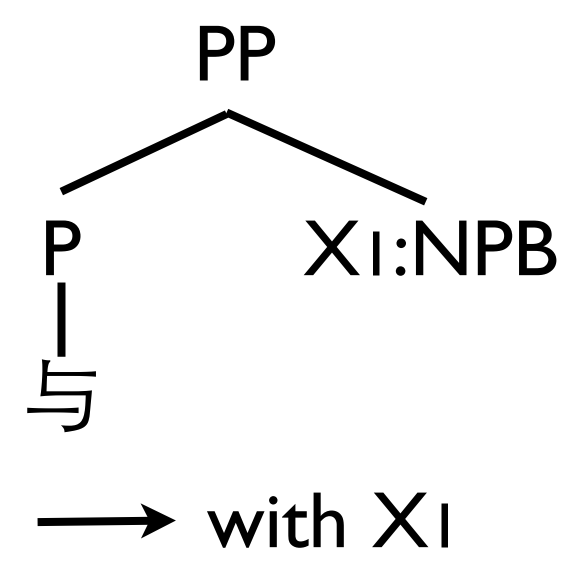
(Liu et al., 2006; Huang et al., 2006)

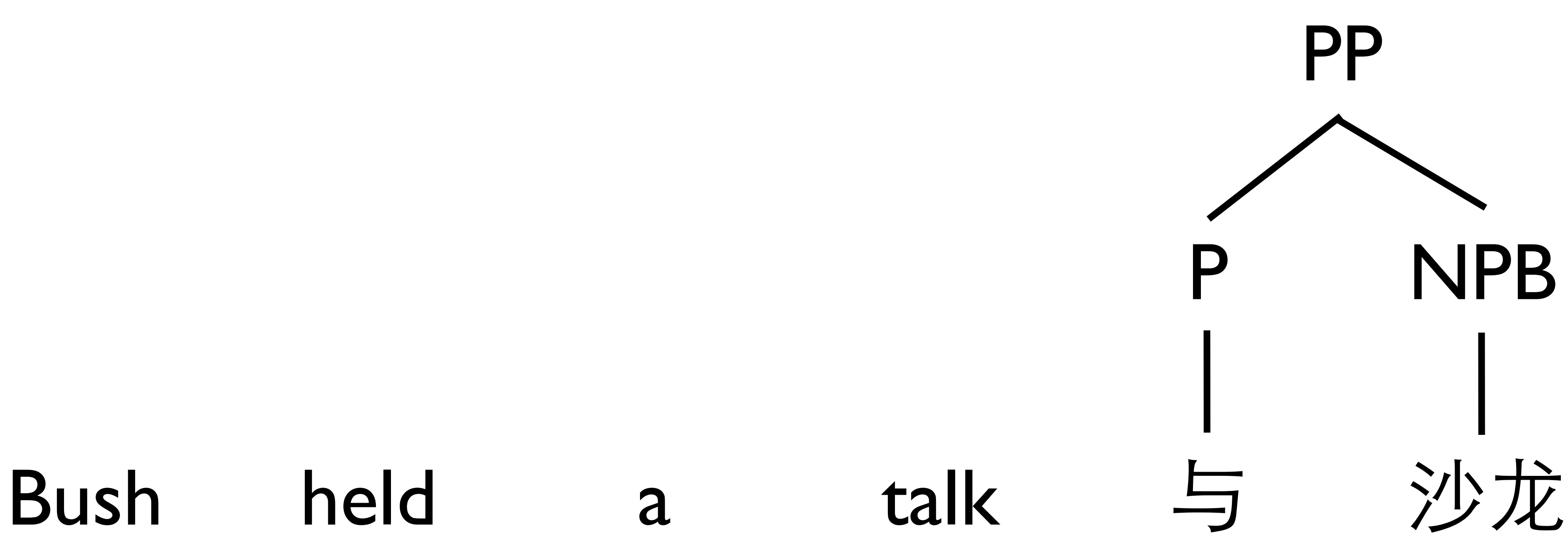
Tree-to-String Translation



(Liu et al., 2006; Huang et al., 2006)

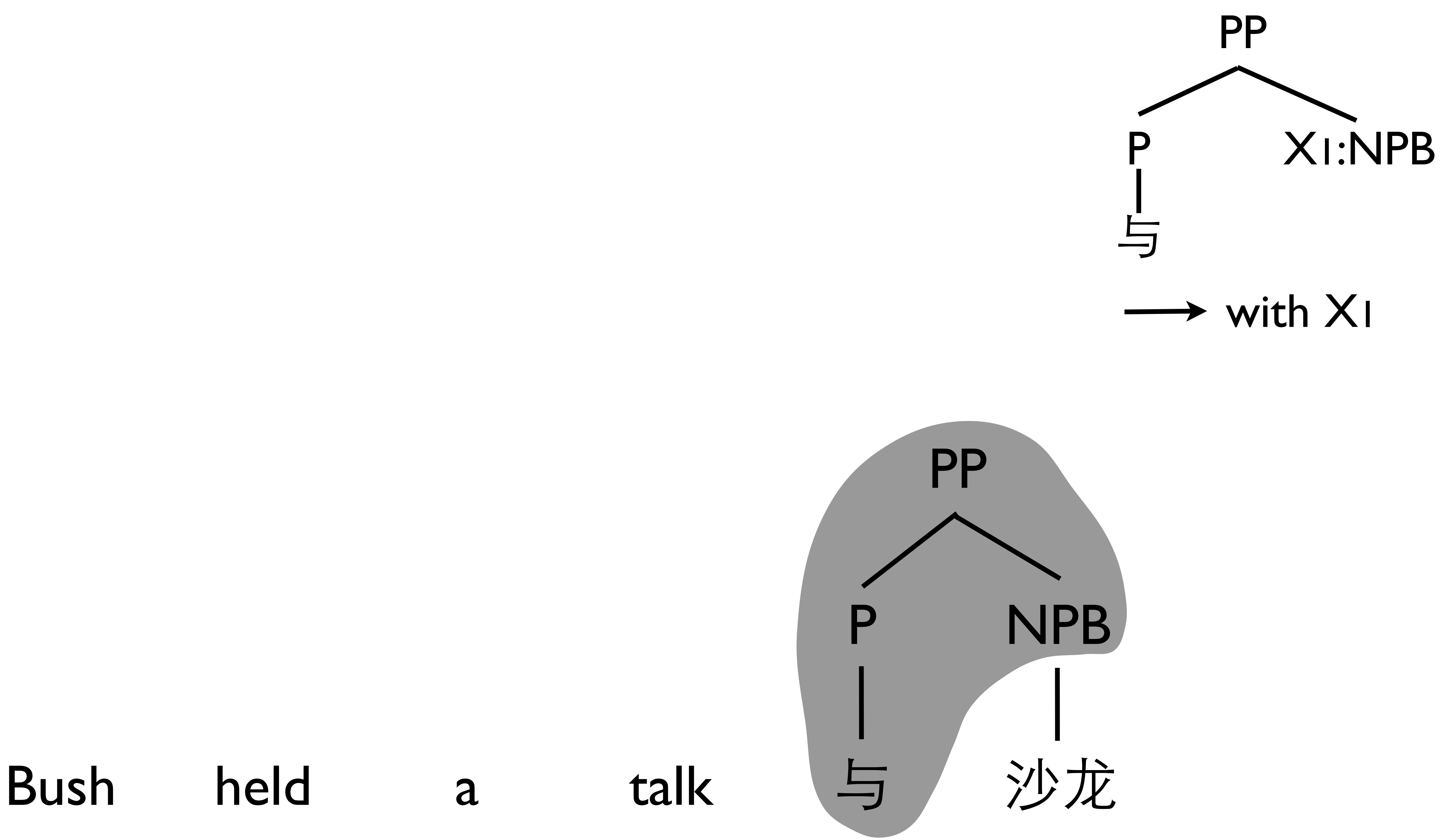
Tree-to-String Translation





(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation



(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation

NPB

Bush held a talk with 沙龙

(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation

NPB
|
沙龙

→ Sharon

NPB
|
Bush held a talk with 沙龙

(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation

NPB
|
沙龙

→ Sharon

Bush held a talk with

NPB
|
沙龙

(Liu et al., 2006; Huang et al., 2006)

Tree-to-String Translation

Bush held a talk with Sharon

(Liu et al., 2006; Huang et al., 2006)

Expressive Power

(Knight and Graehl, 2005)

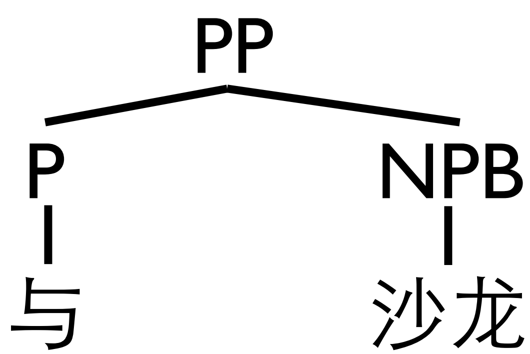
Expressive Power

phrase translation

(Knight and Graehl, 2005)

Expressive Power

phrase translation



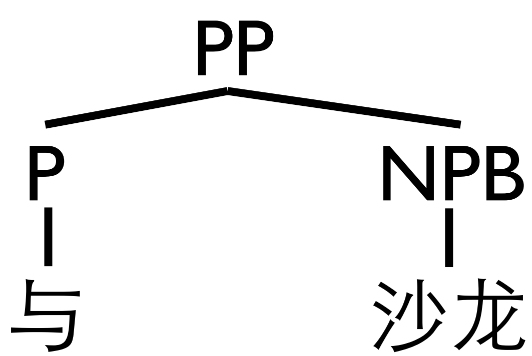
→ with Sharon

(Knight and Graehl, 2005)

Expressive Power

phrase translation

non-constituent

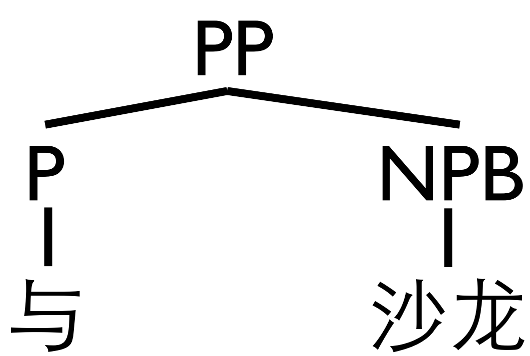


→ with Sharon

(Knight and Graehl, 2005)

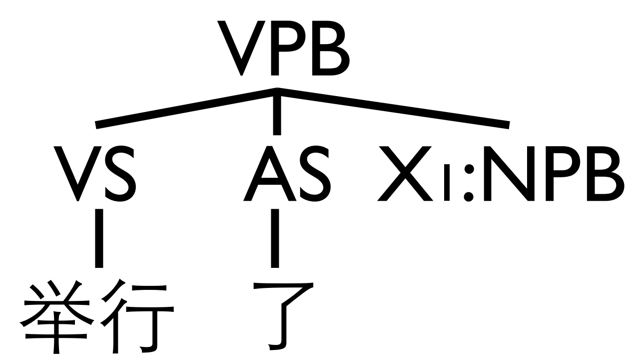
Expressive Power

phrase translation



→ with Sharon

non-constituent

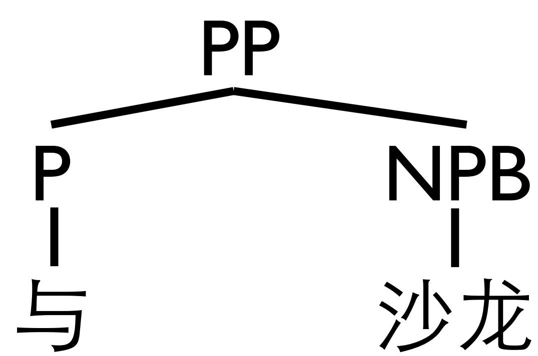


→ held a X_1

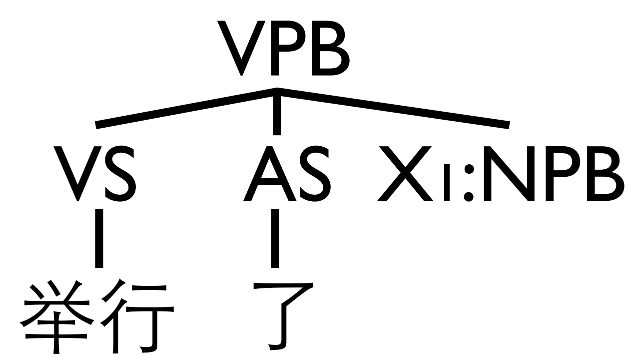
(Knight and Graehl, 2005)

Expressive Power

phrase translation | non-constituent | discontinuous



→ with Sharon

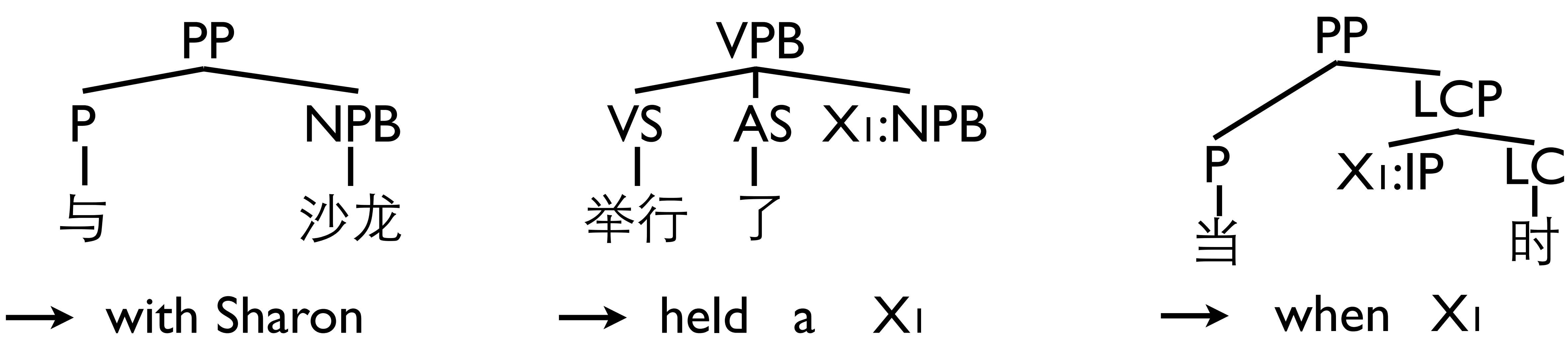


→ held a X_1

(Knight and Graehl, 2005)

Expressive Power

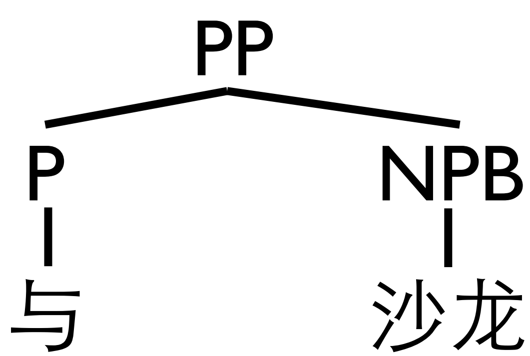
phrase translation | non-constituent | discontinuous



(Knight and Graehl, 2005)

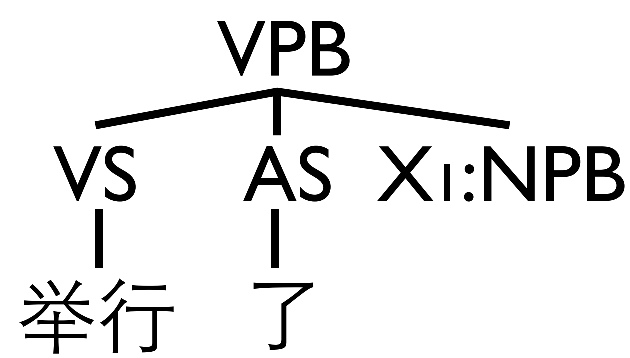
Expressive Power

phrase translation



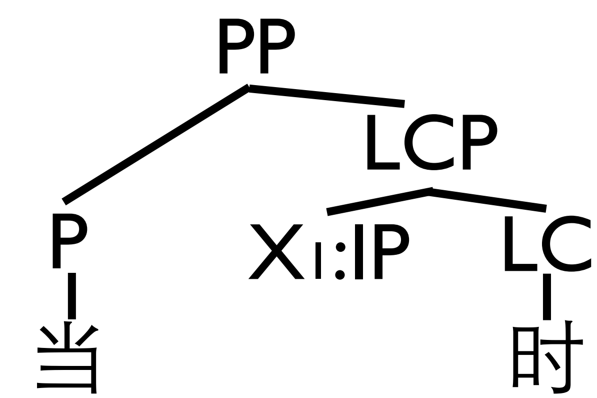
→ with Sharon

non-constituent



→ held a X_1

discontinuous



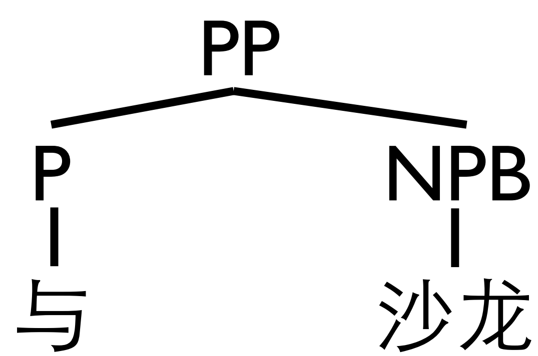
→ when X_1

word deletion

(Knight and Graehl, 2005)

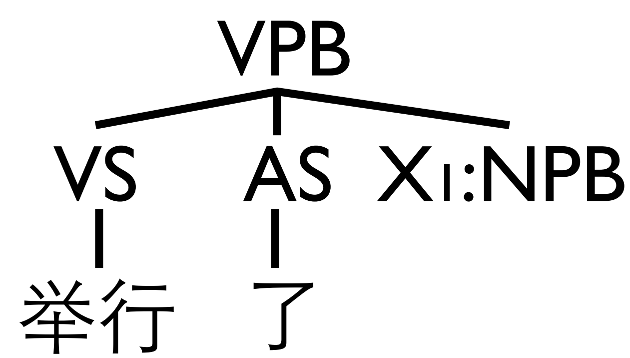
Expressive Power

phrase translation



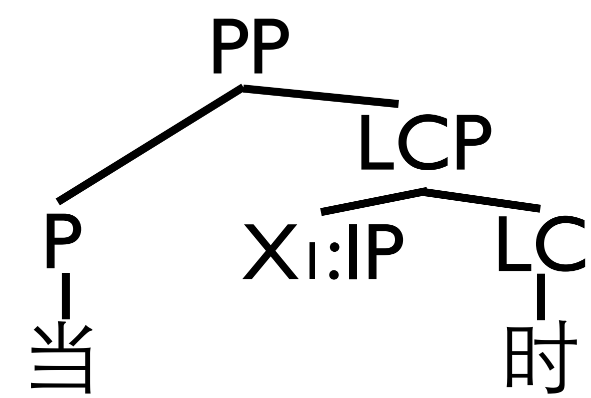
→ with Sharon

non-constituent



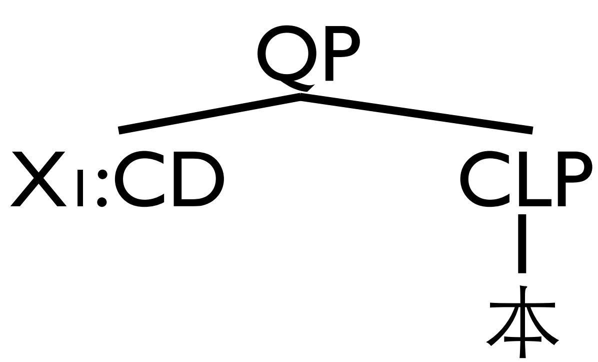
→ held a X1

discontinuous



→ when X1

word deletion

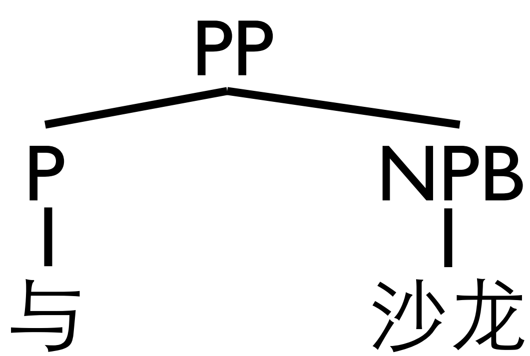


→ X1

(Knight and Graehl, 2005)

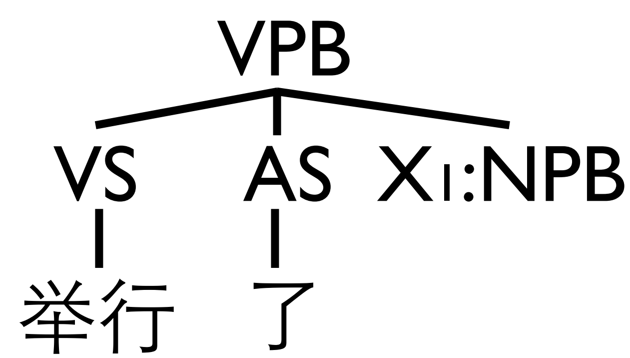
Expressive Power

phrase translation



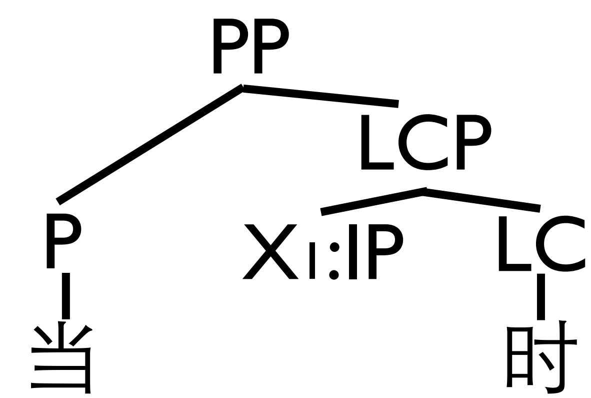
→ with Sharon

non-constituent



→ held a X1

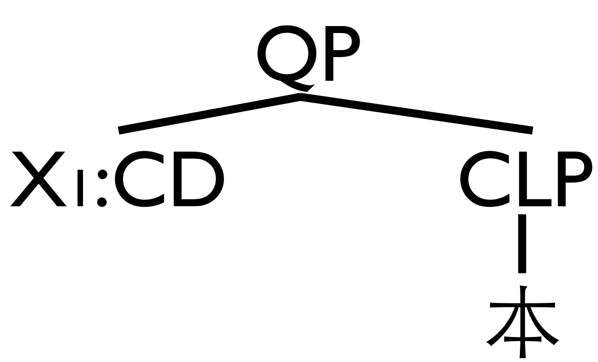
discontinuous



→ when X1

word deletion

multi-level reordering

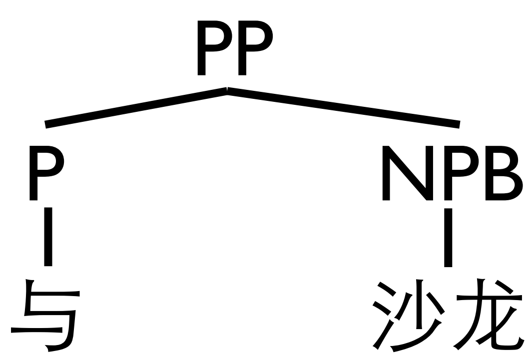


→ X1

(Knight and Graehl, 2005)

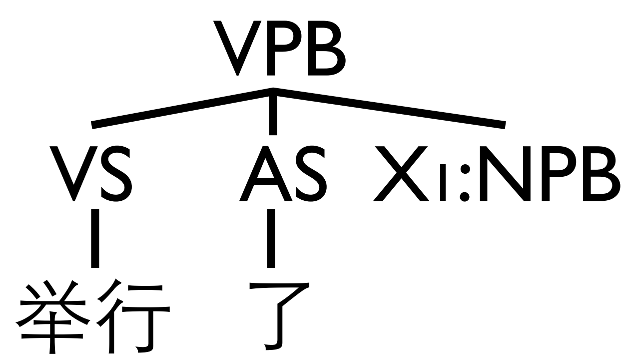
Expressive Power

phrase translation



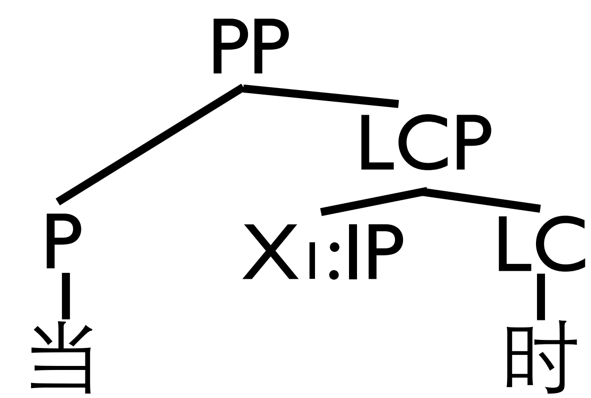
→ with Sharon

non-constituent



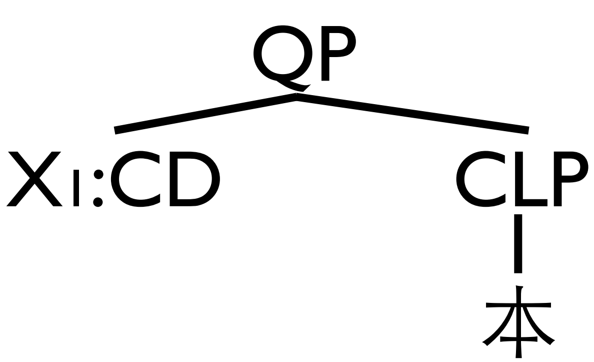
→ held a X_1

discontinuous



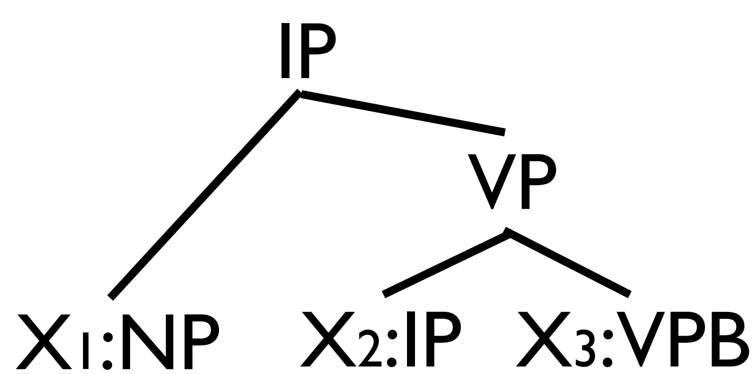
→ when X_1

word deletion



→ X_1

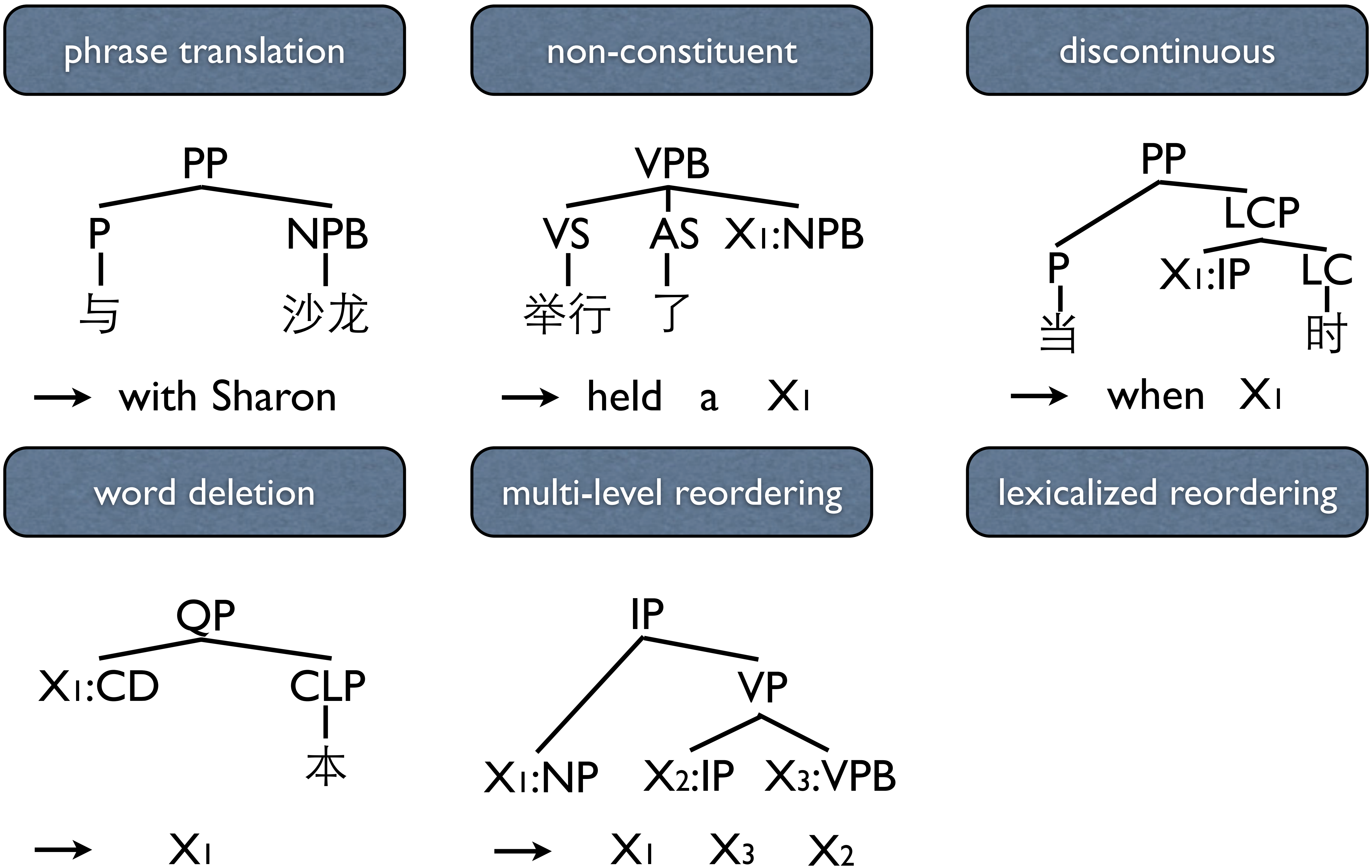
multi-level reordering



→ X_1 X_3 X_2

(Knight and Graehl, 2005)

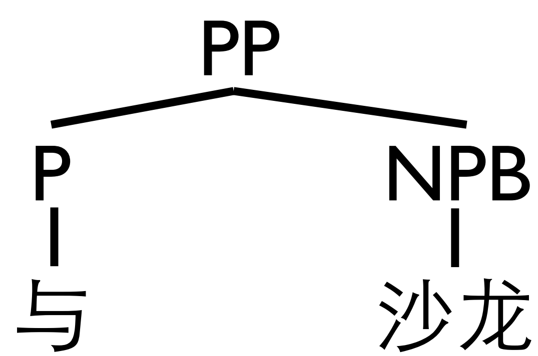
Expressive Power



(Knight and Graehl, 2005)

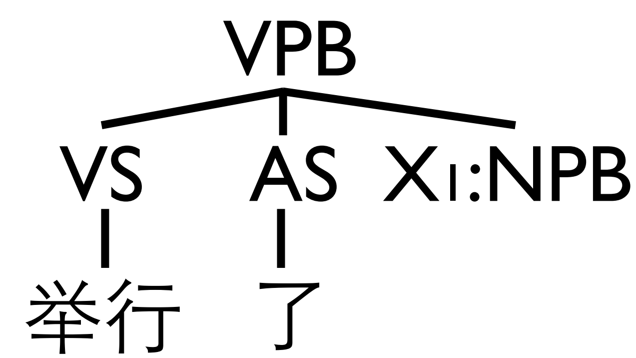
Expressive Power

phrase translation



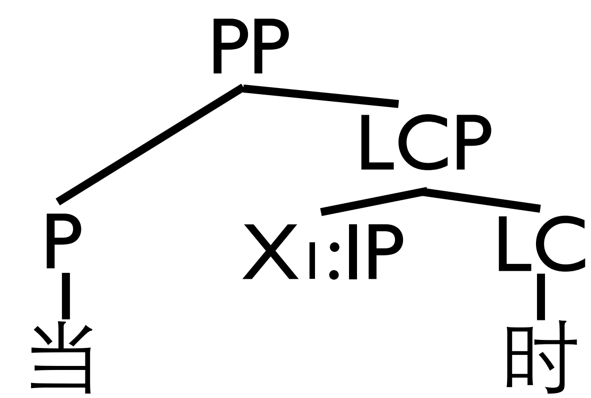
→ with Sharon

non-constituent



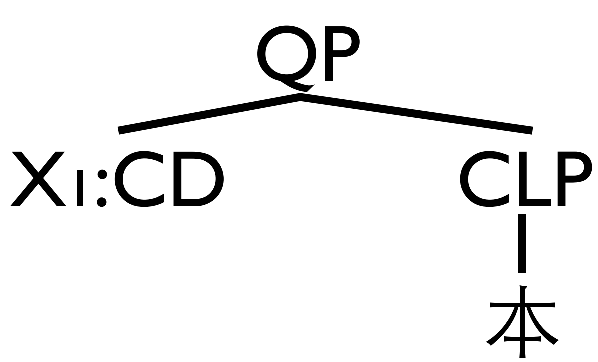
→ held a X_1

discontinuous



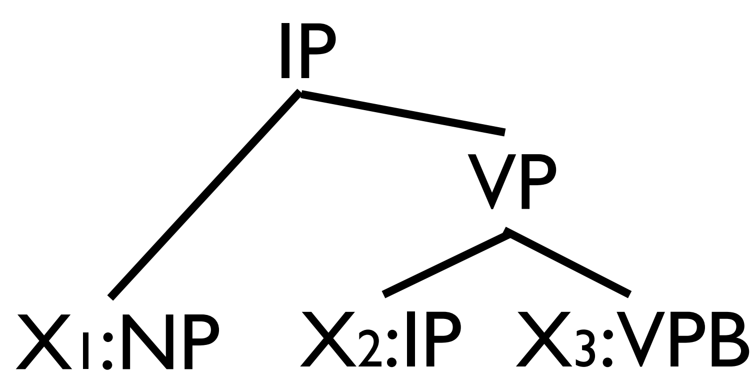
→ when X_1

word deletion



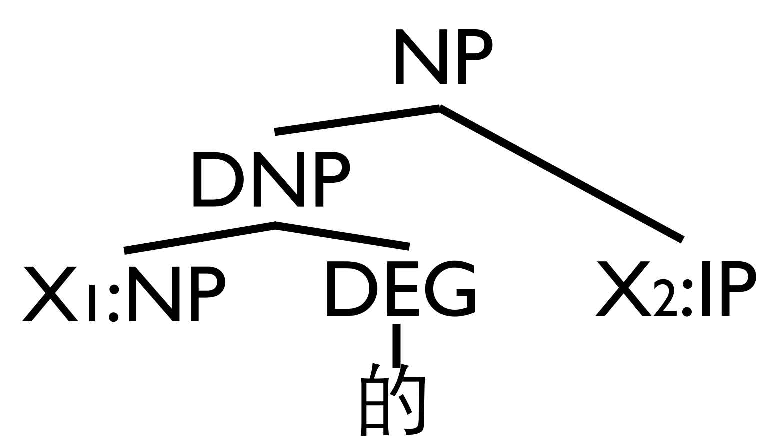
→ X_1

multi-level reordering



→ X_1 X_3 X_2

lexicalized reordering



→ X_2 of X_1

(Knight and Graehl, 2005)

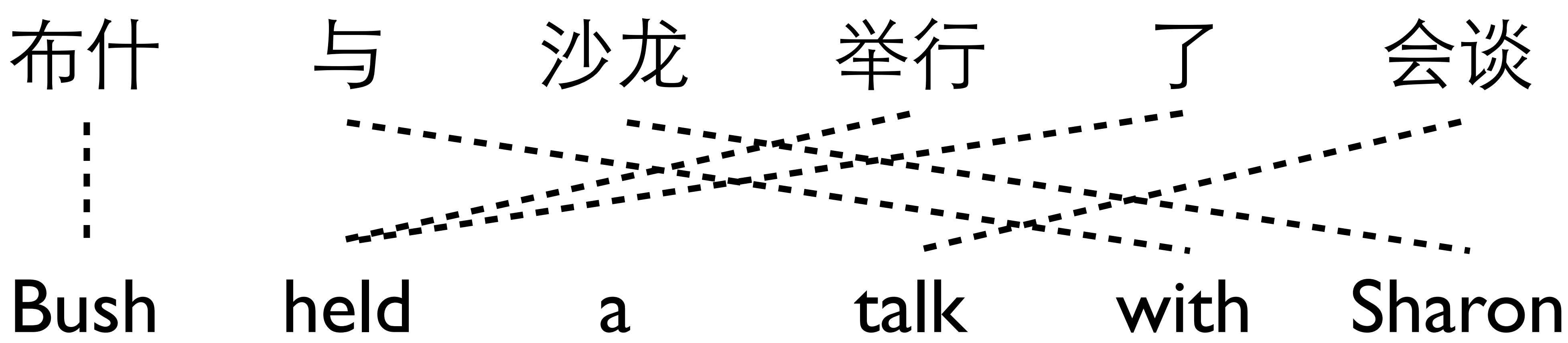
Rule Extraction

布什 与 沙龙 举行 了 会谈

Bush held a talk with Sharon

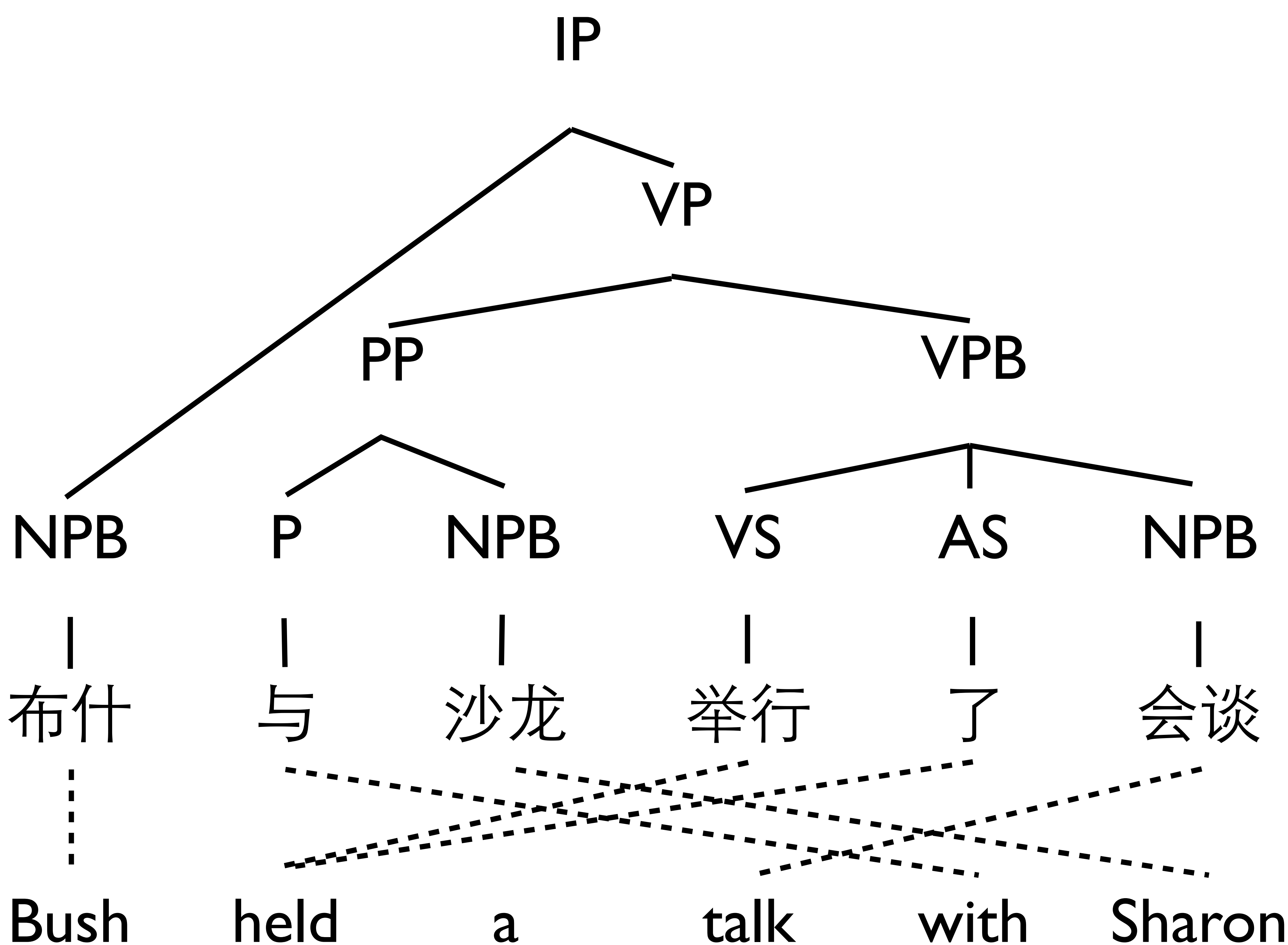
(Galley et al., 2004)

Rule Extraction



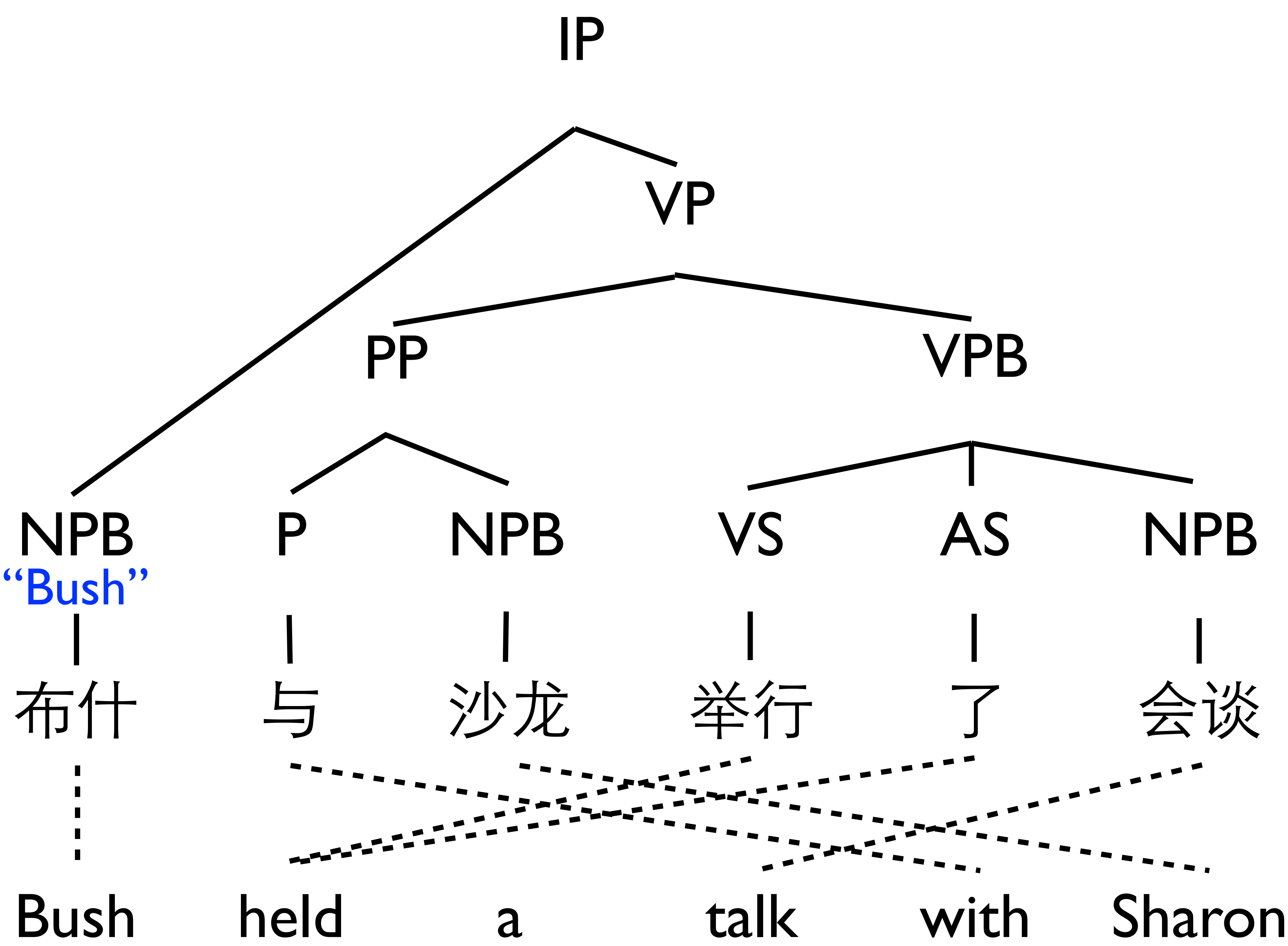
(Galley et al., 2004)

Rule Extraction



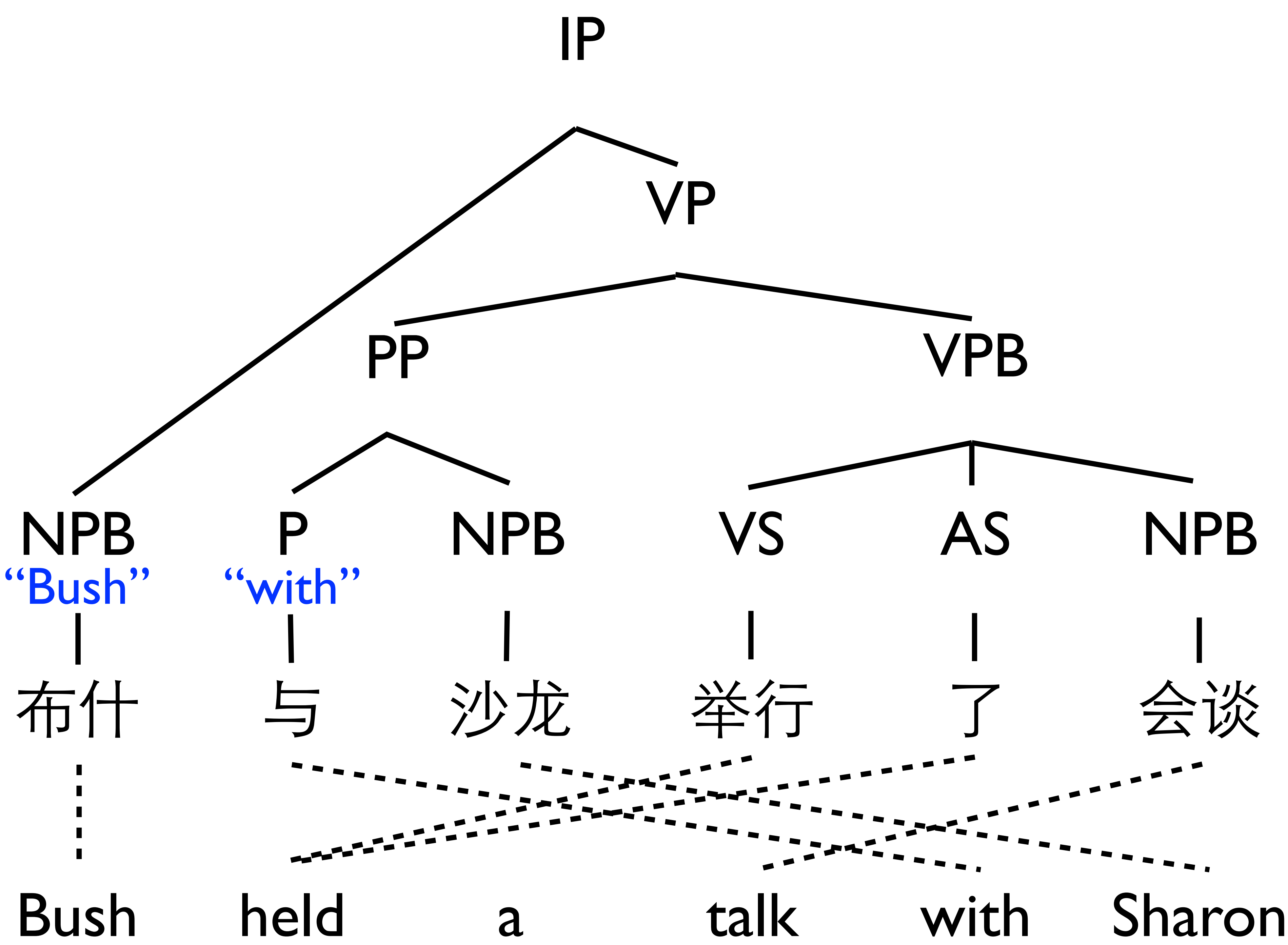
(Galley et al., 2004)

Rule Extraction



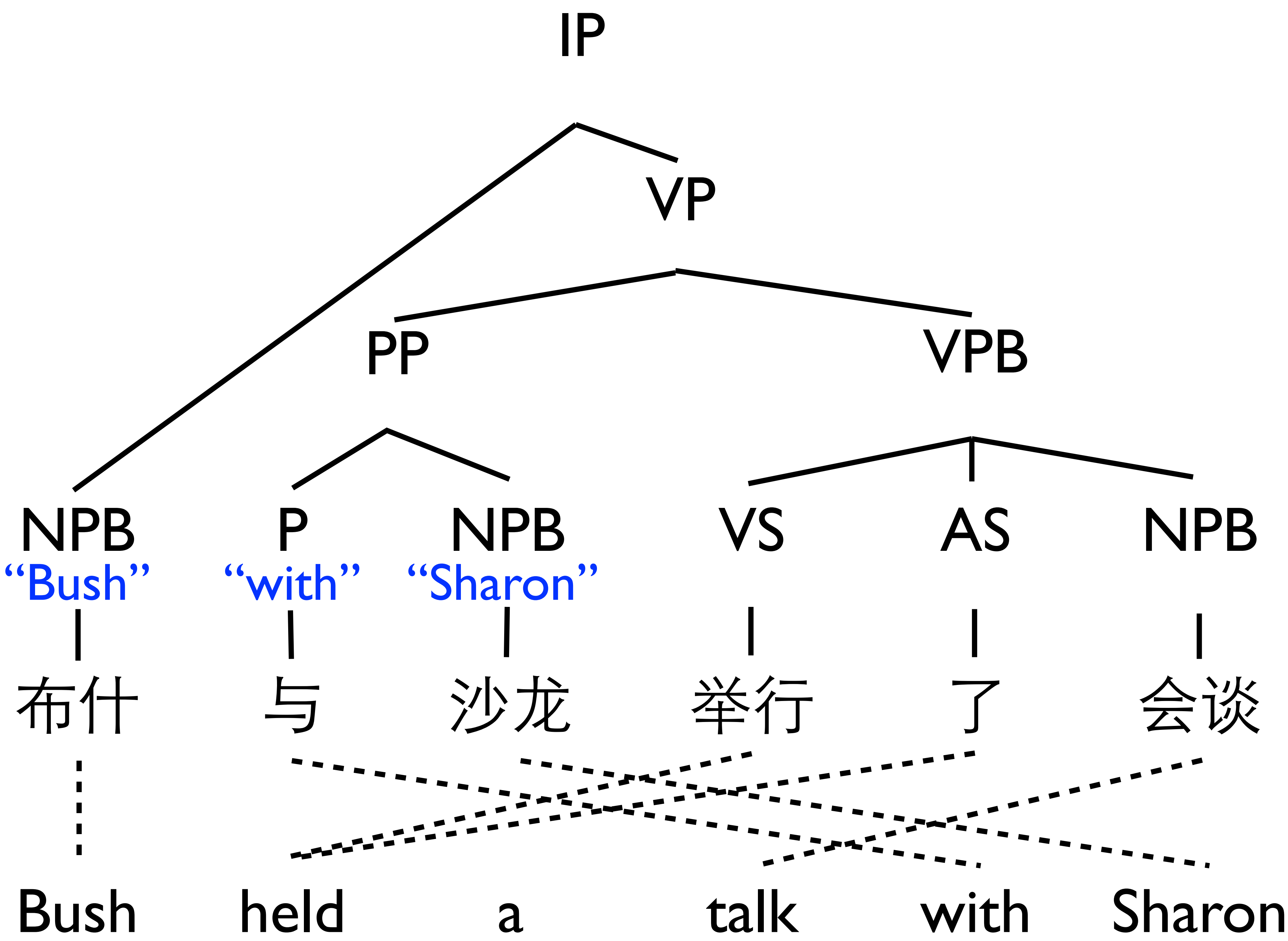
(Galley et al., 2004)

Rule Extraction



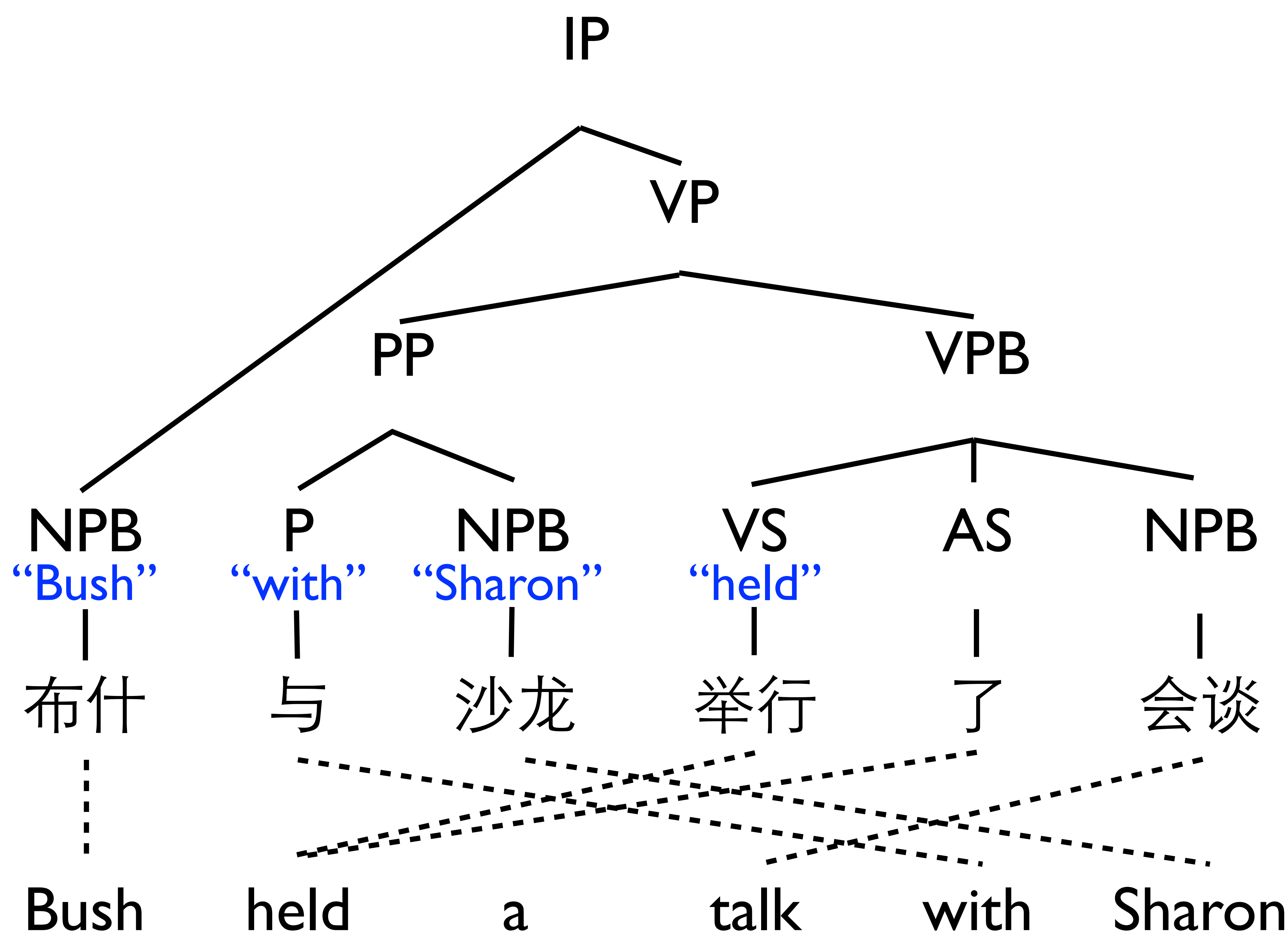
(Galley et al., 2004)

Rule Extraction



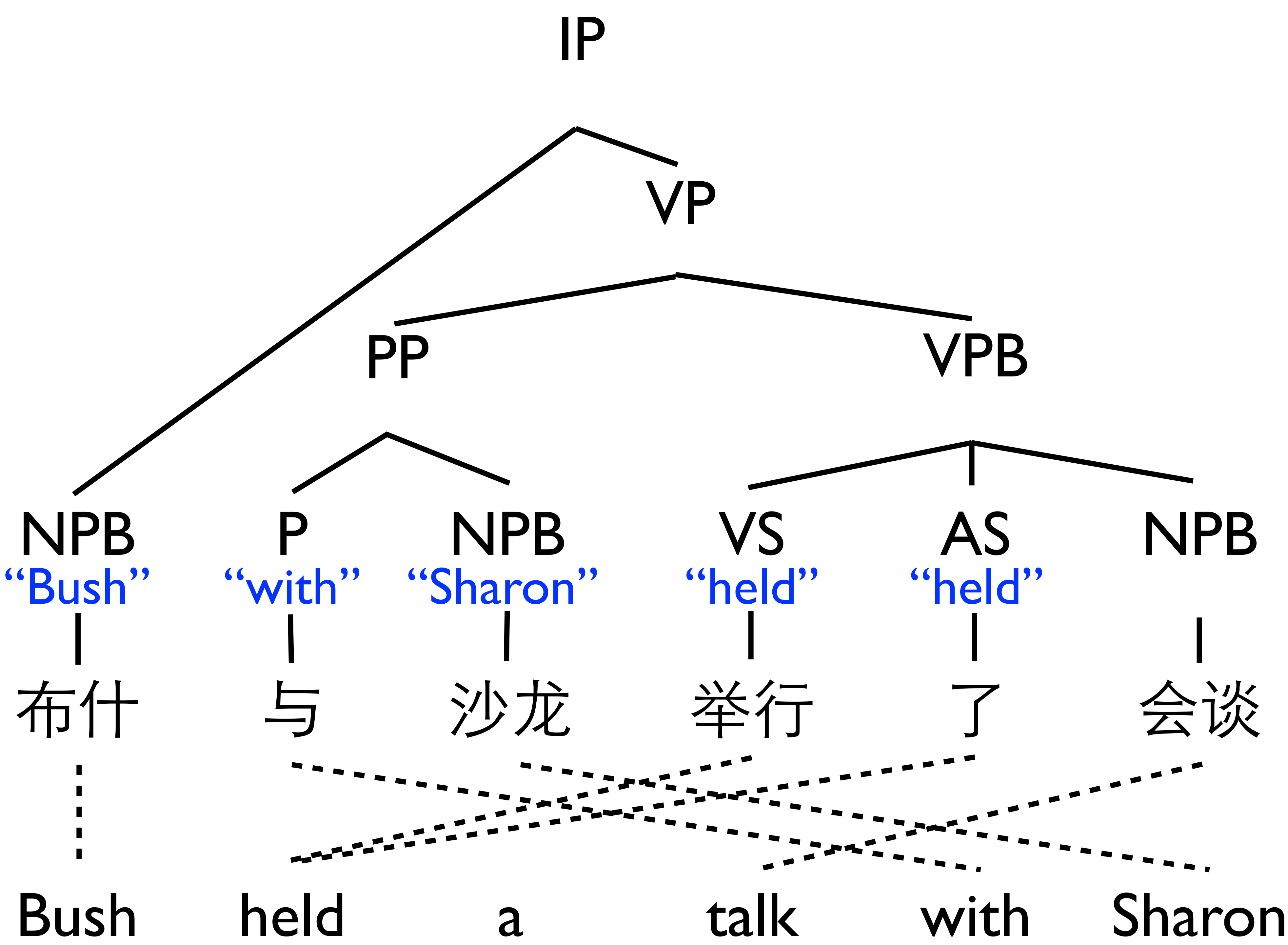
(Galley et al., 2004)

Rule Extraction



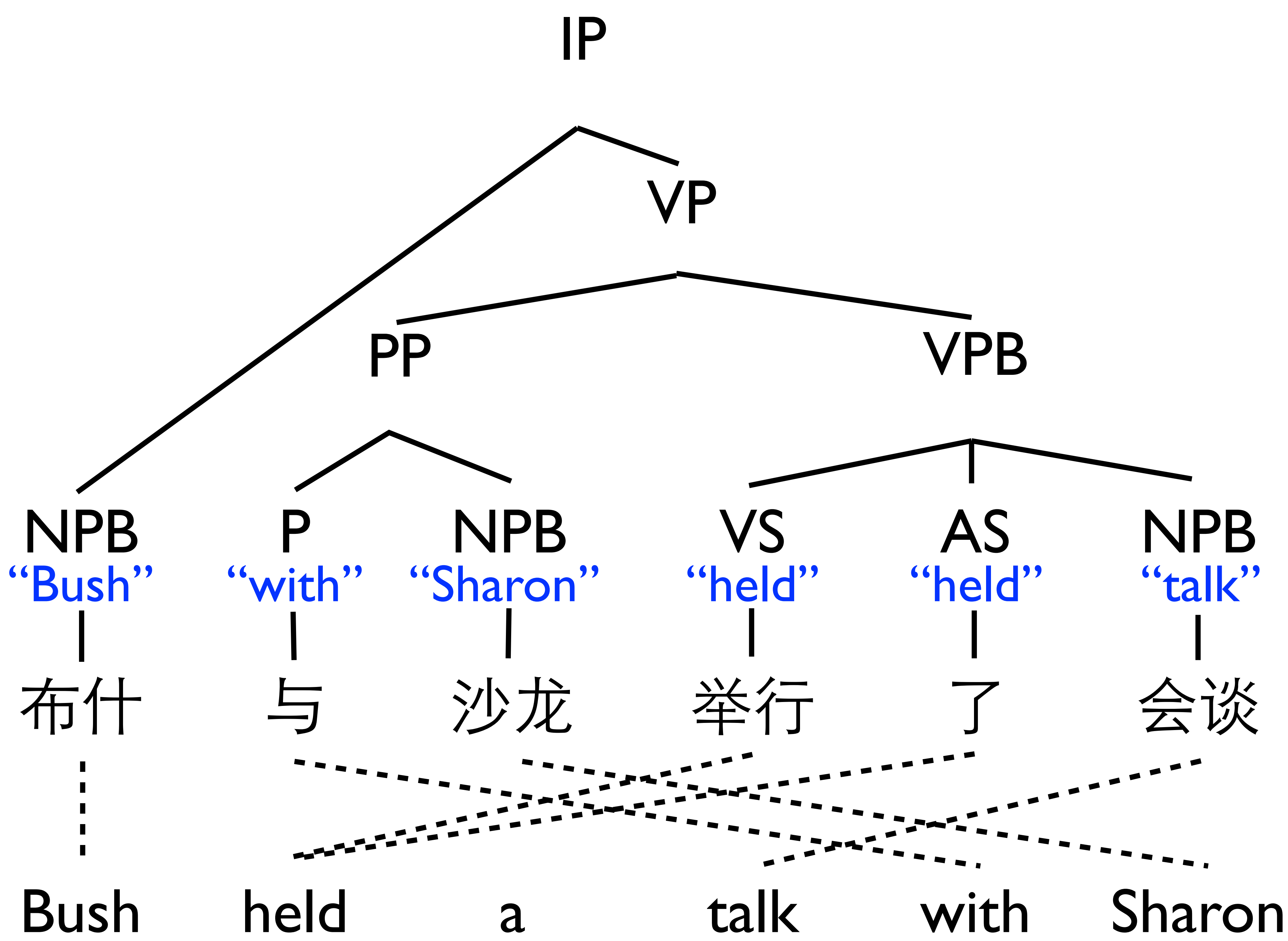
(Galley et al., 2004)

Rule Extraction



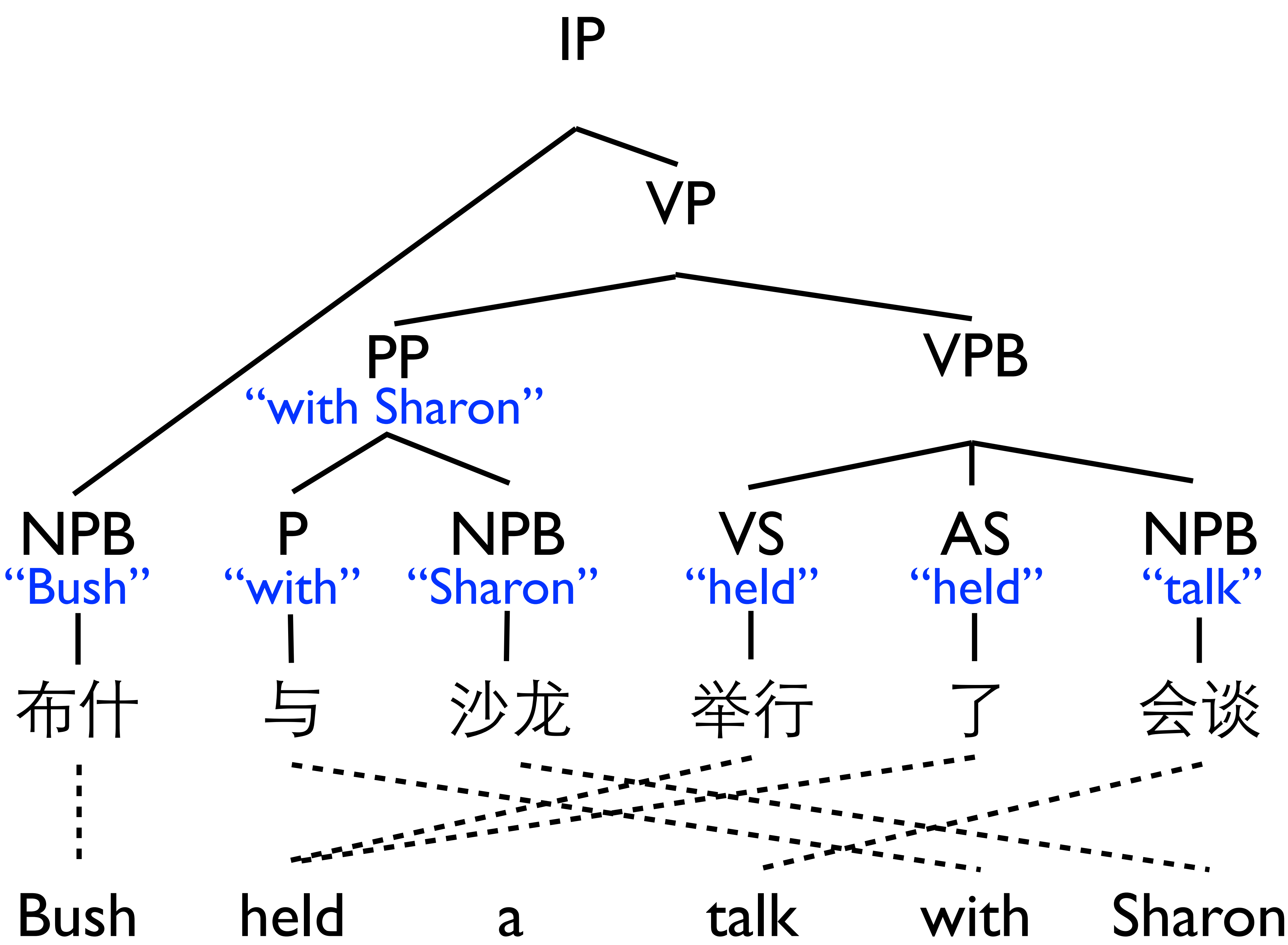
(Galley et al., 2004)

Rule Extraction



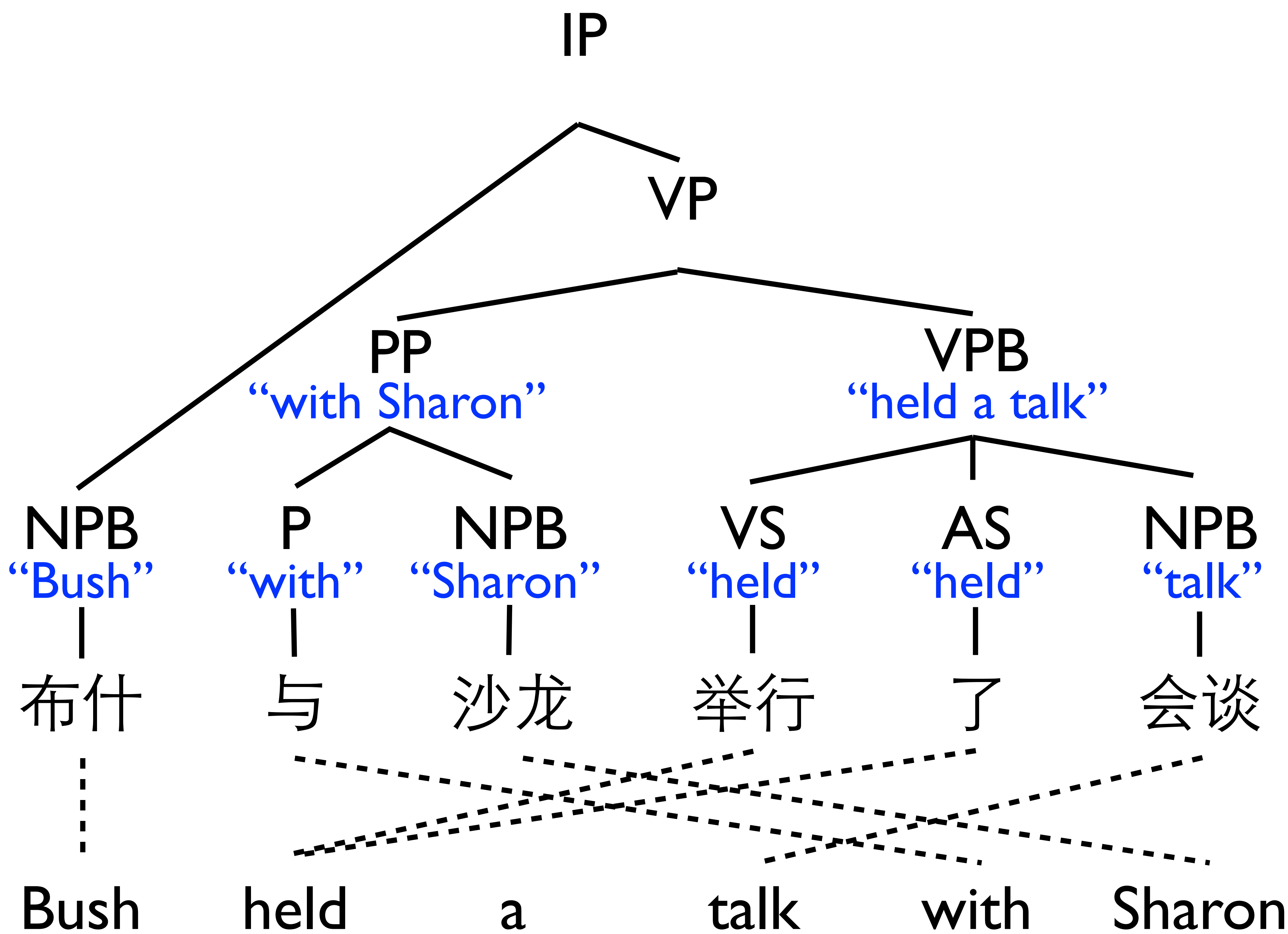
(Galley et al., 2004)

Rule Extraction



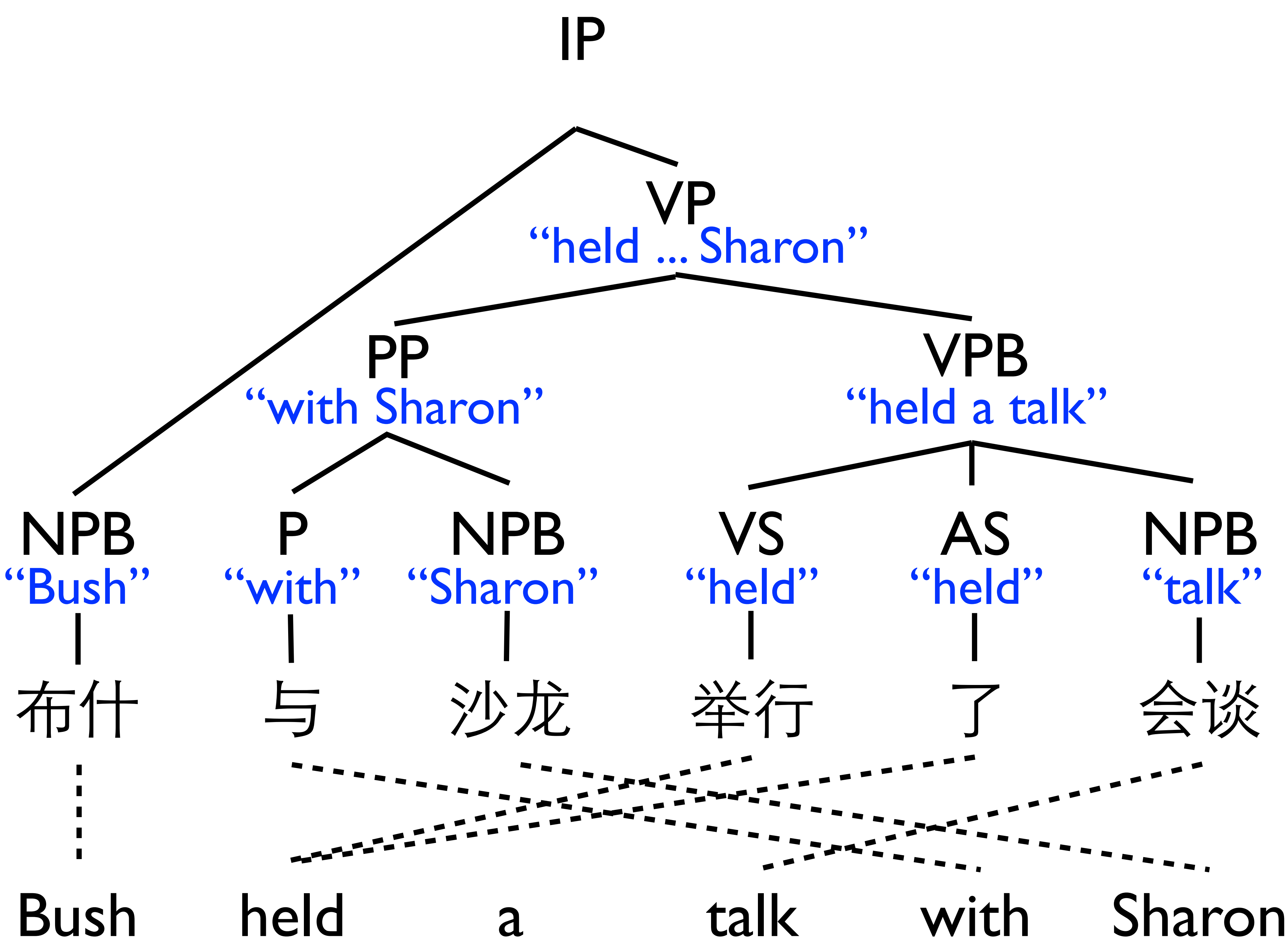
(Galley et al., 2004)

Rule Extraction



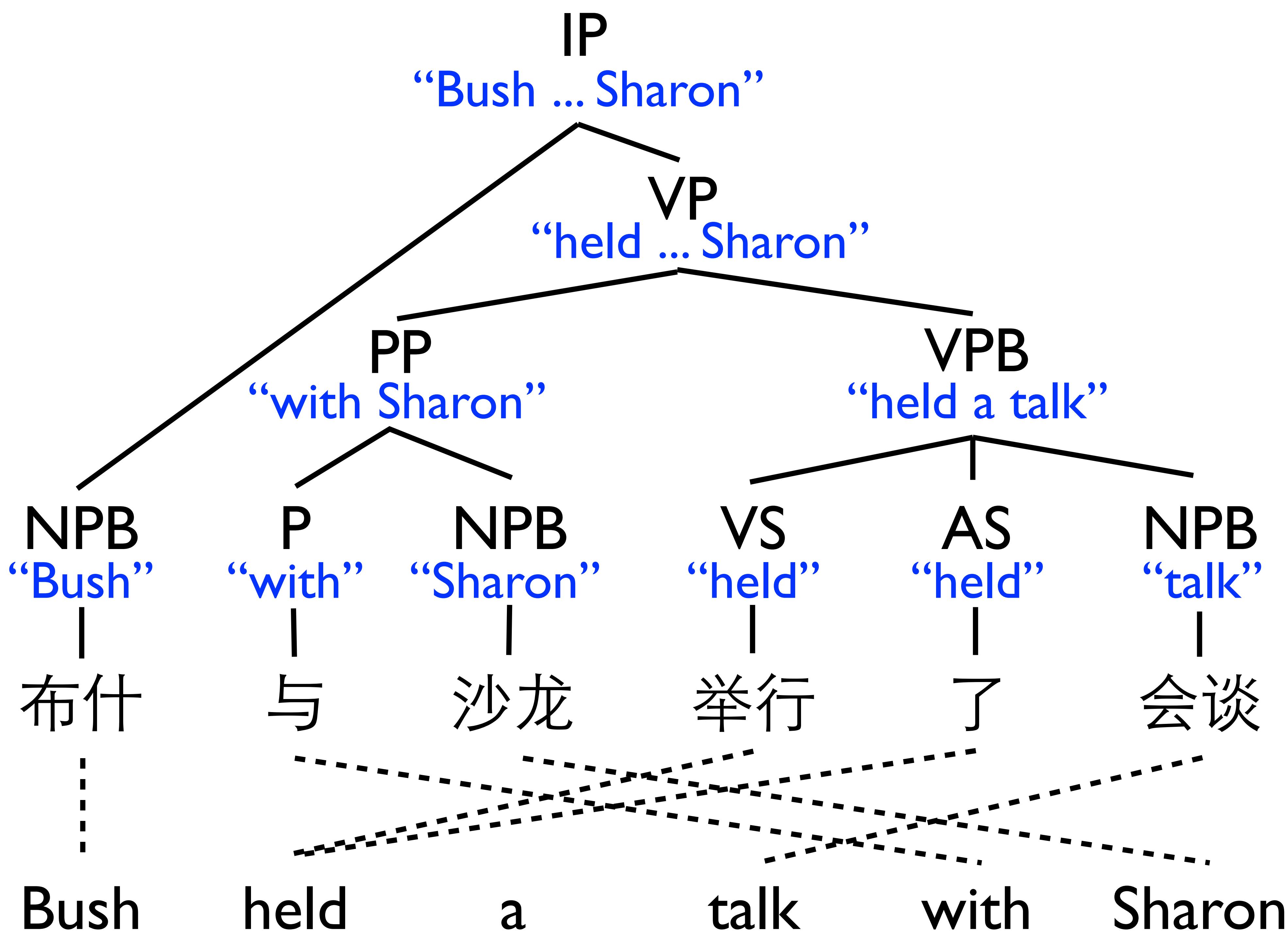
(Galley et al., 2004)

Rule Extraction



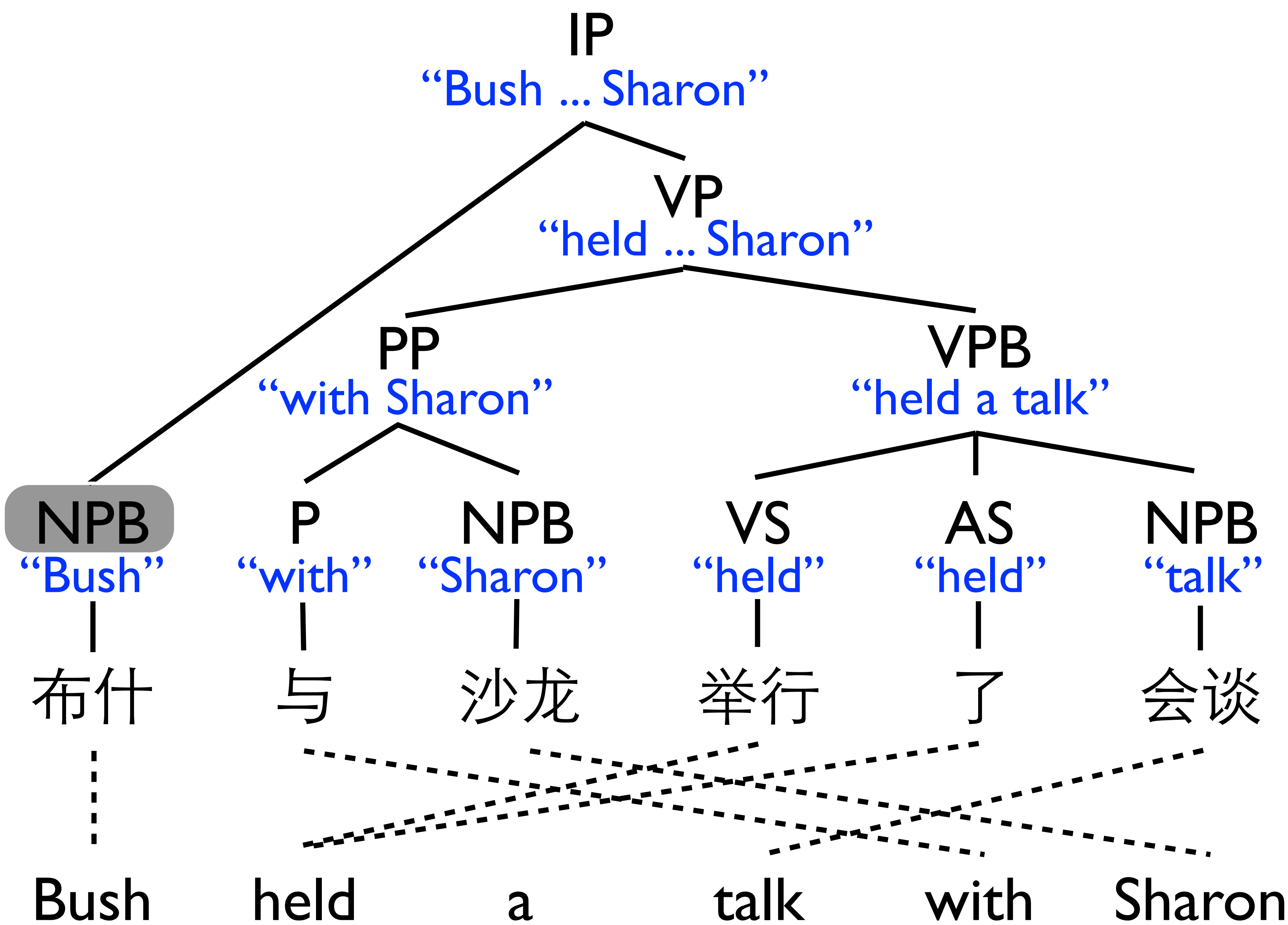
(Galley et al., 2004)

Rule Extraction



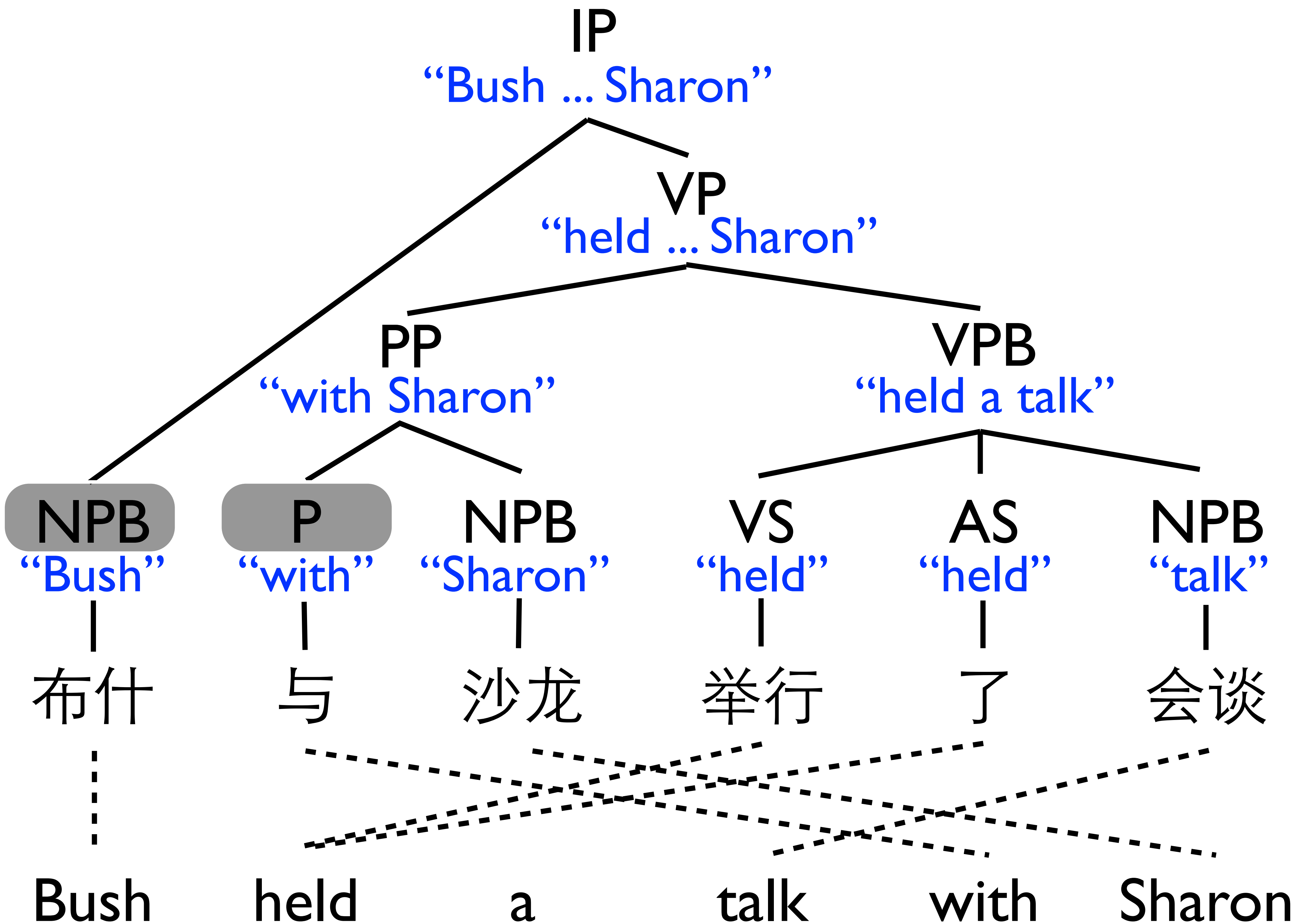
(Galley et al., 2004)

Rule Extraction



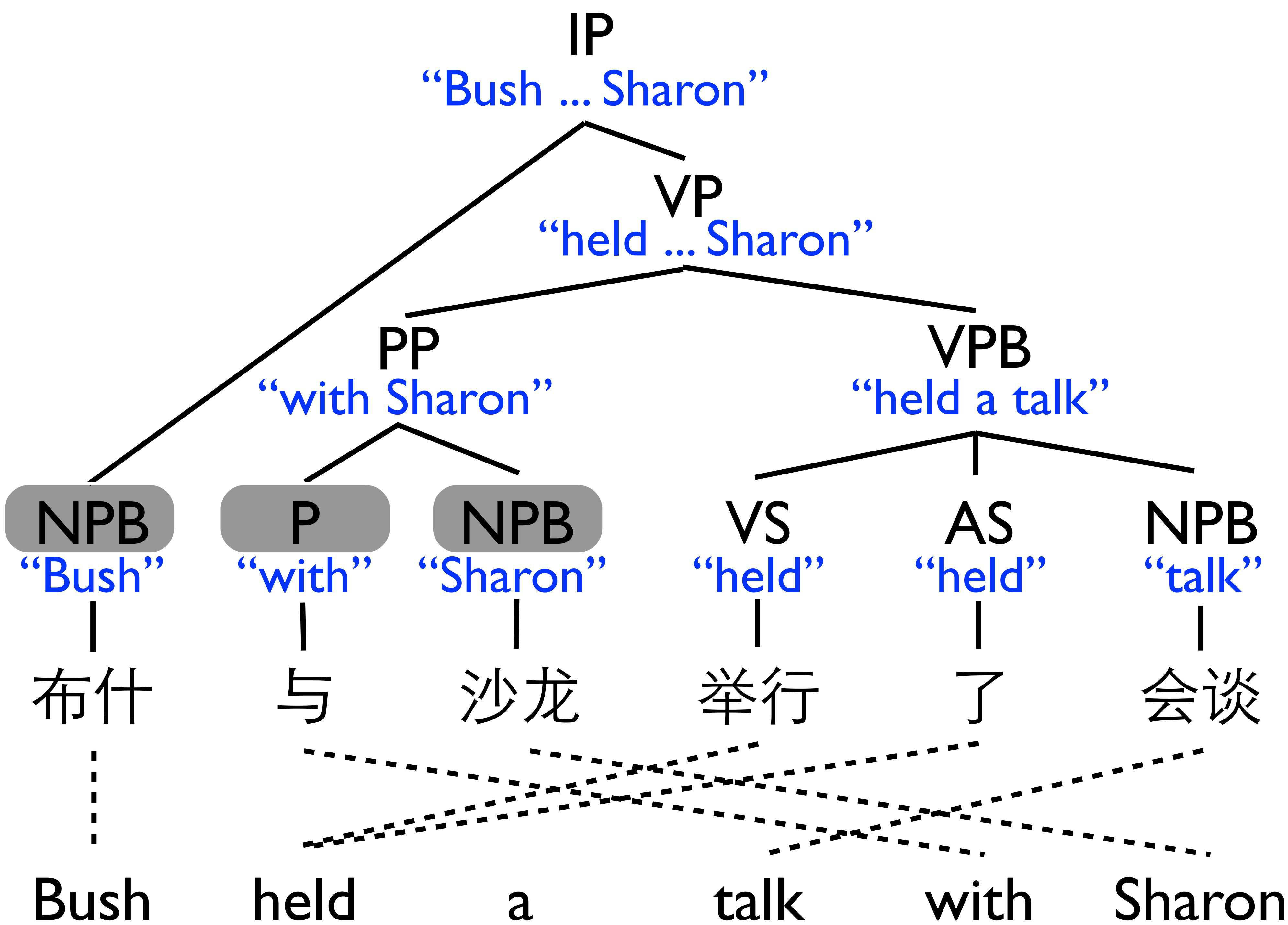
(Galley et al., 2004)

Rule Extraction



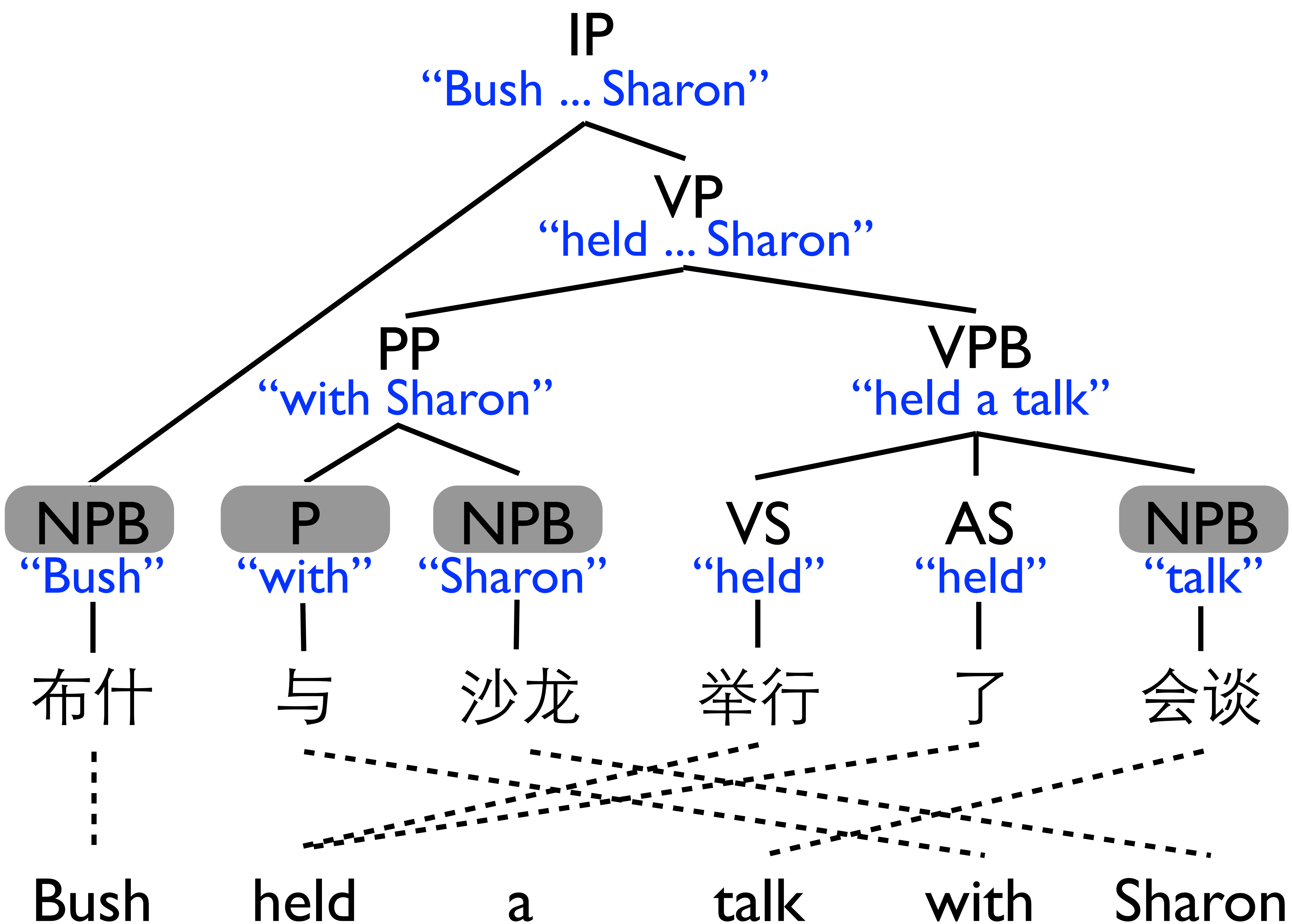
(Galley et al., 2004)

Rule Extraction



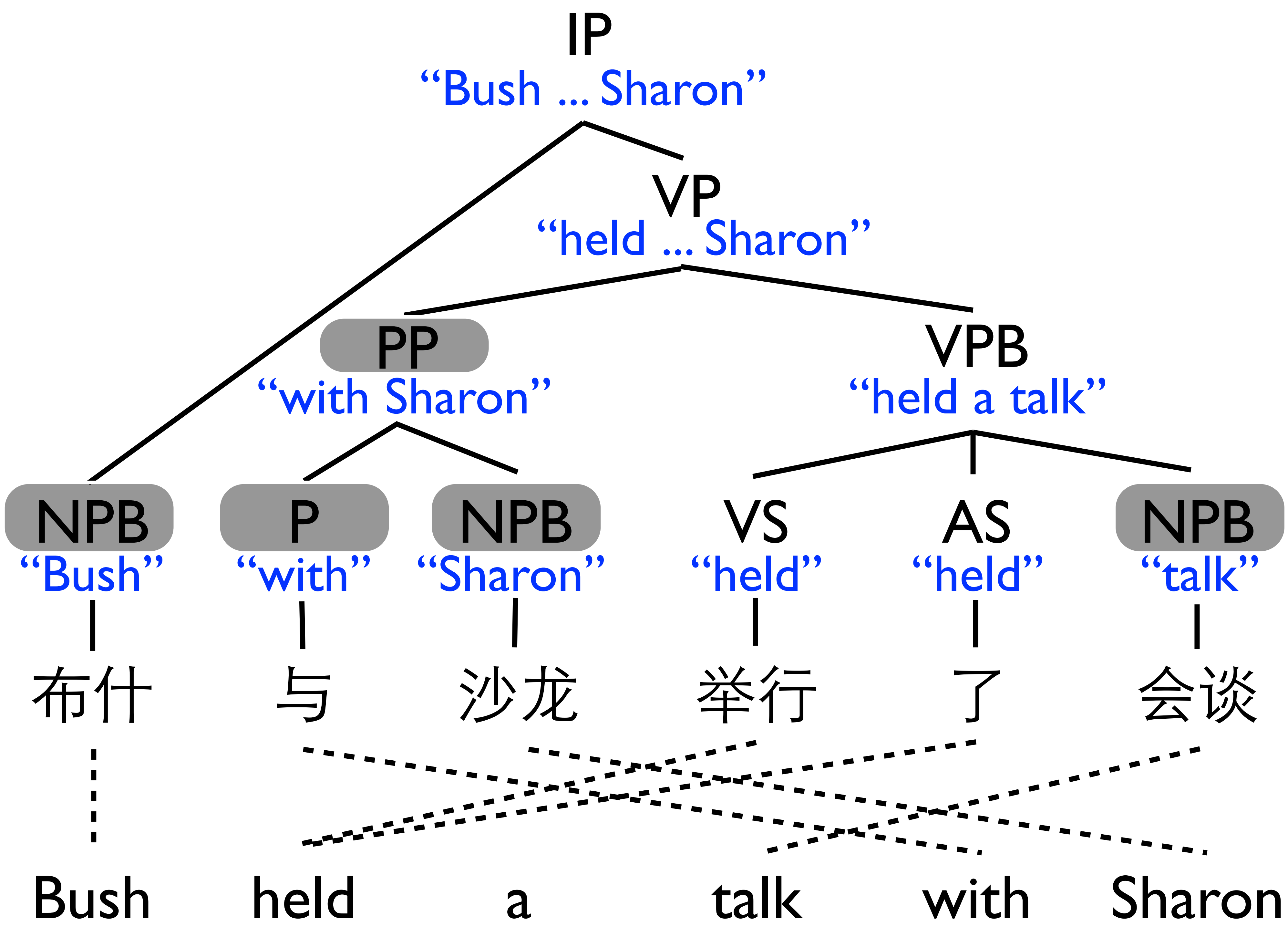
(Galley et al., 2004)

Rule Extraction



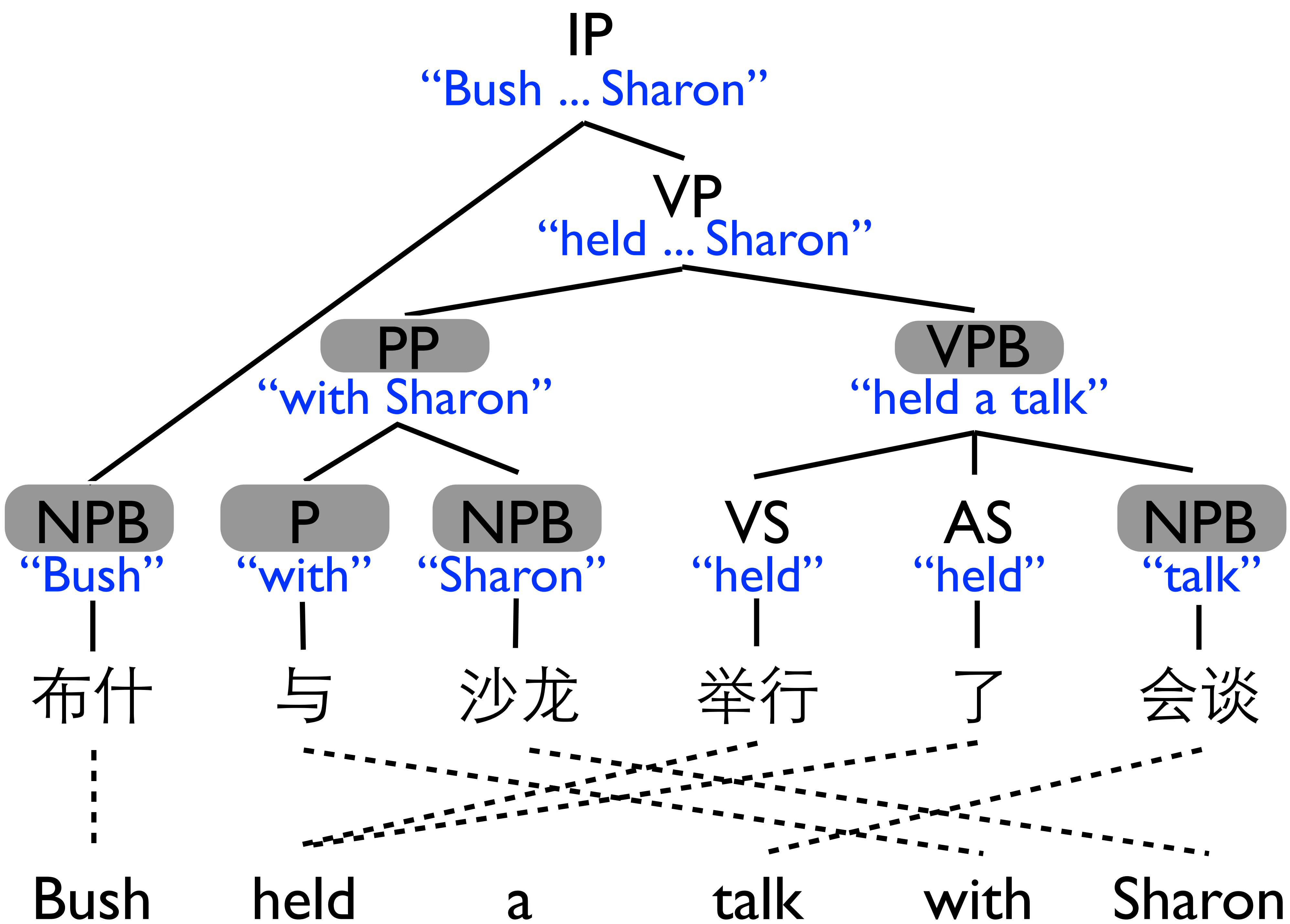
(Galley et al., 2004)

Rule Extraction



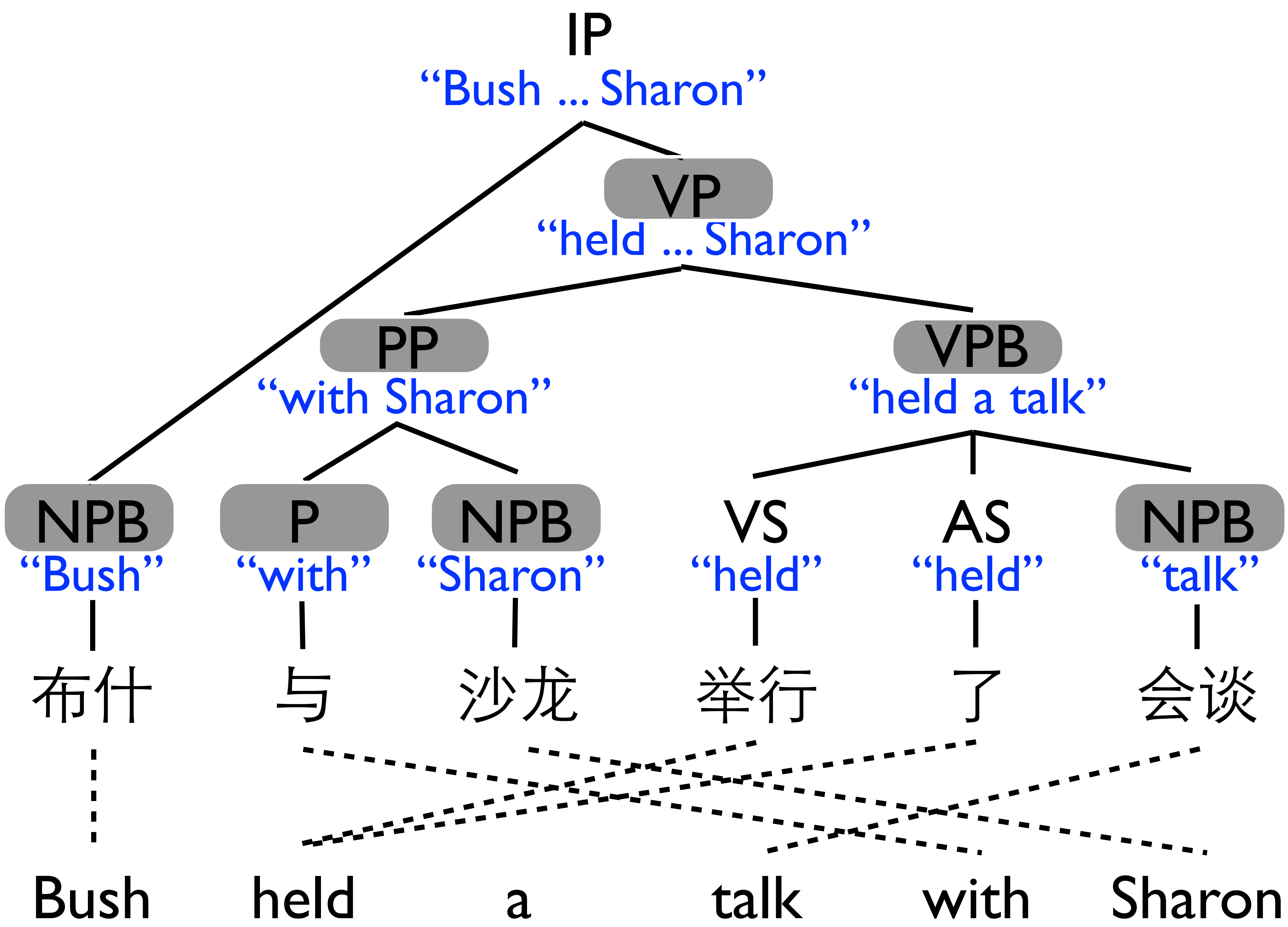
(Galley et al., 2004)

Rule Extraction



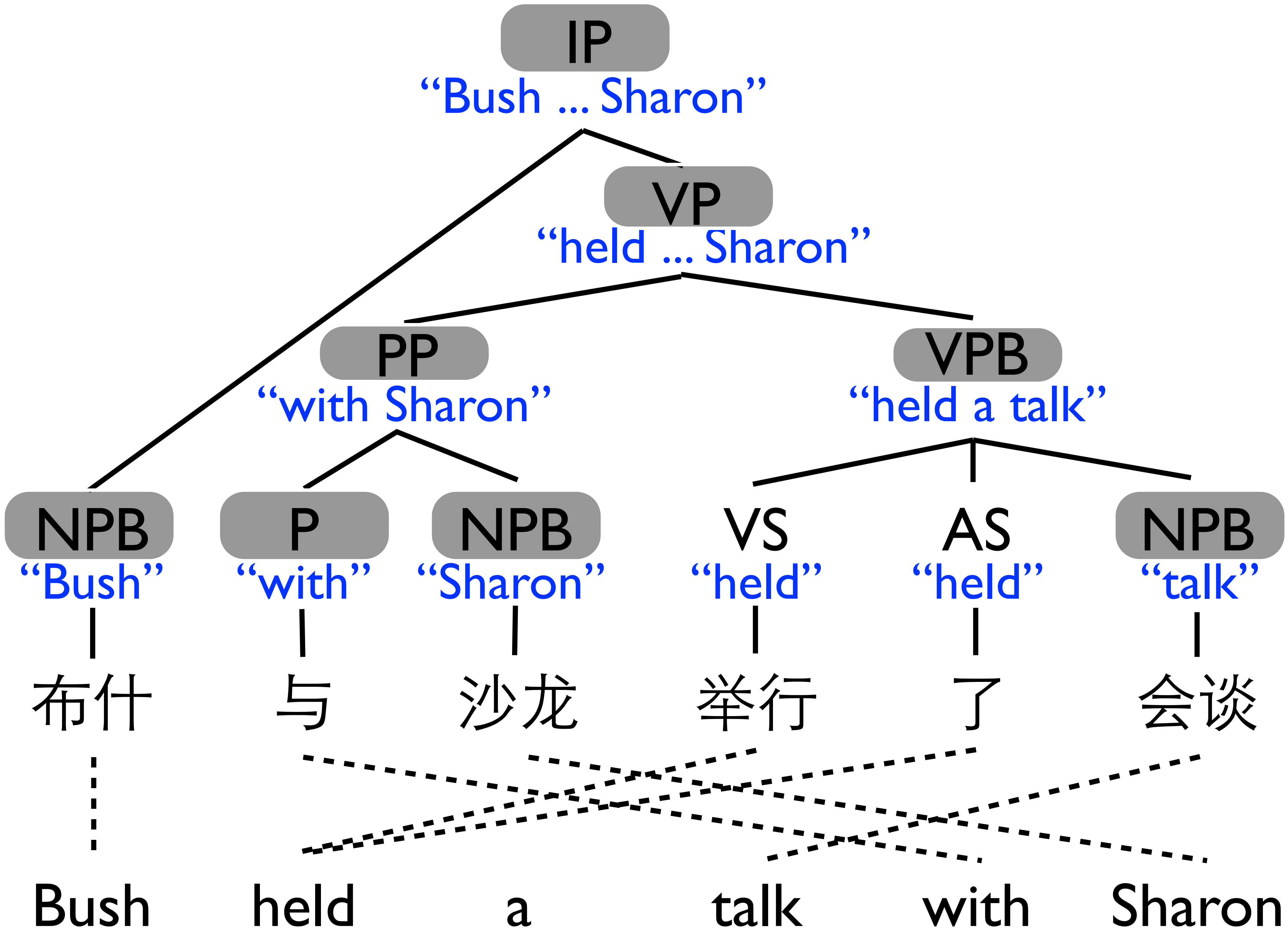
(Galley et al., 2004)

Rule Extraction



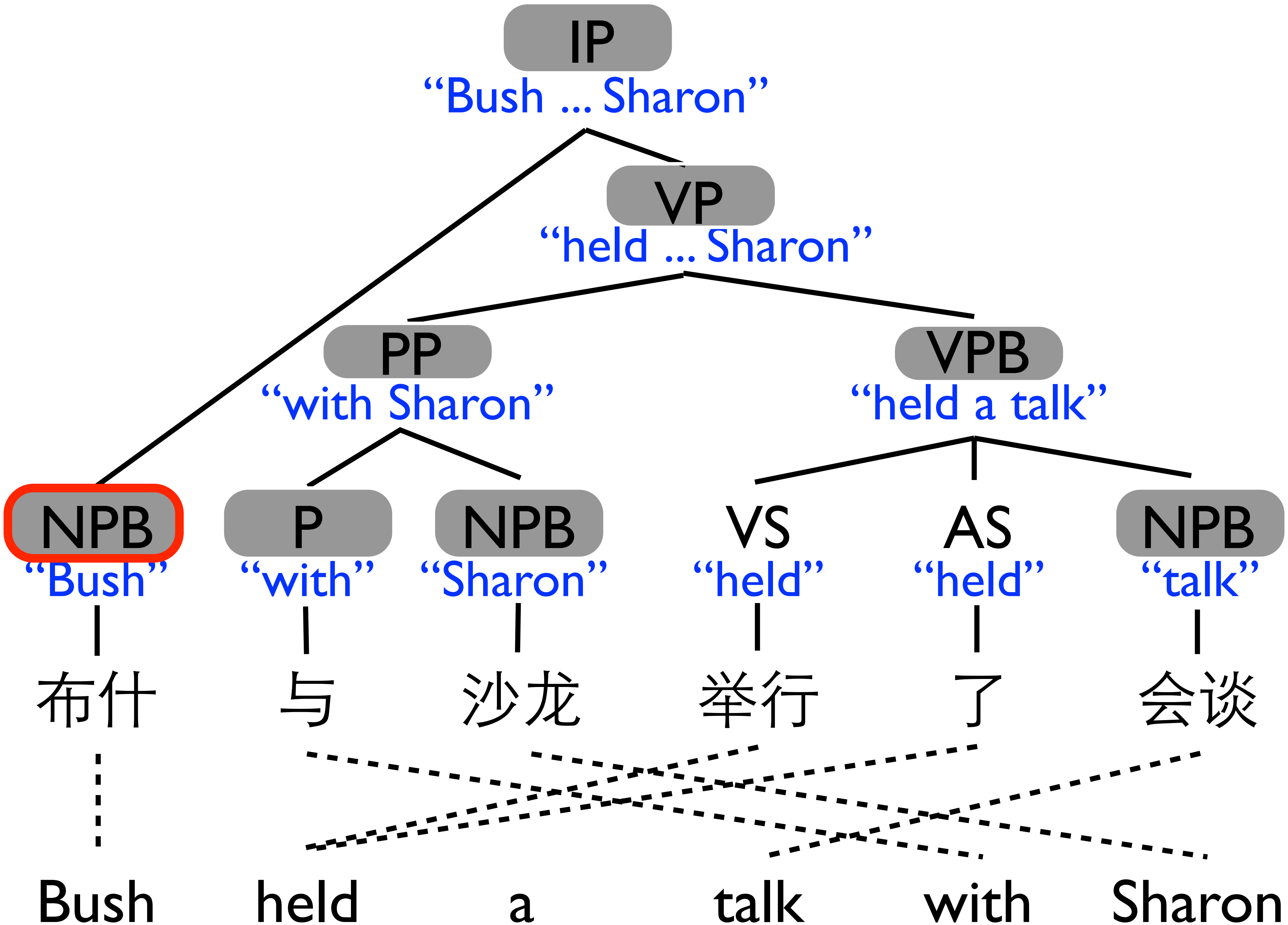
(Galley et al., 2004)

Rule Extraction



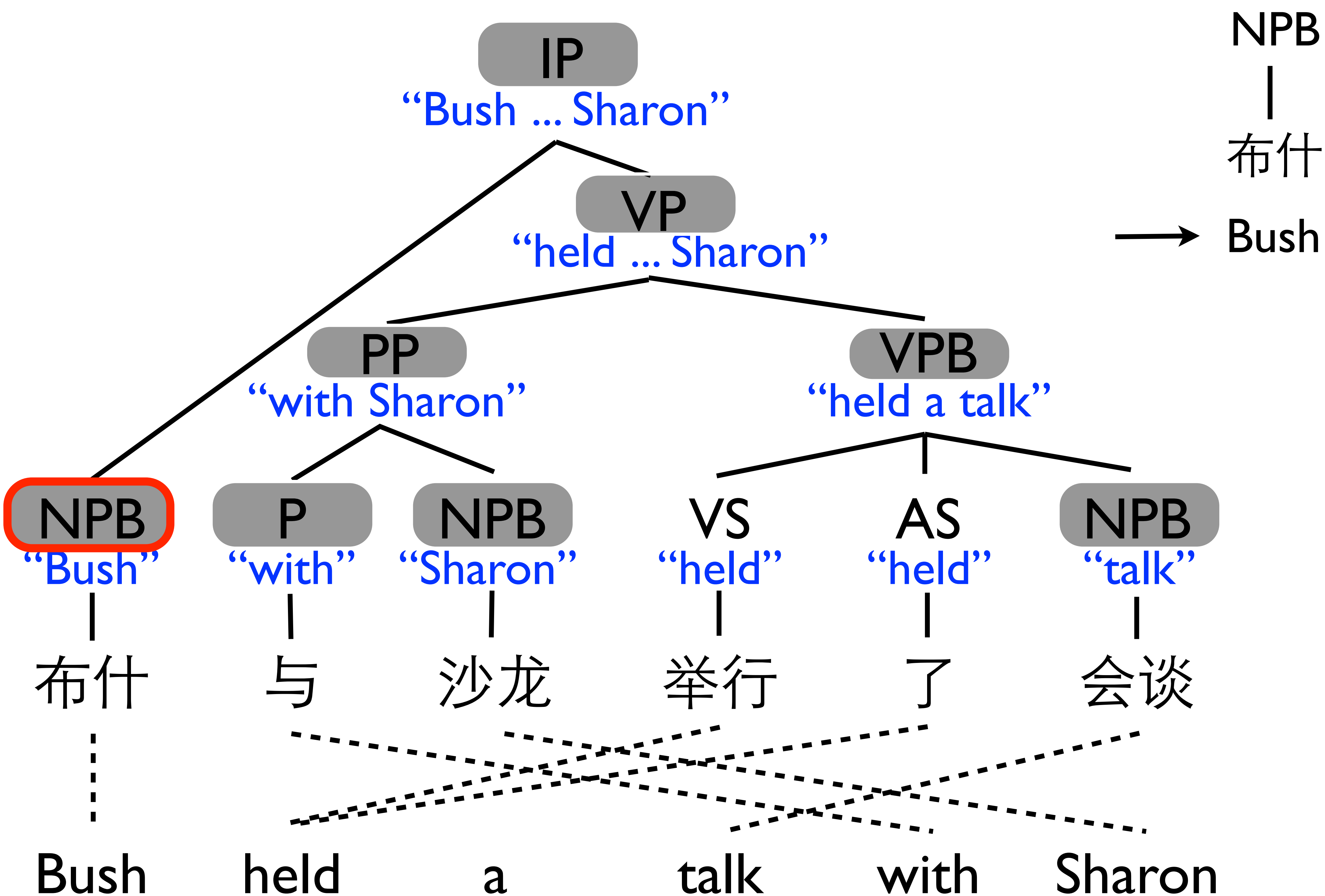
(Galley et al., 2004)

Rule Extraction



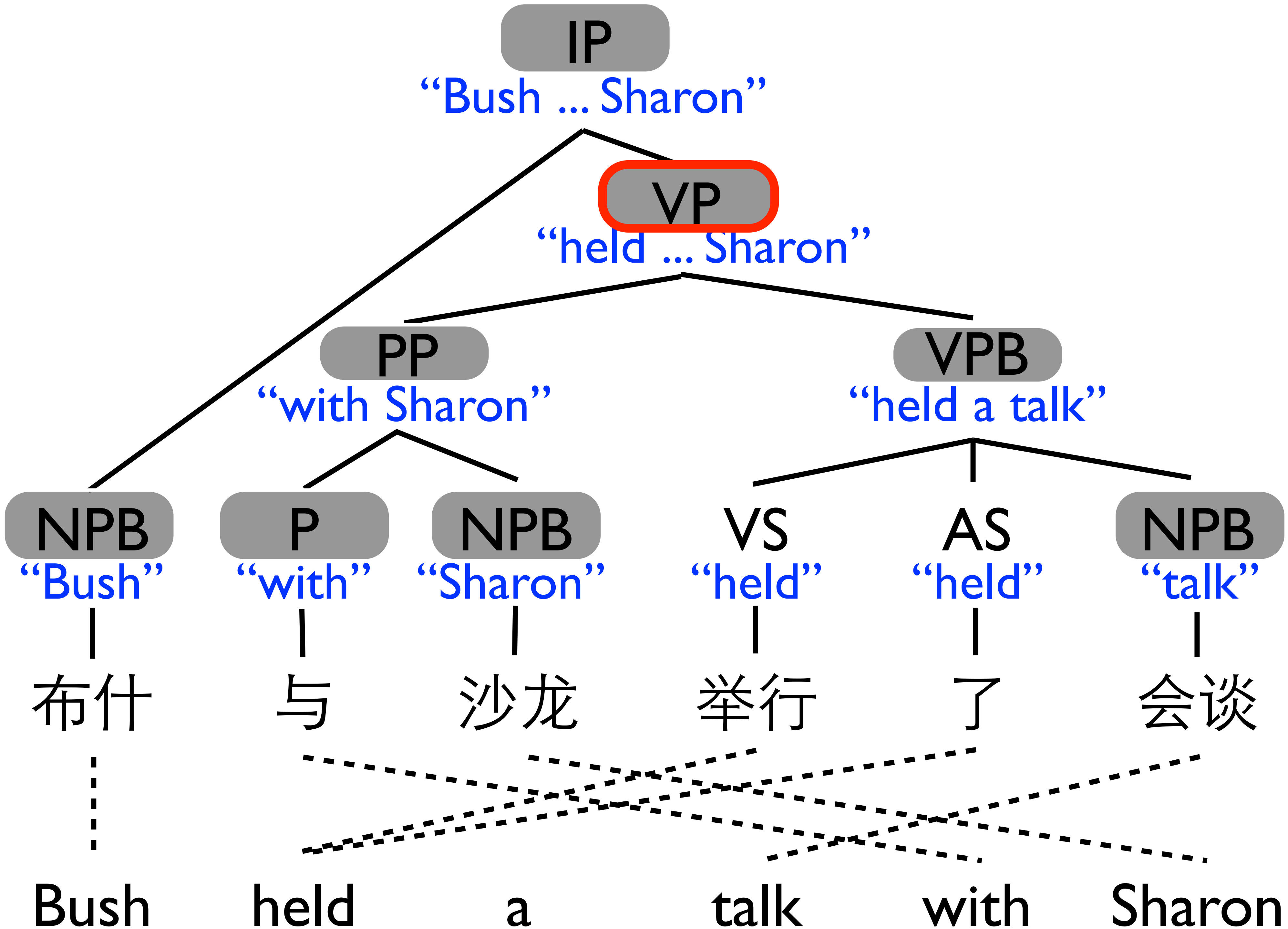
(Galley et al., 2004)

Rule Extraction



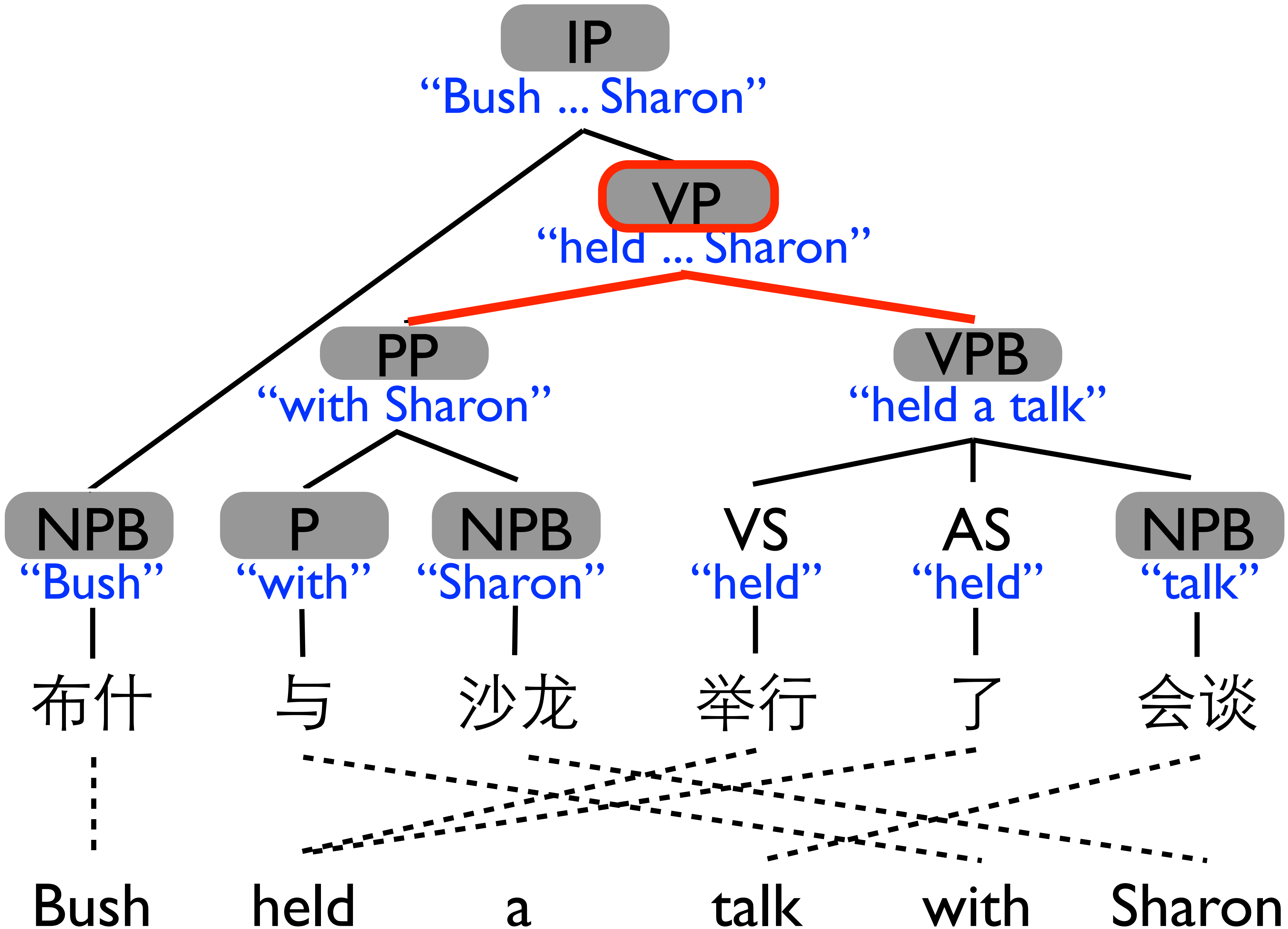
(Galley et al., 2004)

Rule Extraction



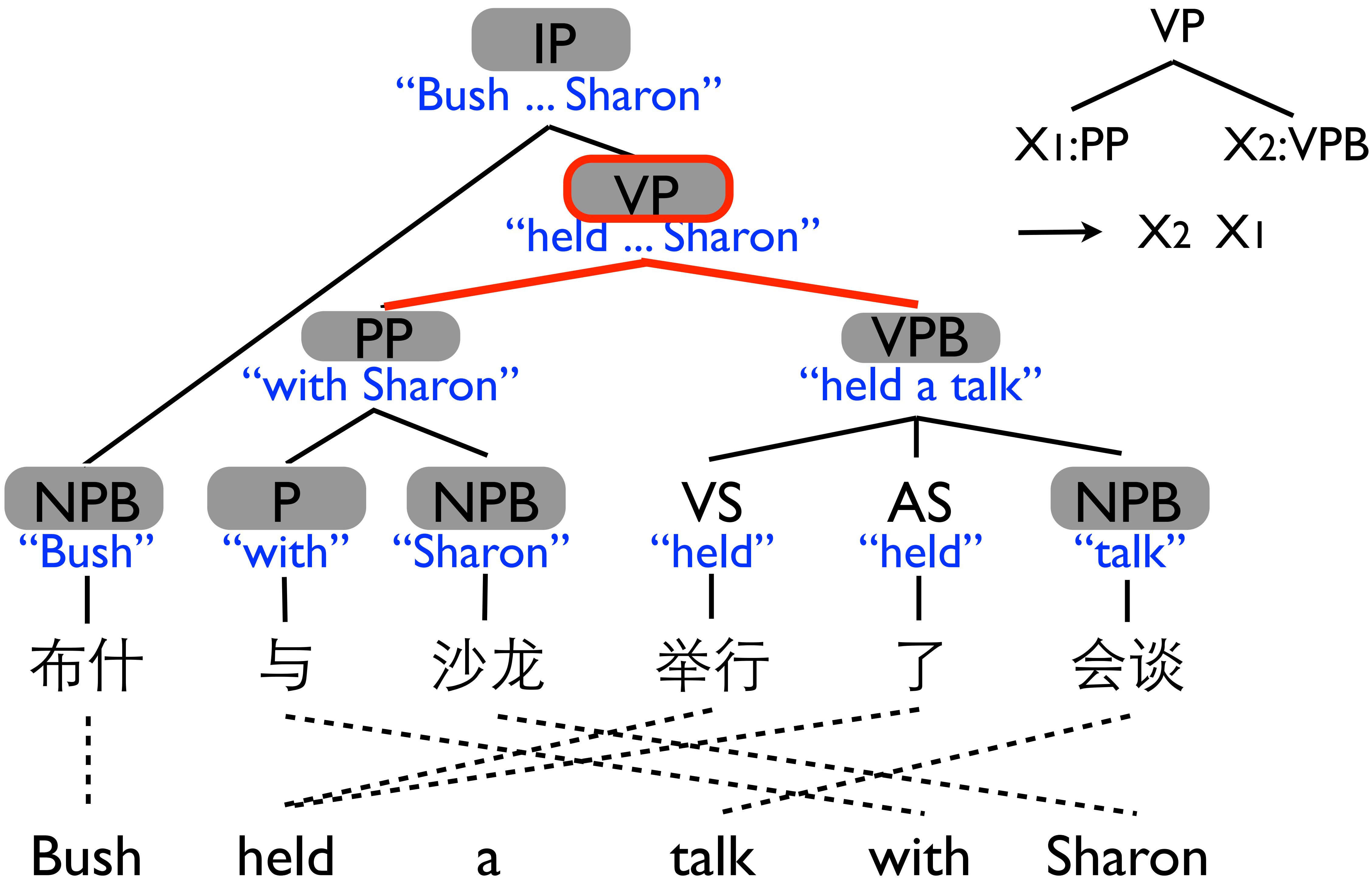
(Galley et al., 2004)

Rule Extraction



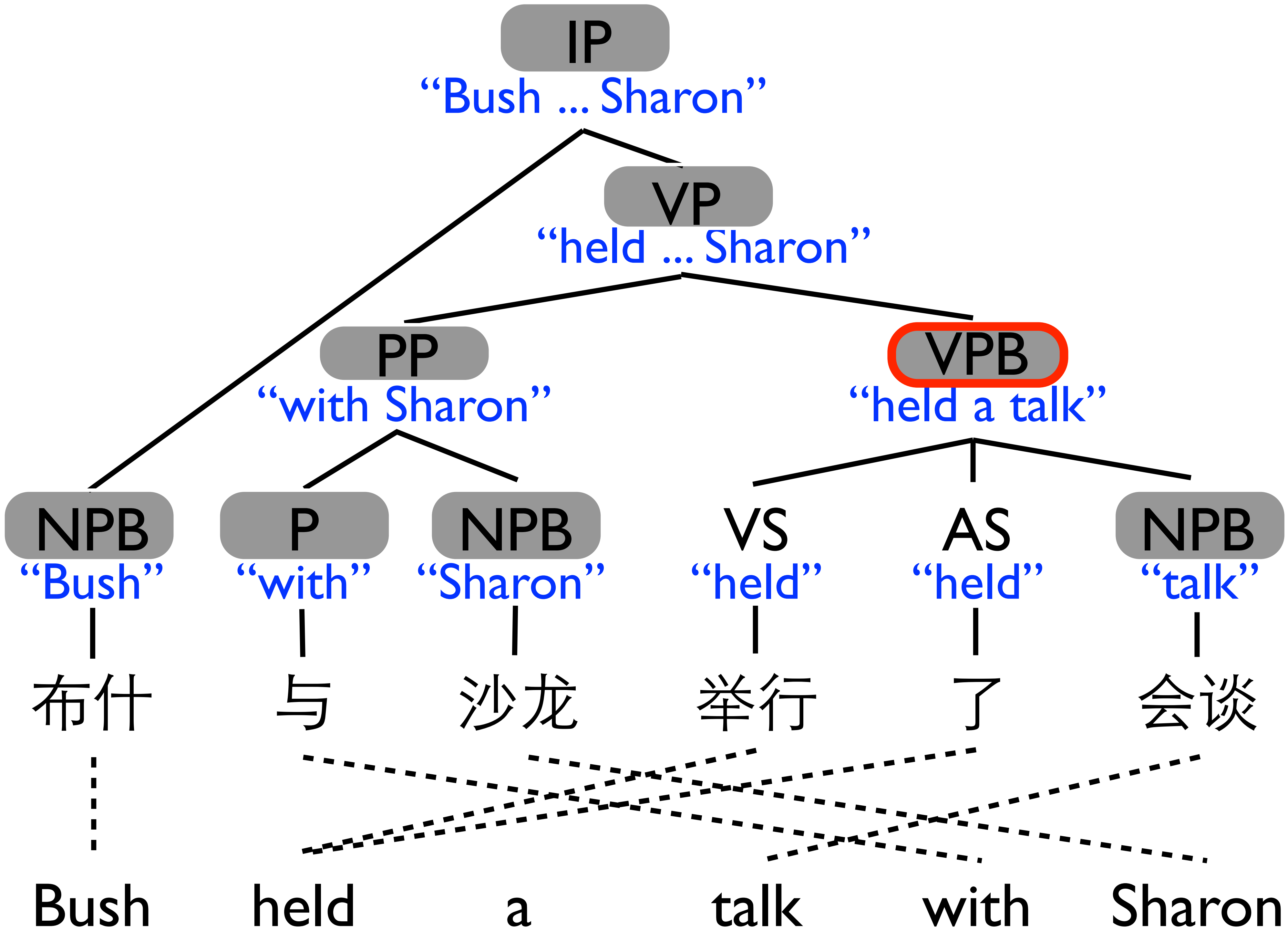
(Galley et al., 2004)

Rule Extraction



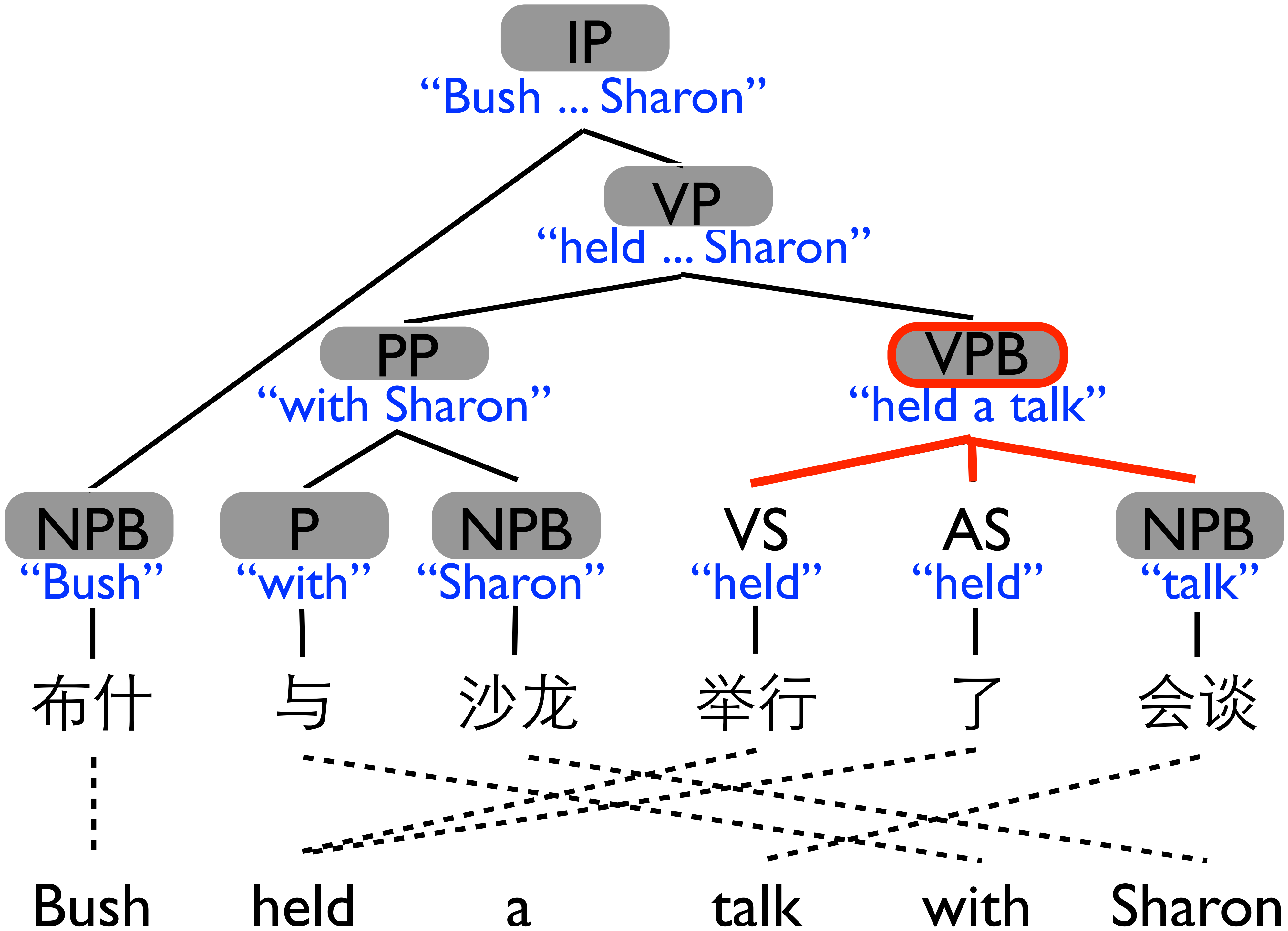
(Galley et al., 2004)

Rule Extraction



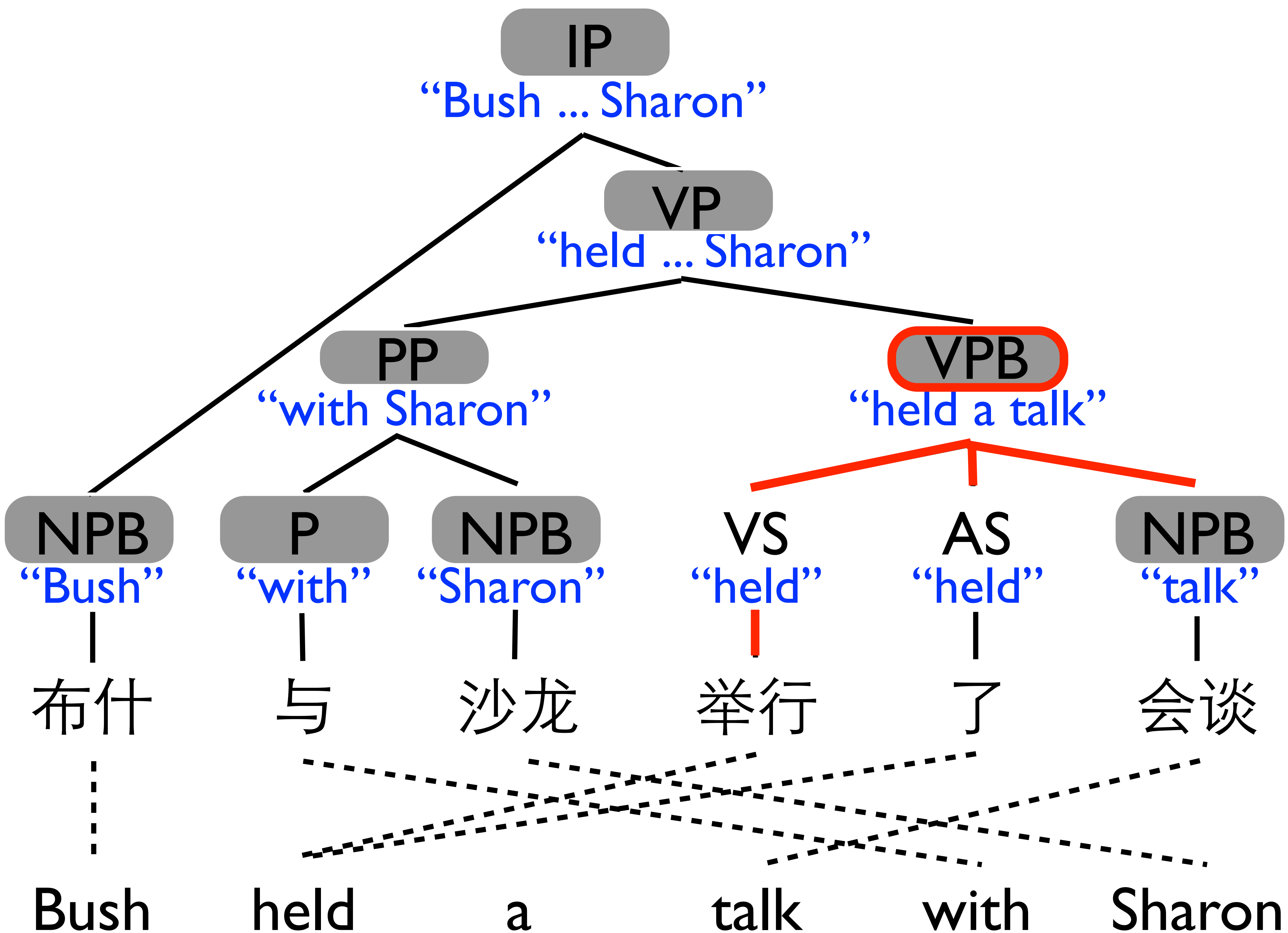
(Galley et al., 2004)

Rule Extraction



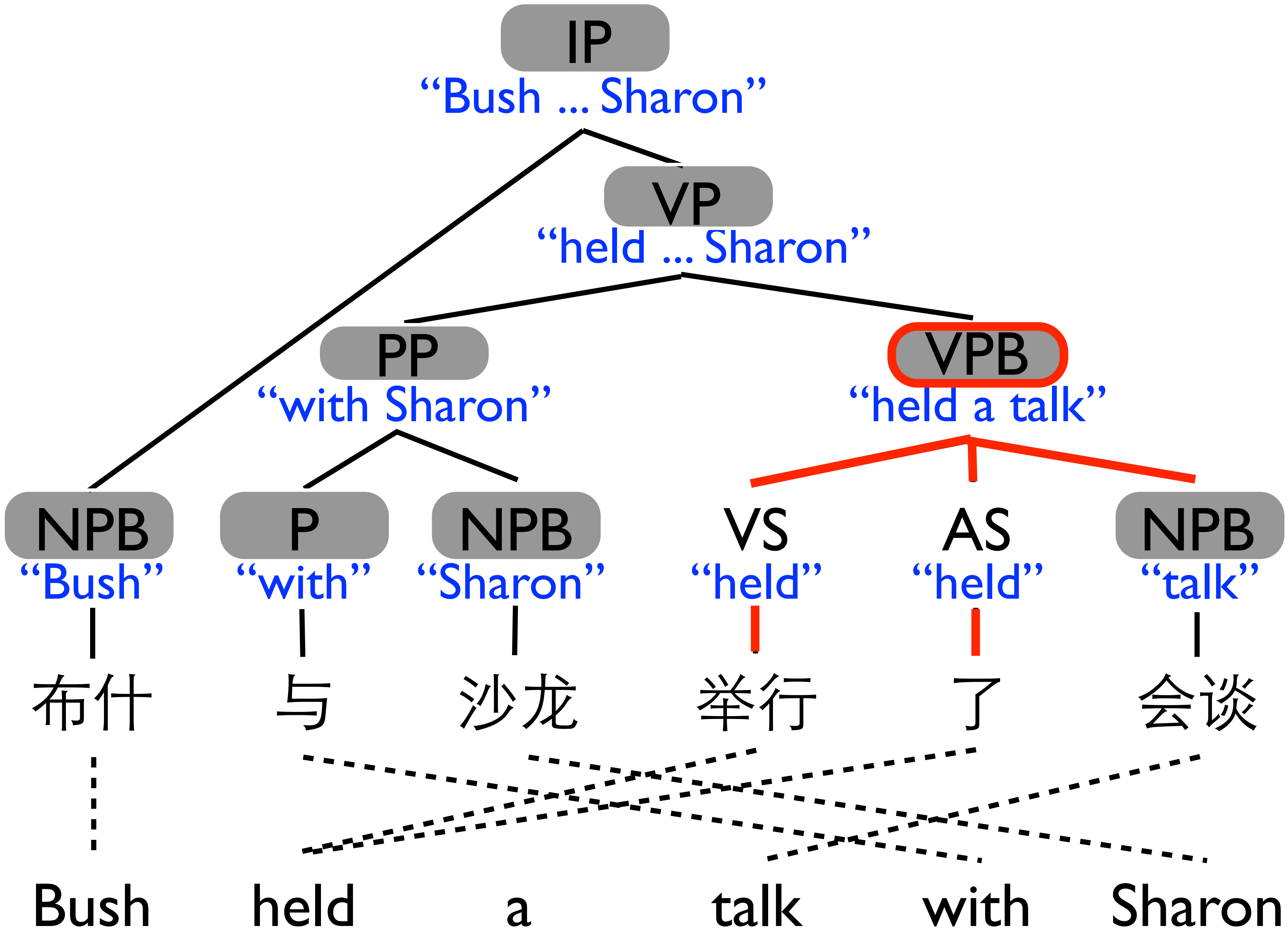
(Galley et al., 2004)

Rule Extraction



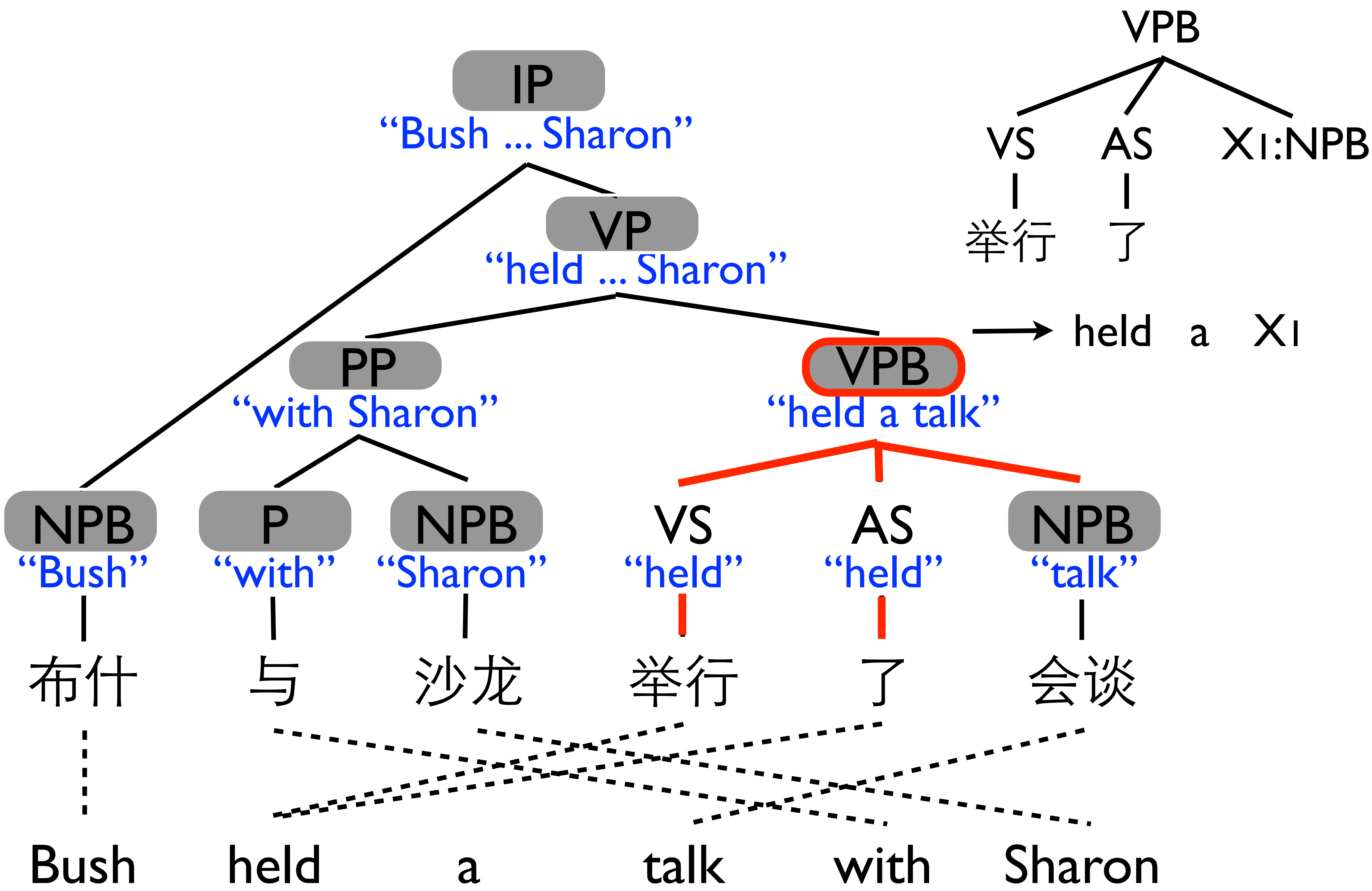
(Galley et al., 2004)

Rule Extraction



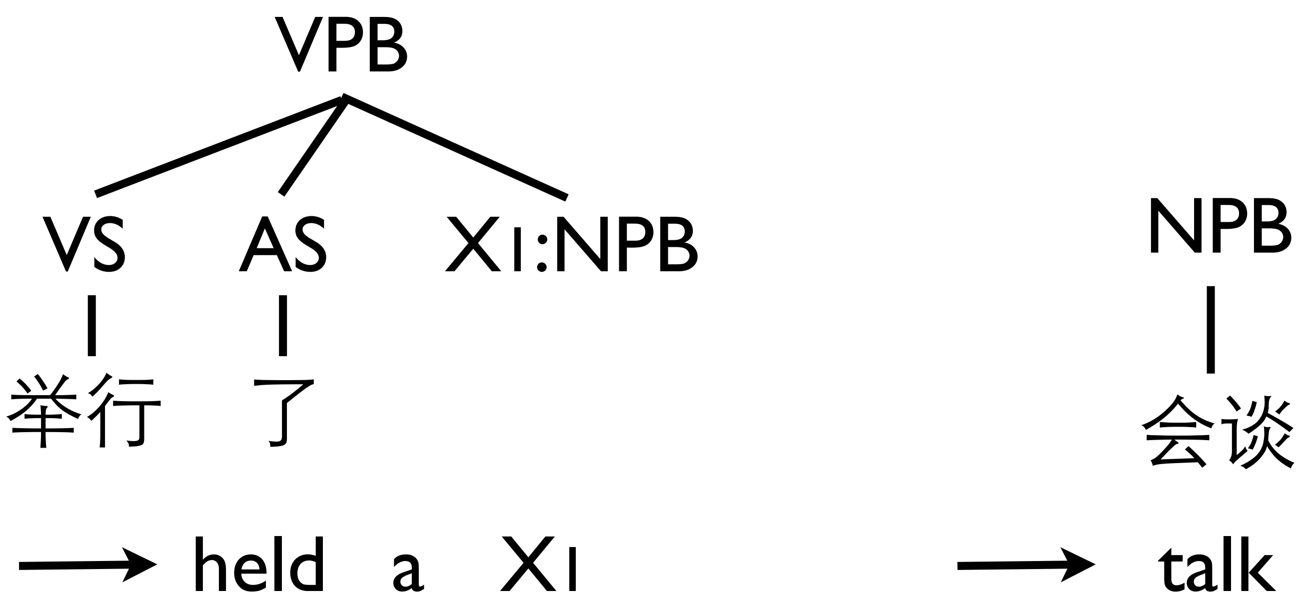
(Galley et al., 2004)

Rule Extraction



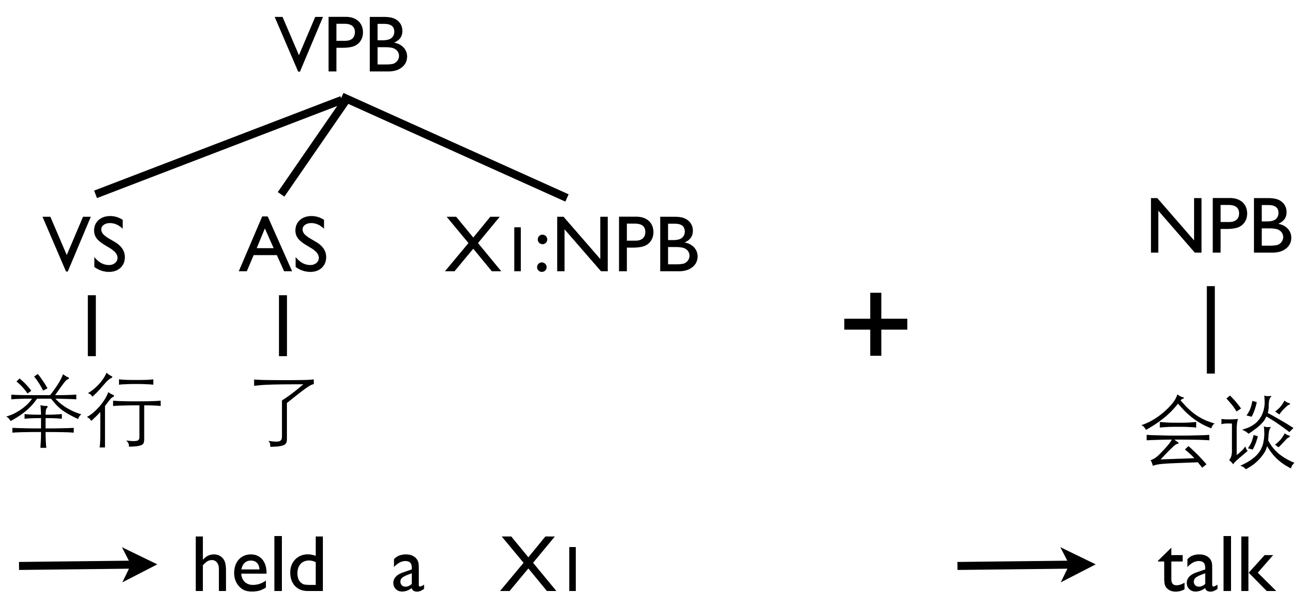
(Galley et al., 2004)

Rule Composition



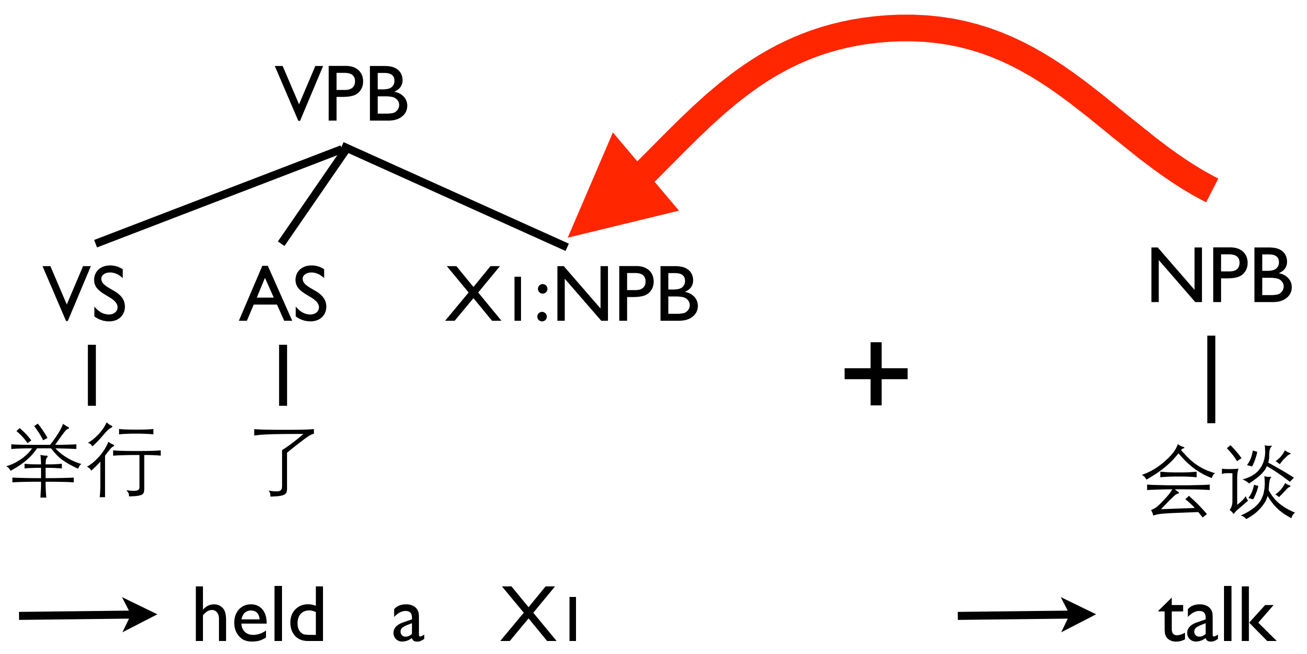
(Galley et al., 2006)

Rule Composition



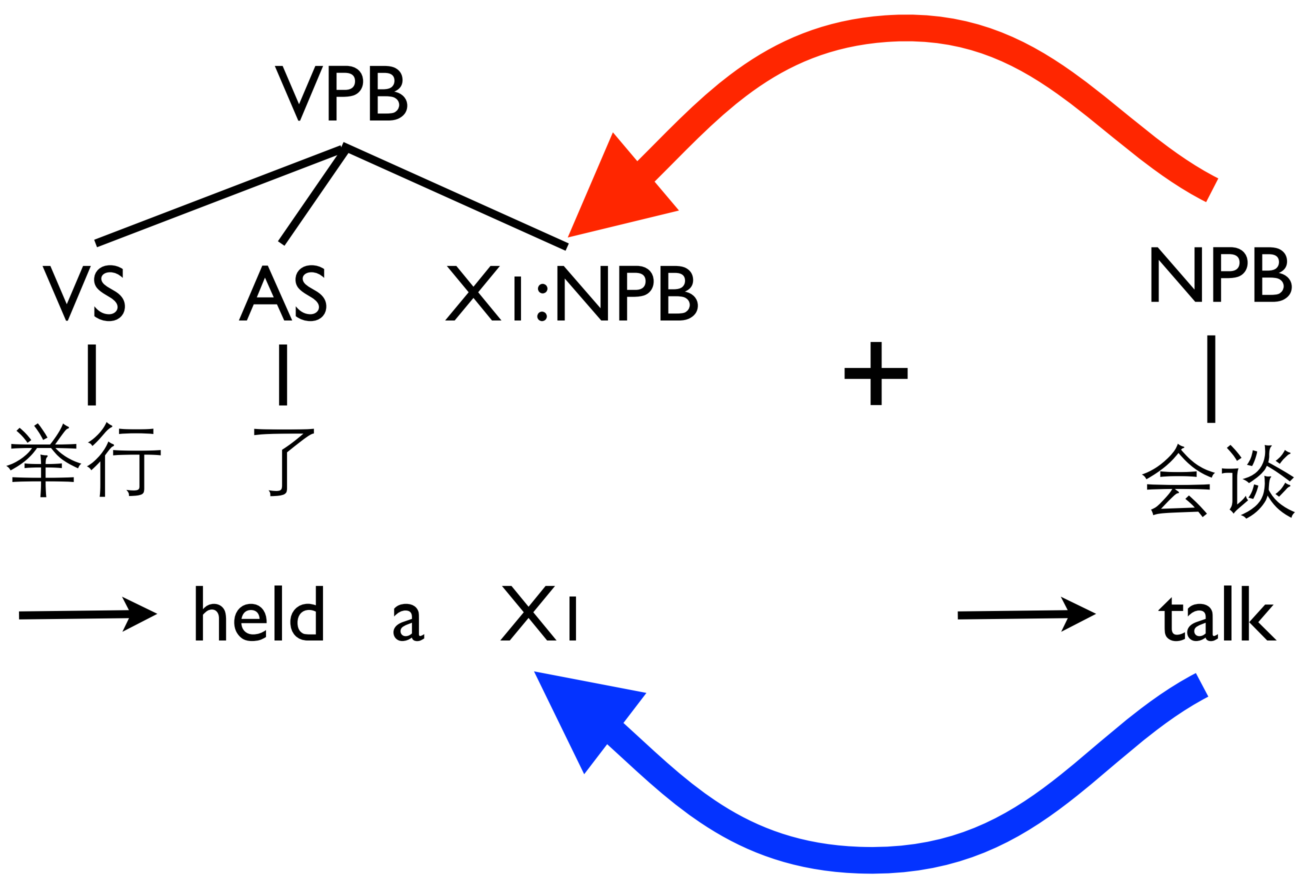
(Galley et al., 2006)

Rule Composition



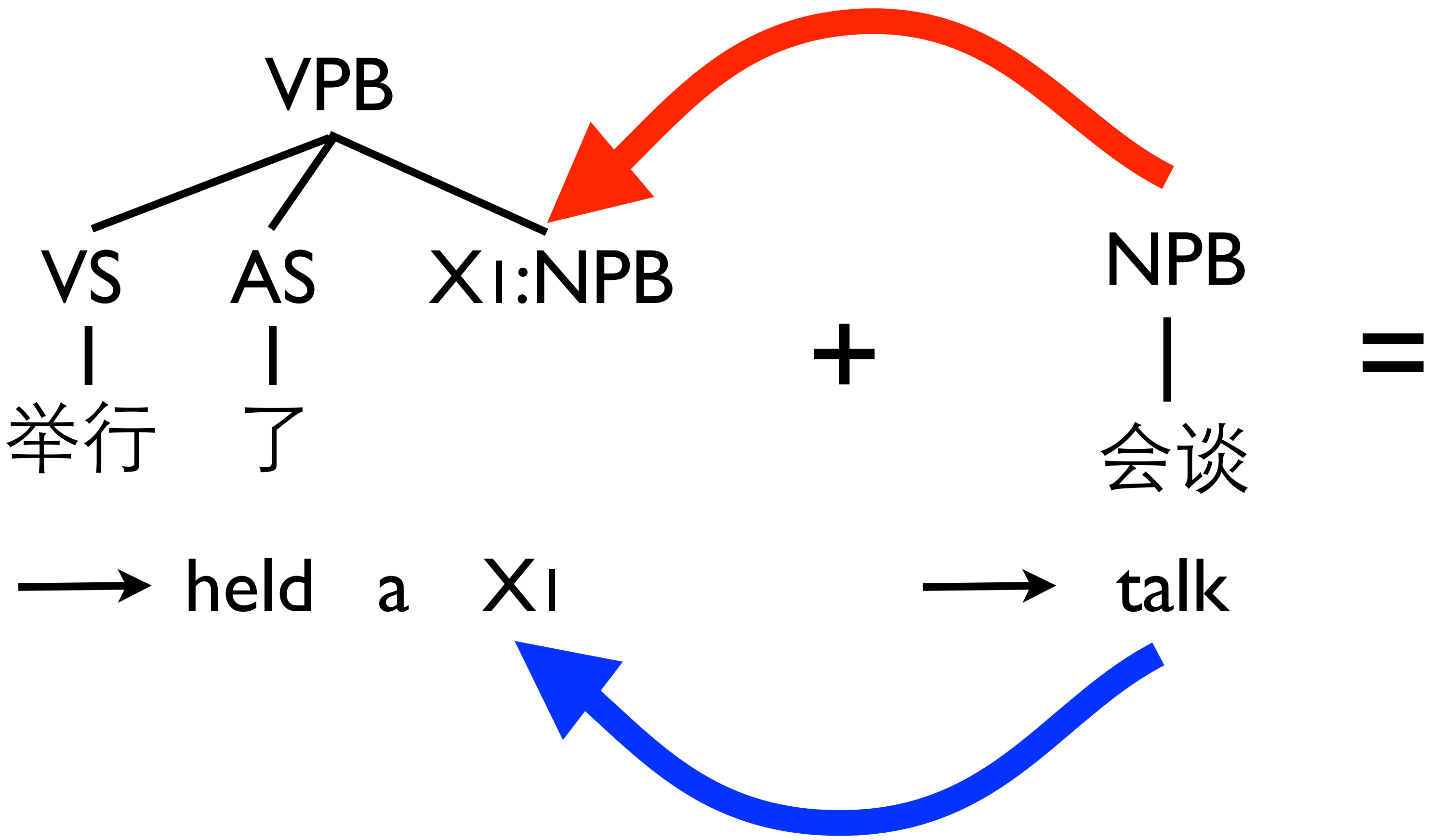
(Galley et al., 2006)

Rule Composition



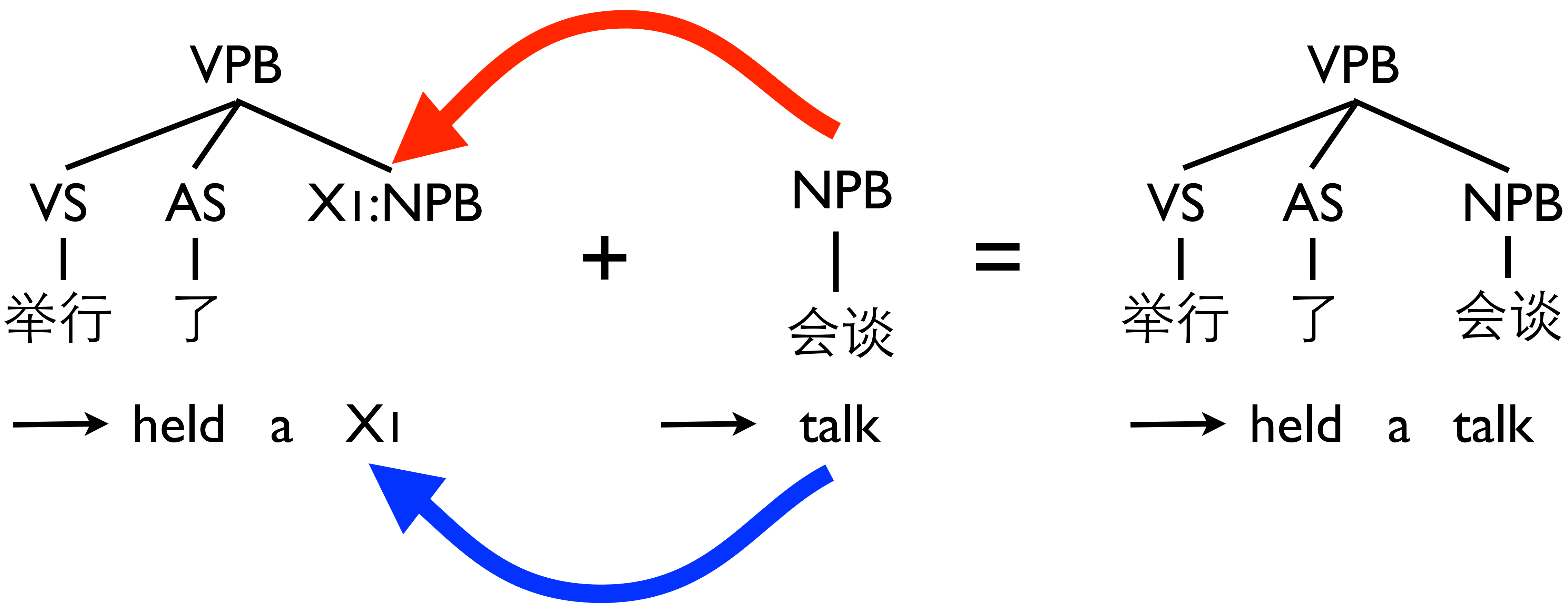
(Galley et al., 2006)

Rule Composition



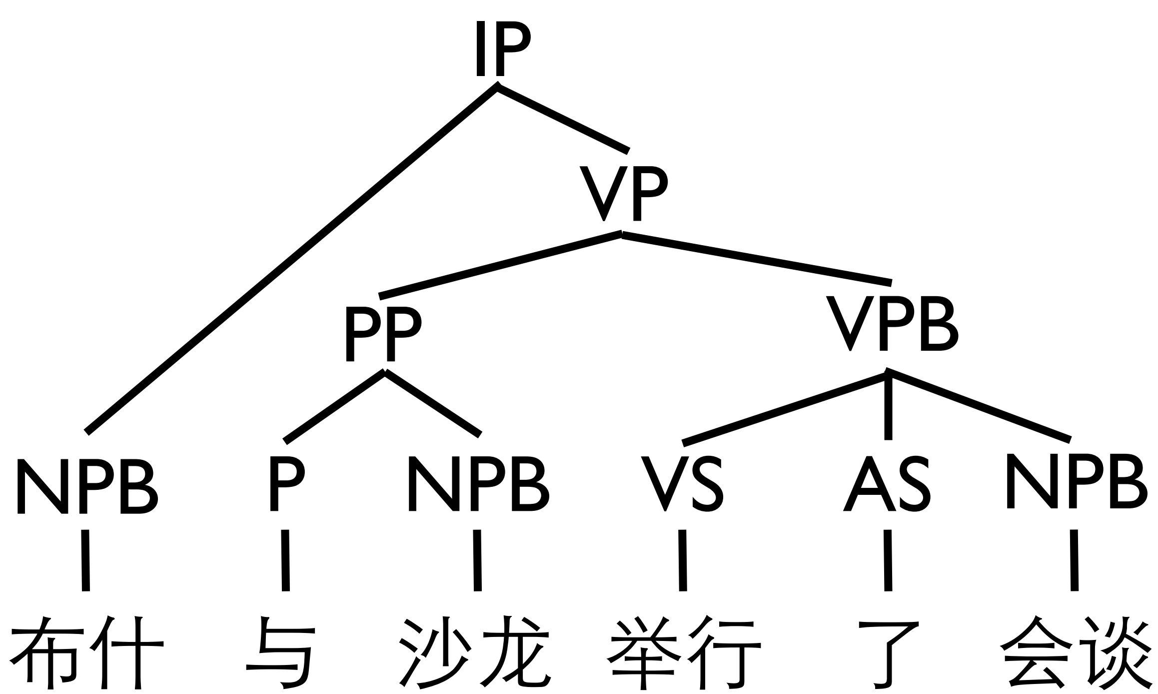
(Galley et al., 2006)

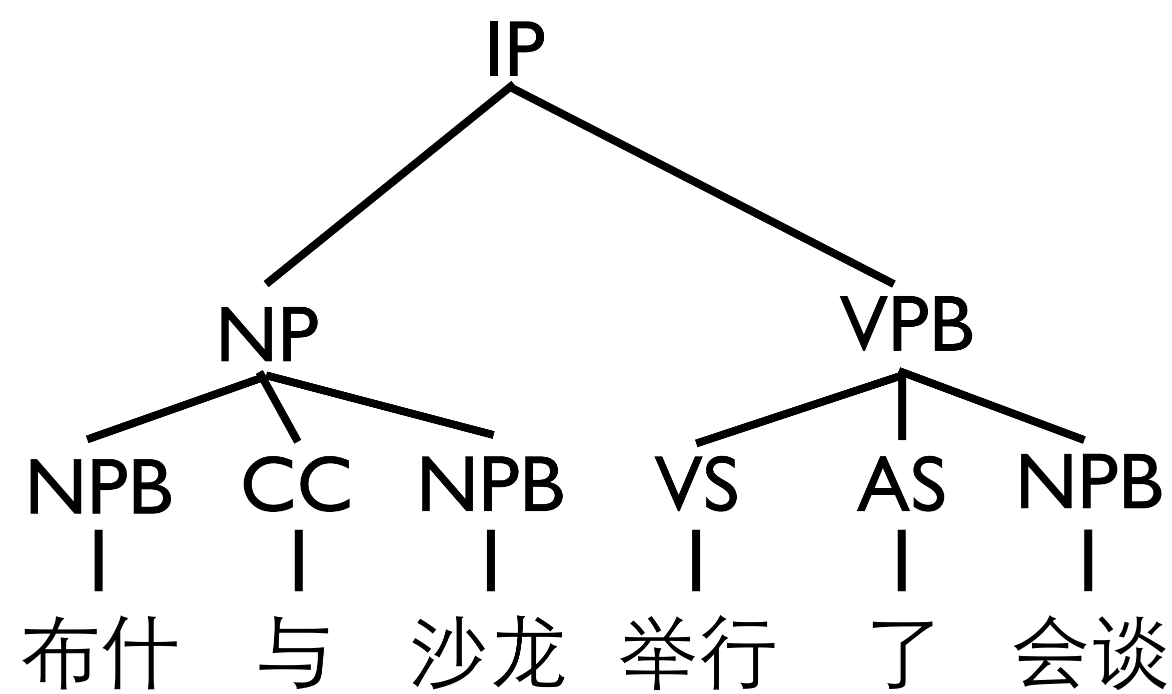
Rule Composition



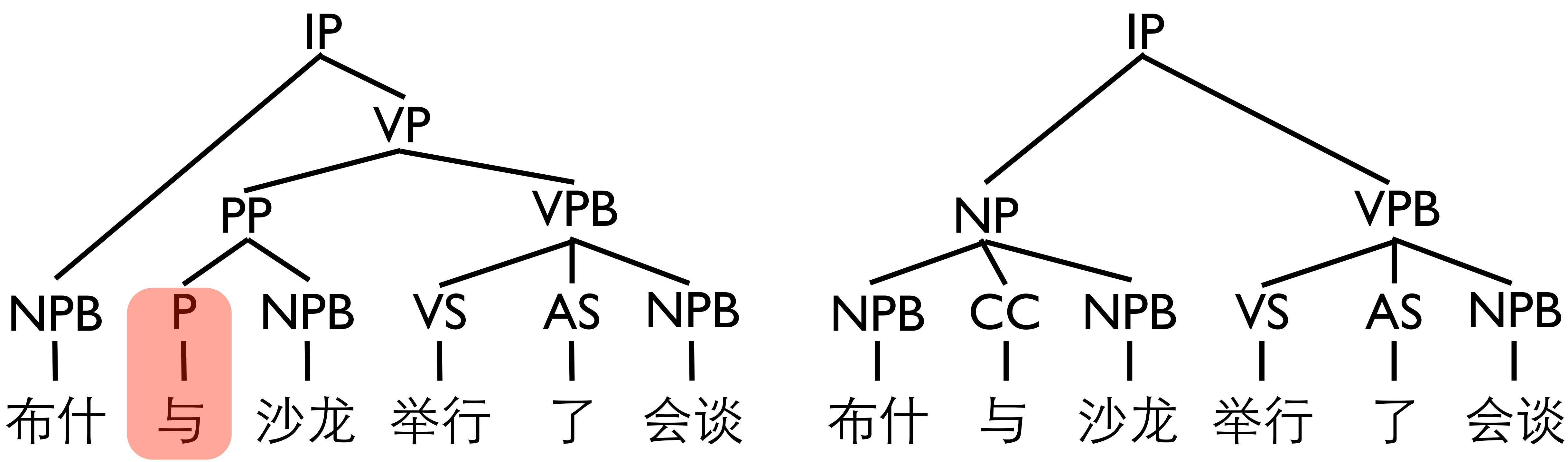
(Galley et al., 2006)

Parsing Ambiguity

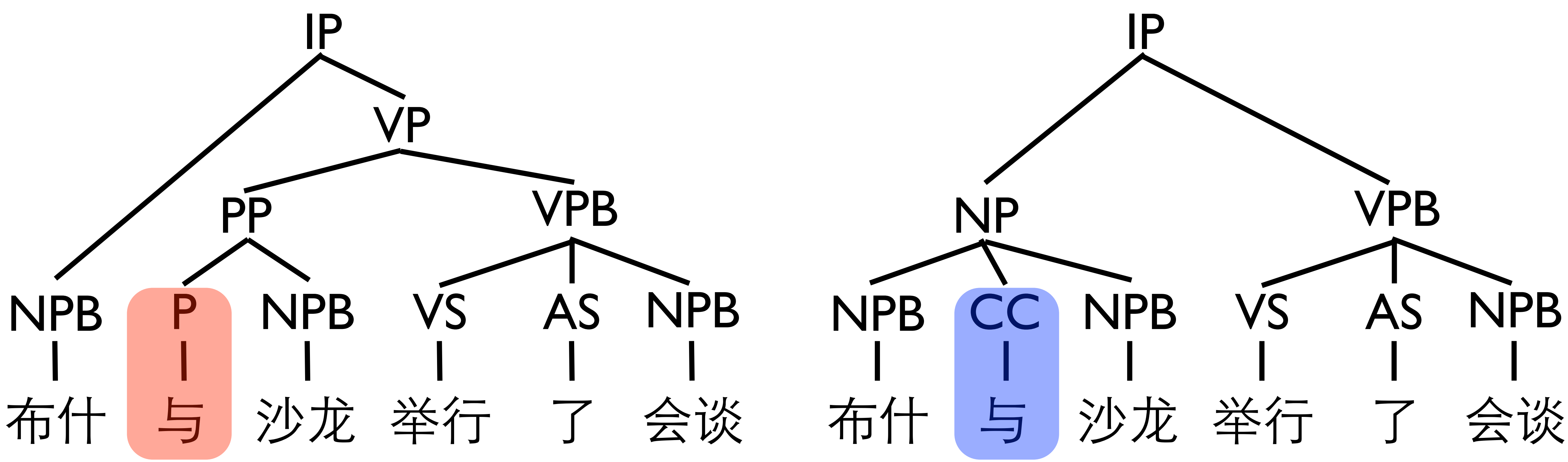




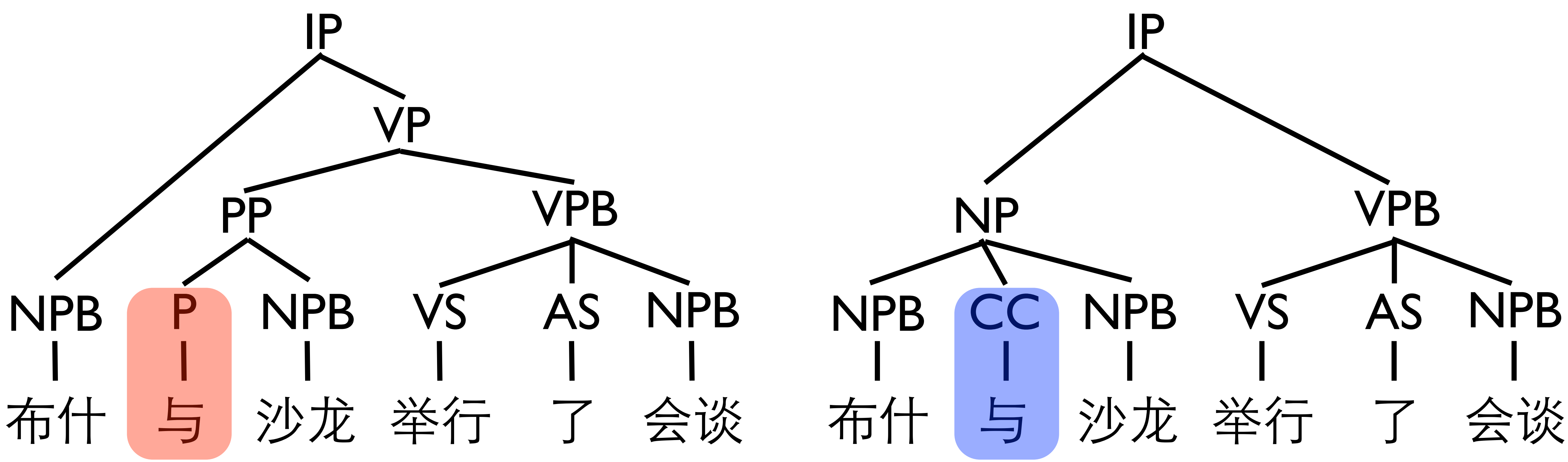
Parsing Ambiguity



Parsing Ambiguity

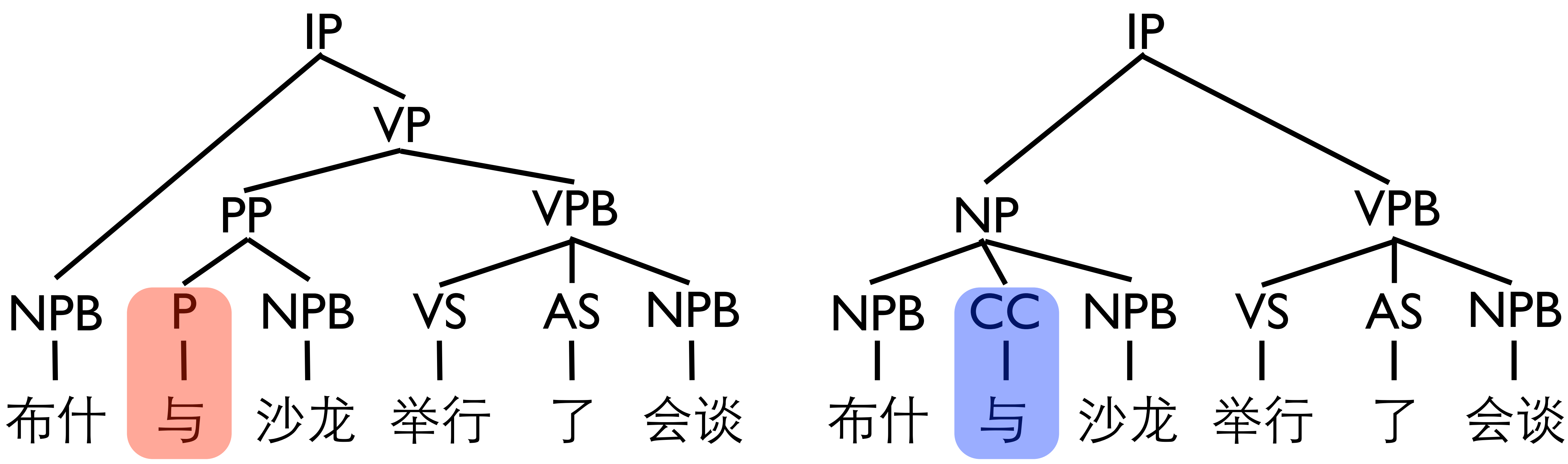


Parsing Ambiguity



Bush held a talk with Sharon

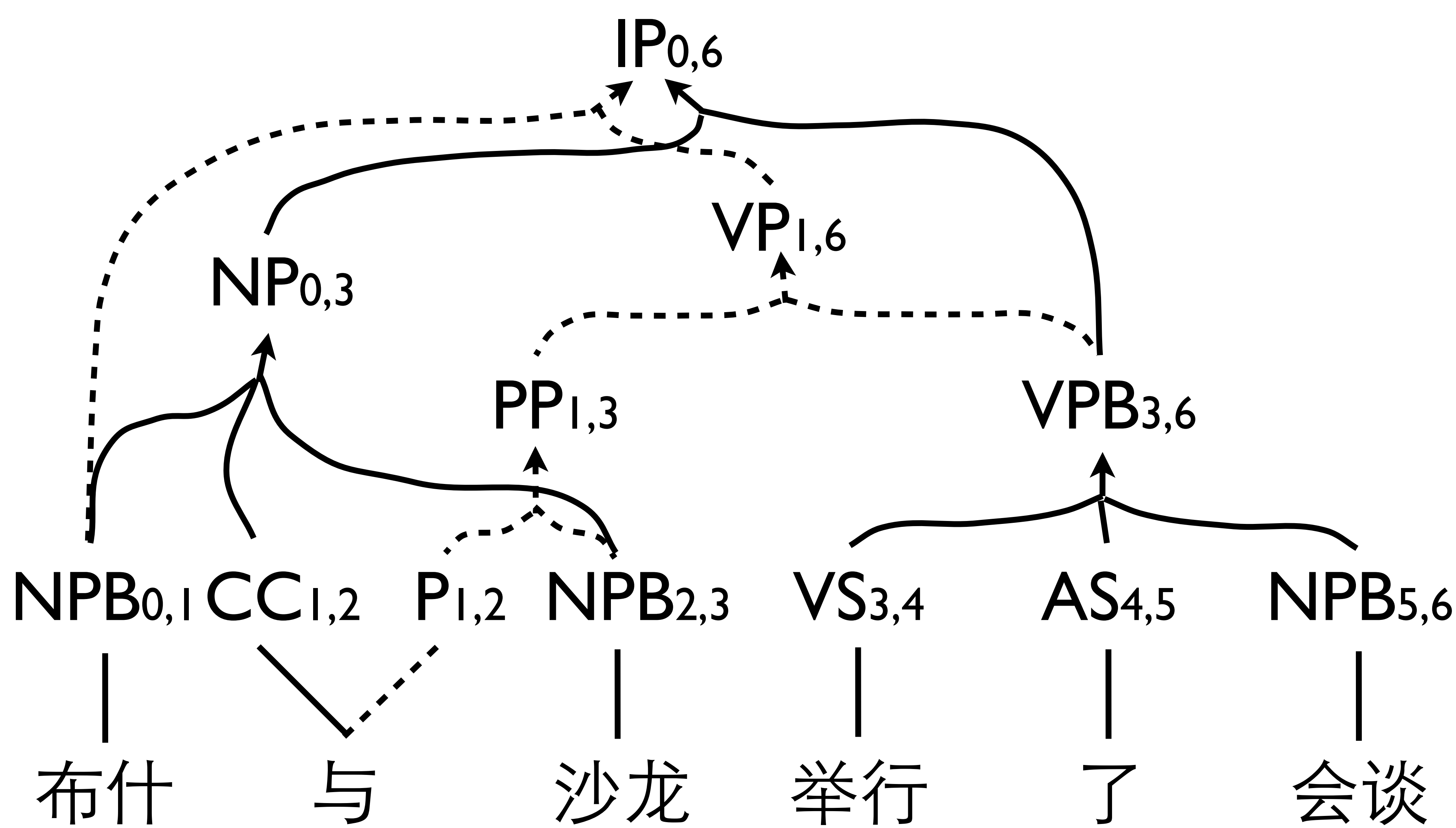
Parsing Ambiguity



Bush held a talk with Sharon

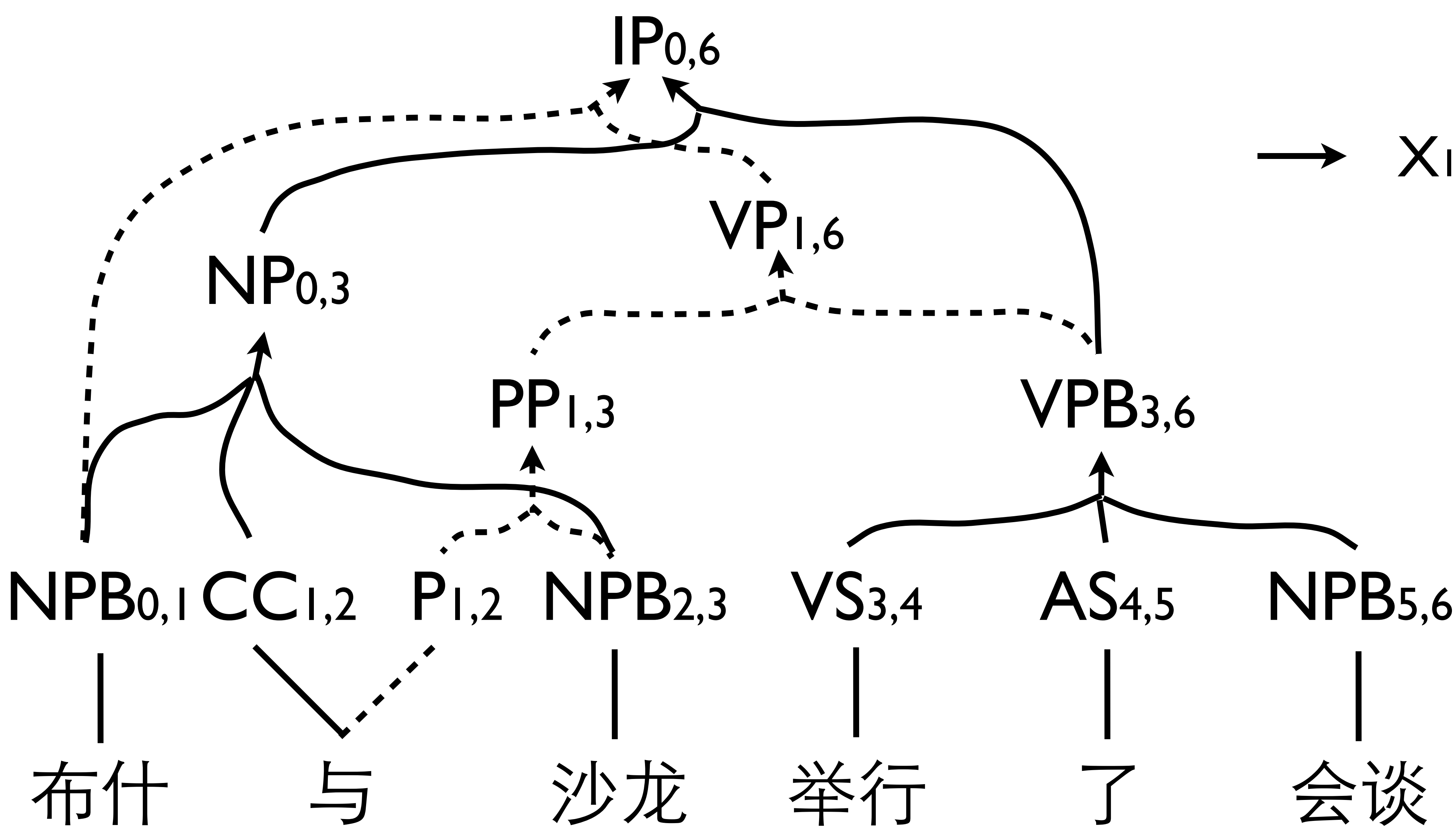
Bush and Sharon held a talk

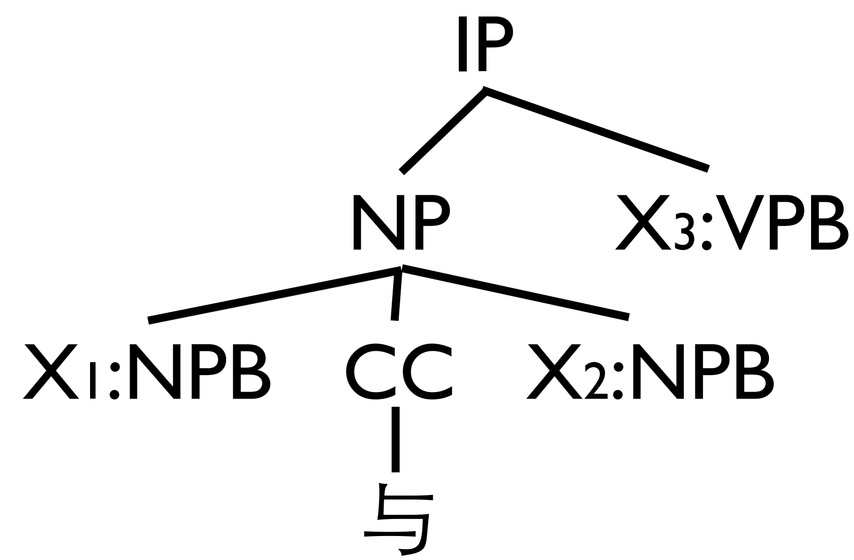
Packed Forest



(Billot and Lang, 1989)

Forest-based Decoding

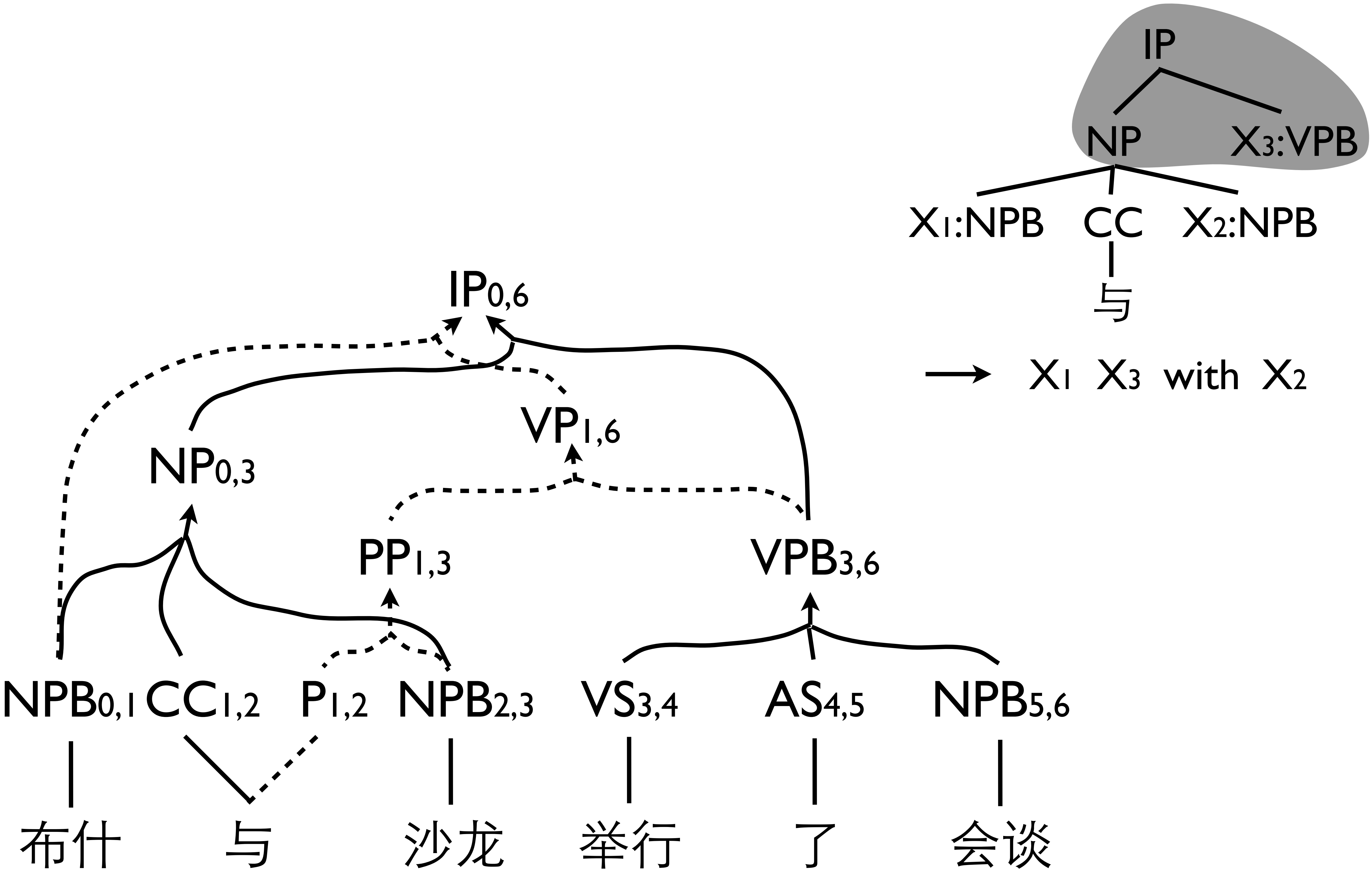




$\rightarrow$ X_1 X_3 with X_2

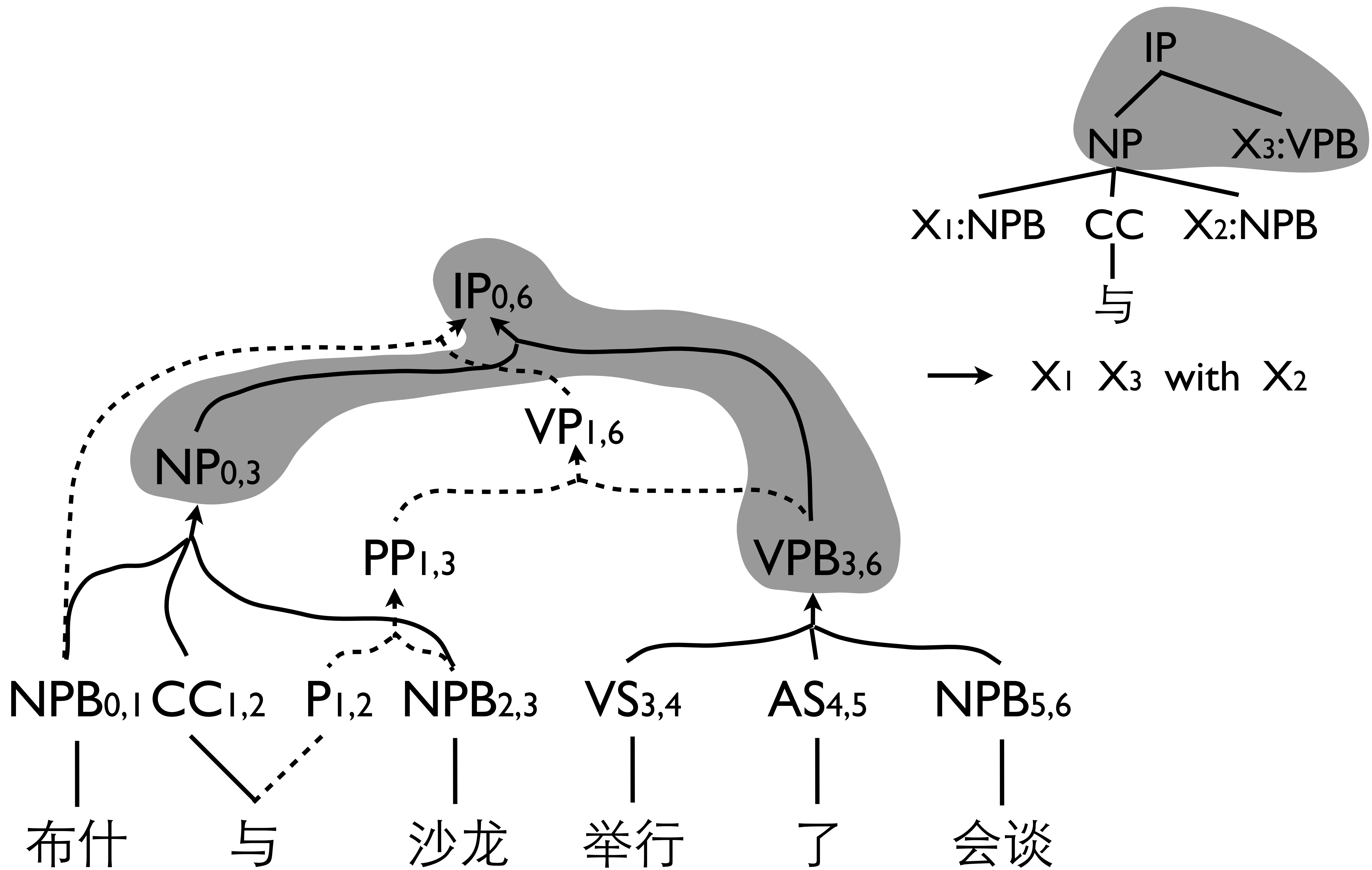
(Mi et al., 2008)

Forest-based Decoding



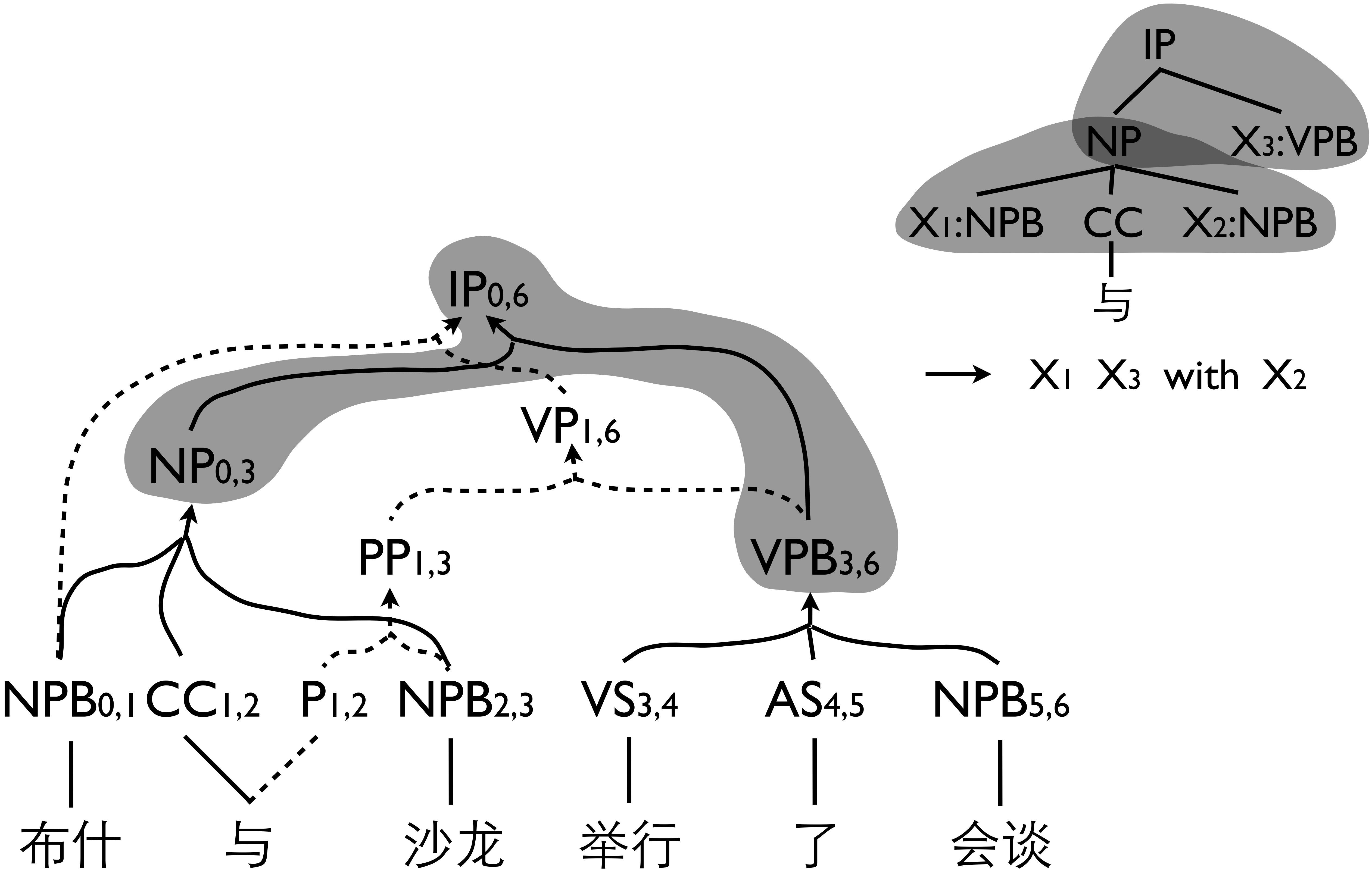
(Mi et al., 2008)

Forest-based Decoding



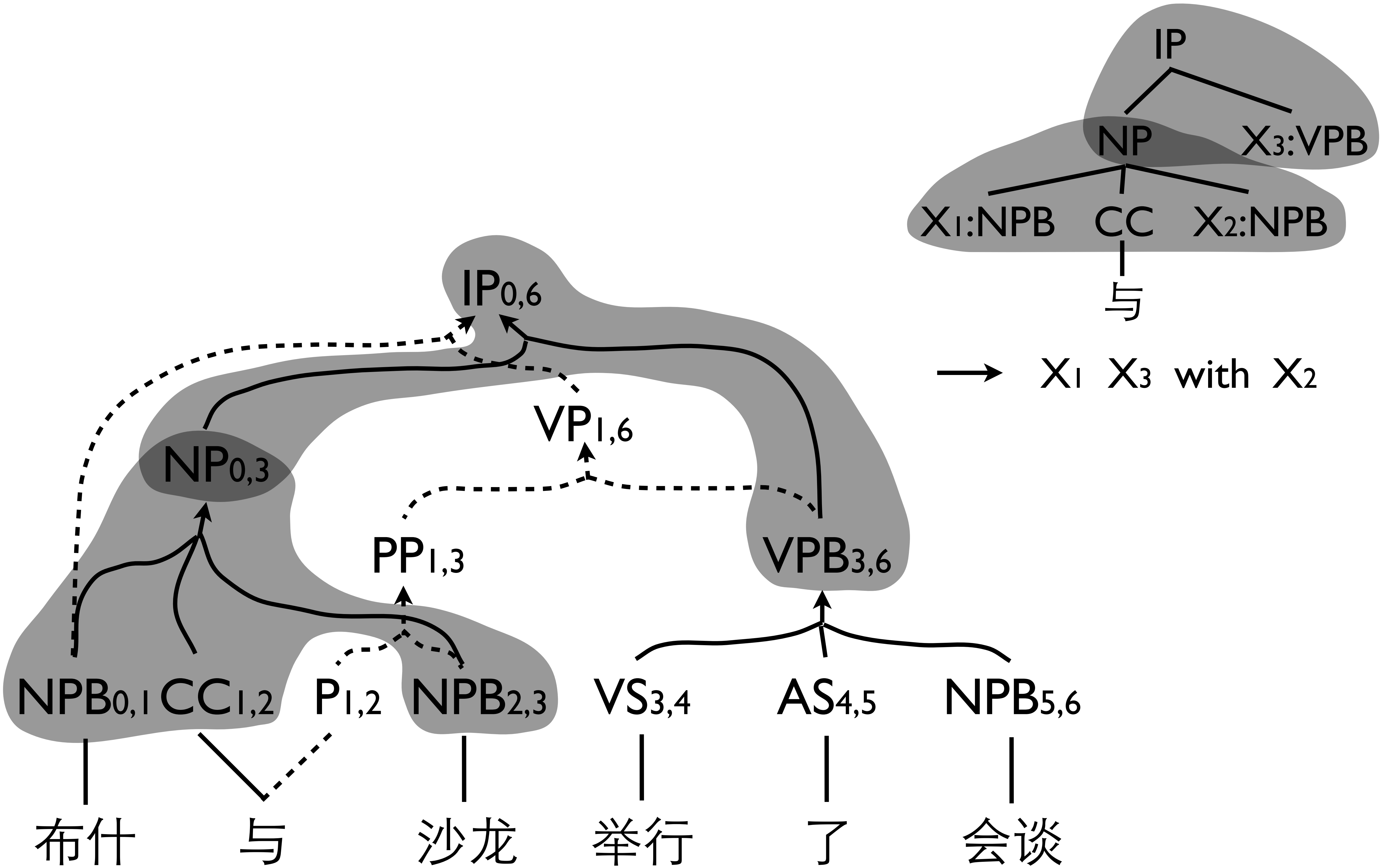
(Mi et al., 2008)

Forest-based Decoding



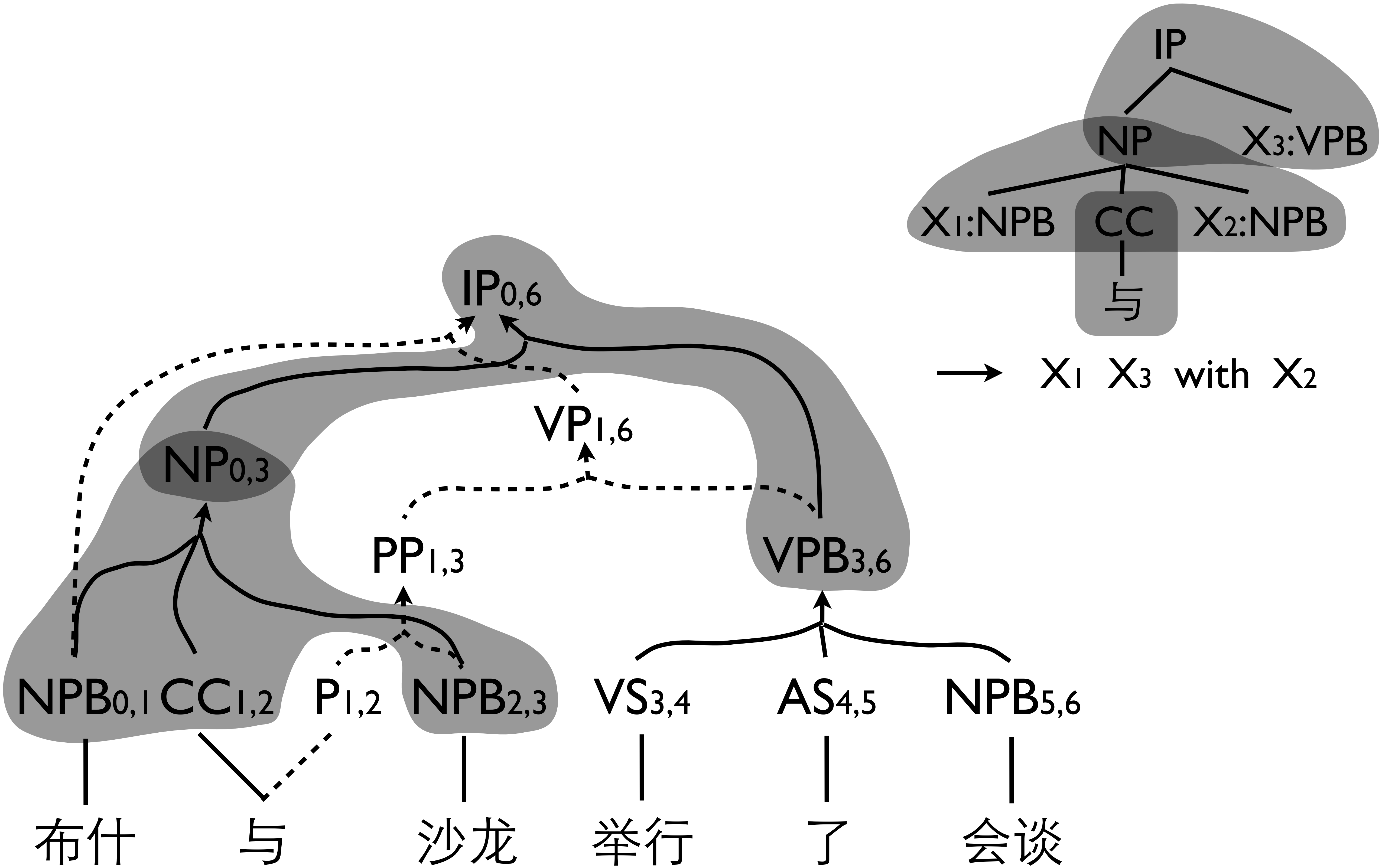
(Mi et al., 2008)

Forest-based Decoding



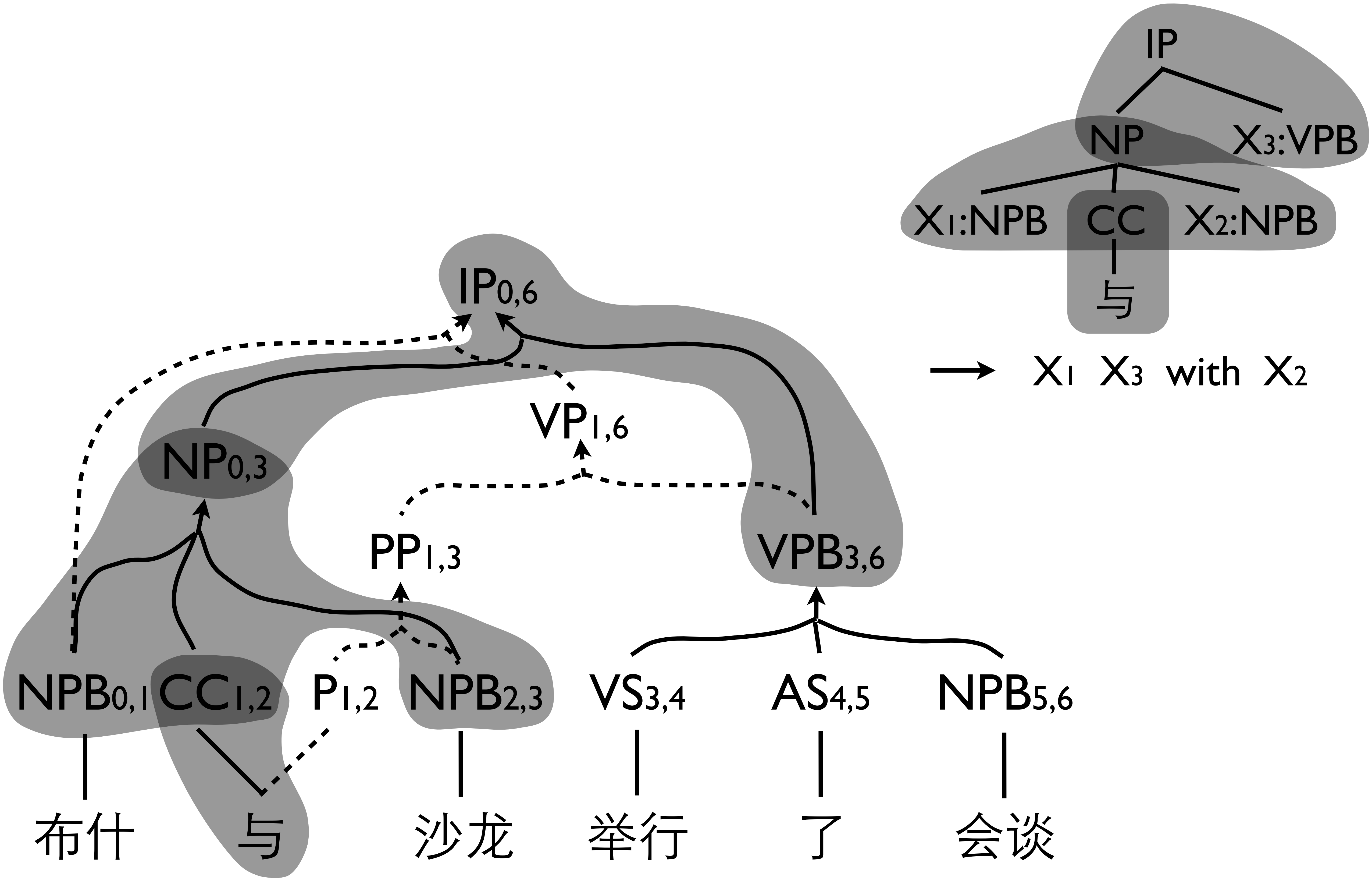
(Mi et al., 2008)

Forest-based Decoding



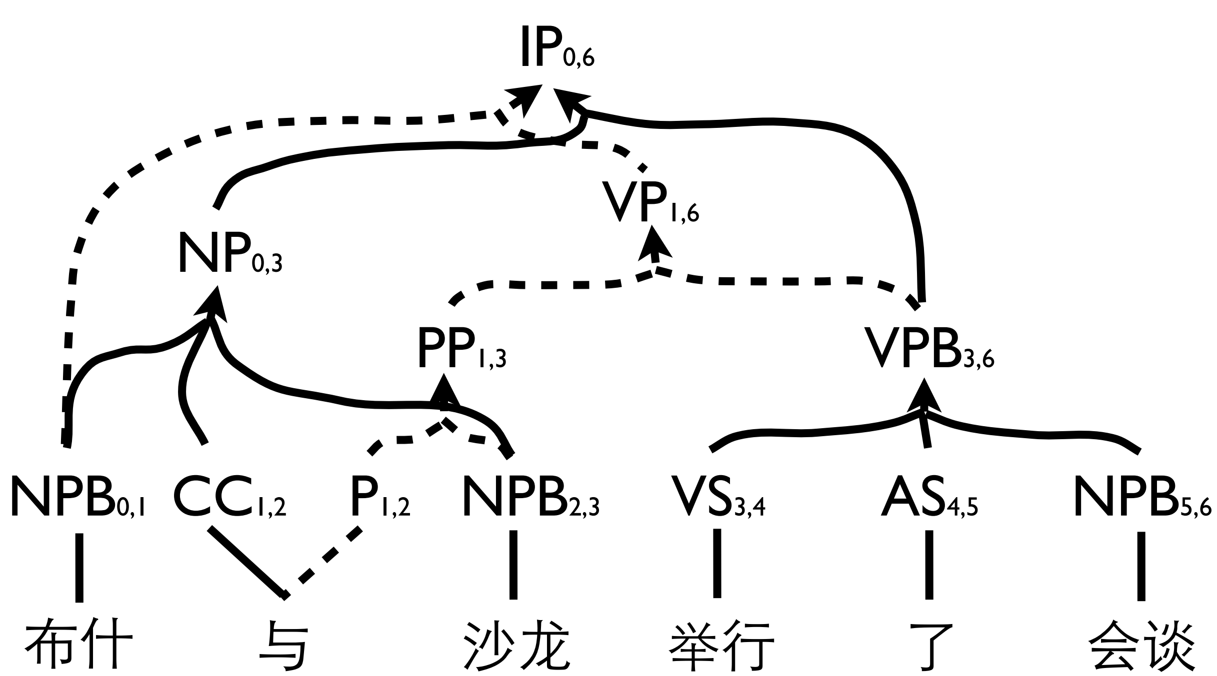
(Mi et al., 2008)

Forest-based Decoding



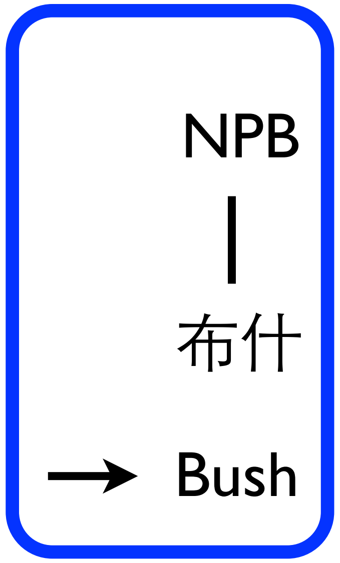
(Mi et al., 2008)

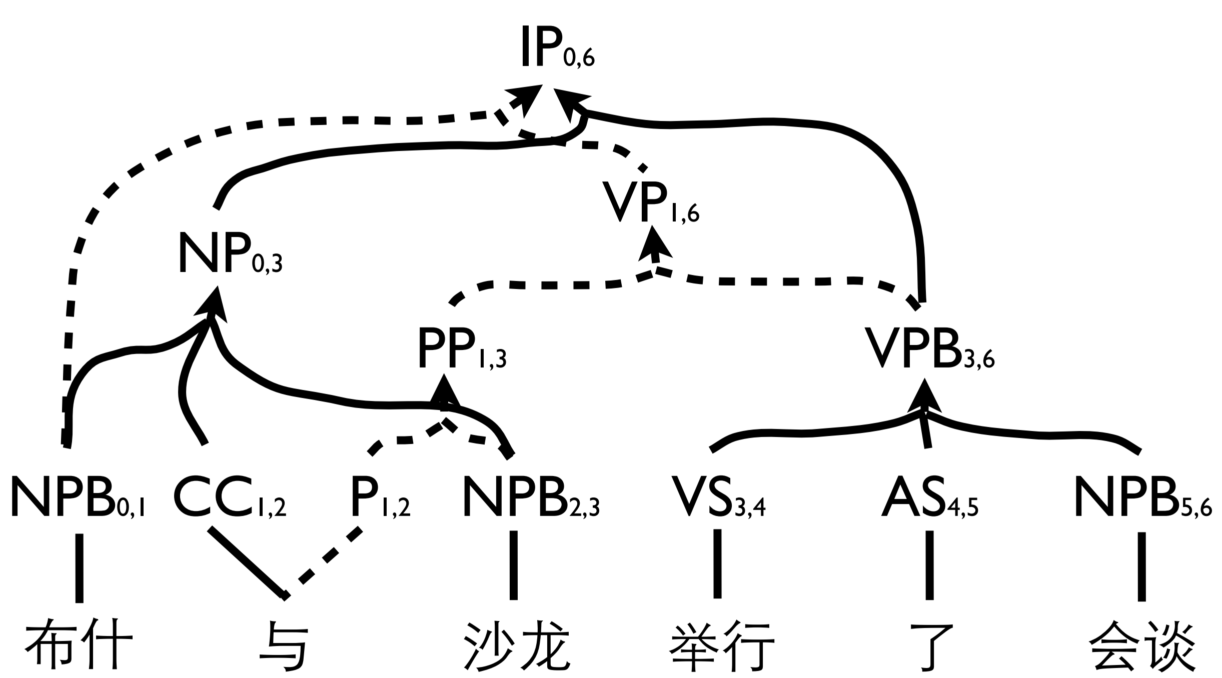
Forest-based Decoding



(Mi et al., 2008)

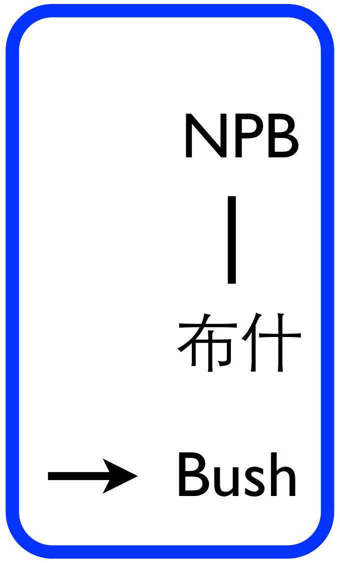
Forest-based Decoding

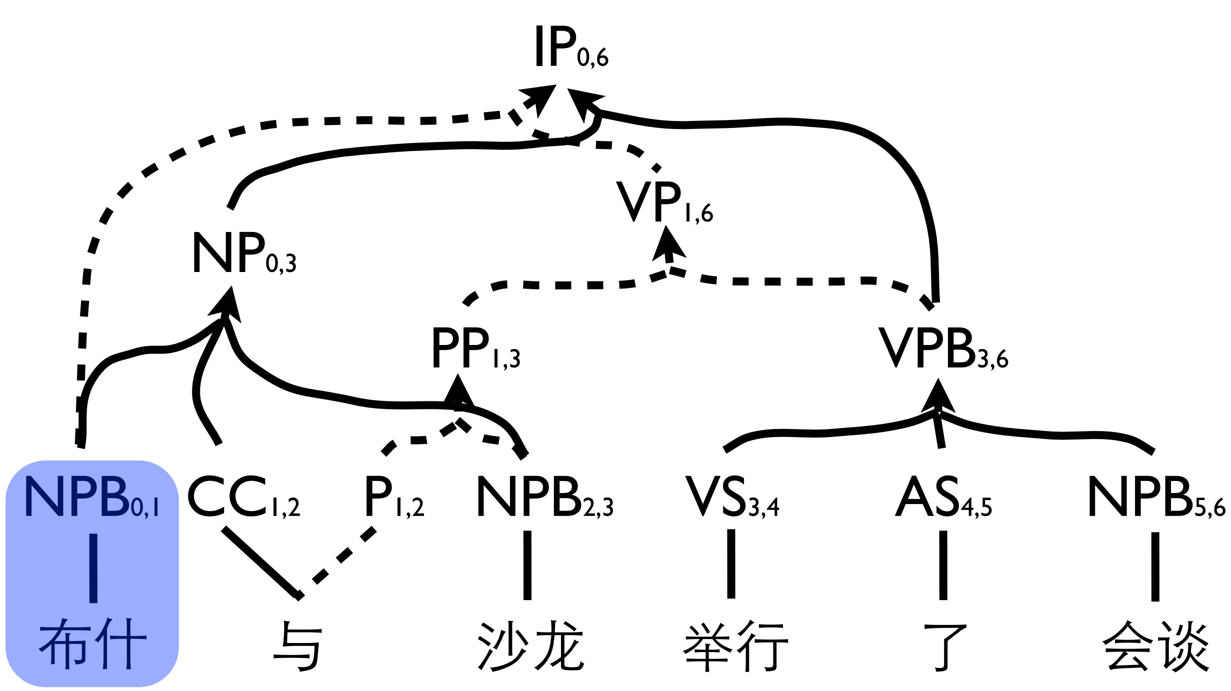




(Mi et al., 2008)

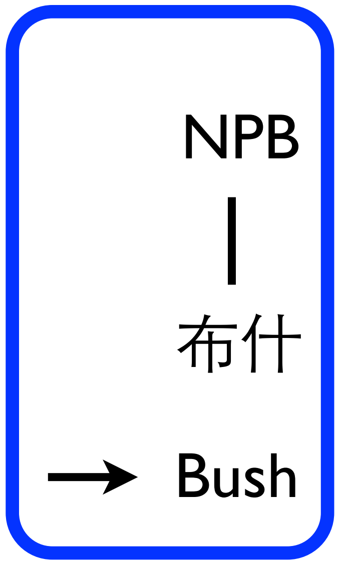
Forest-based Decoding

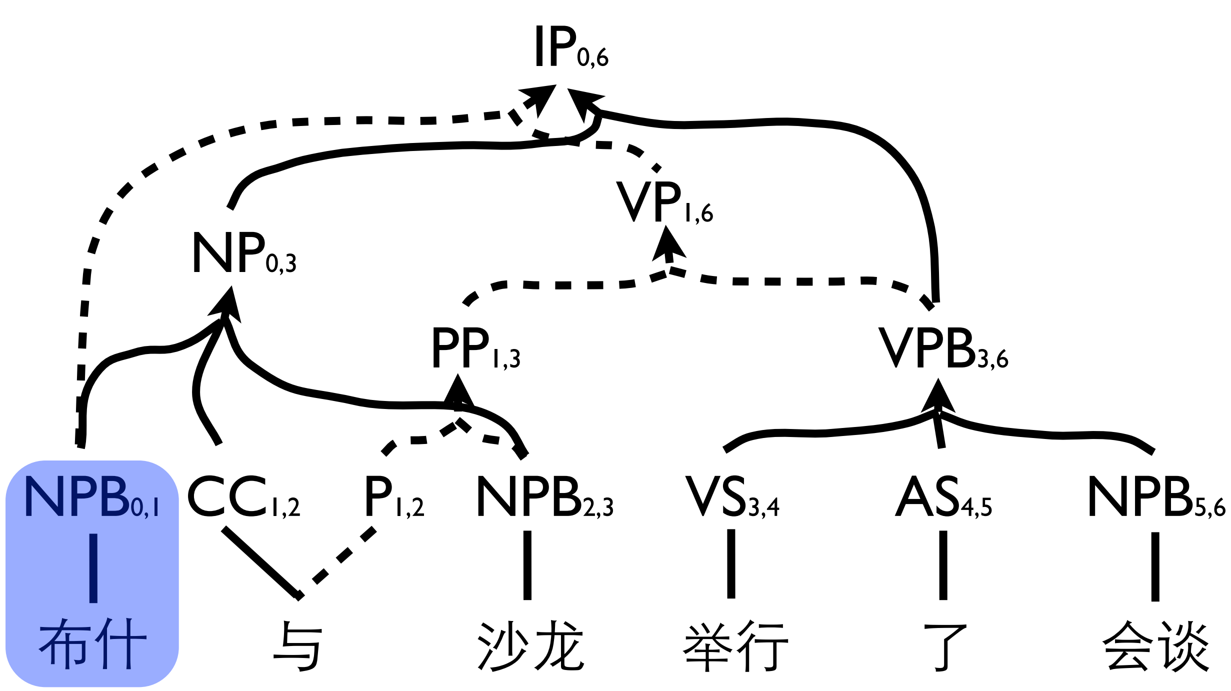




(Mi et al., 2008)

Forest-based Decoding

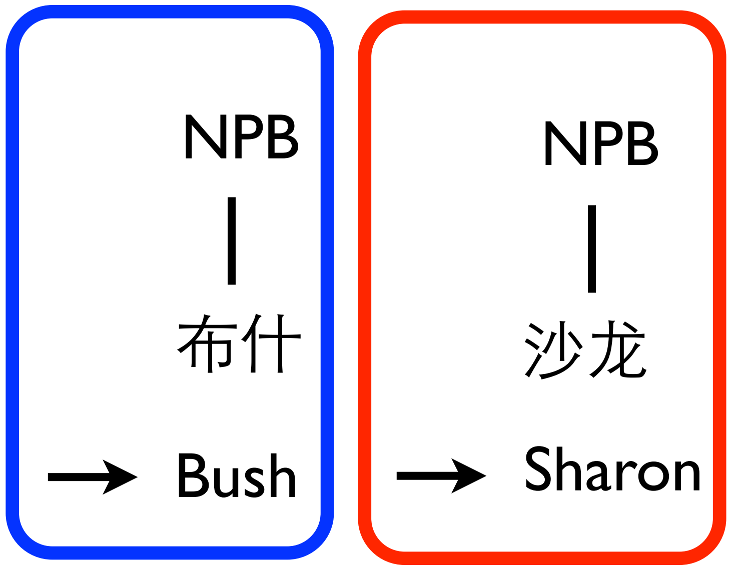


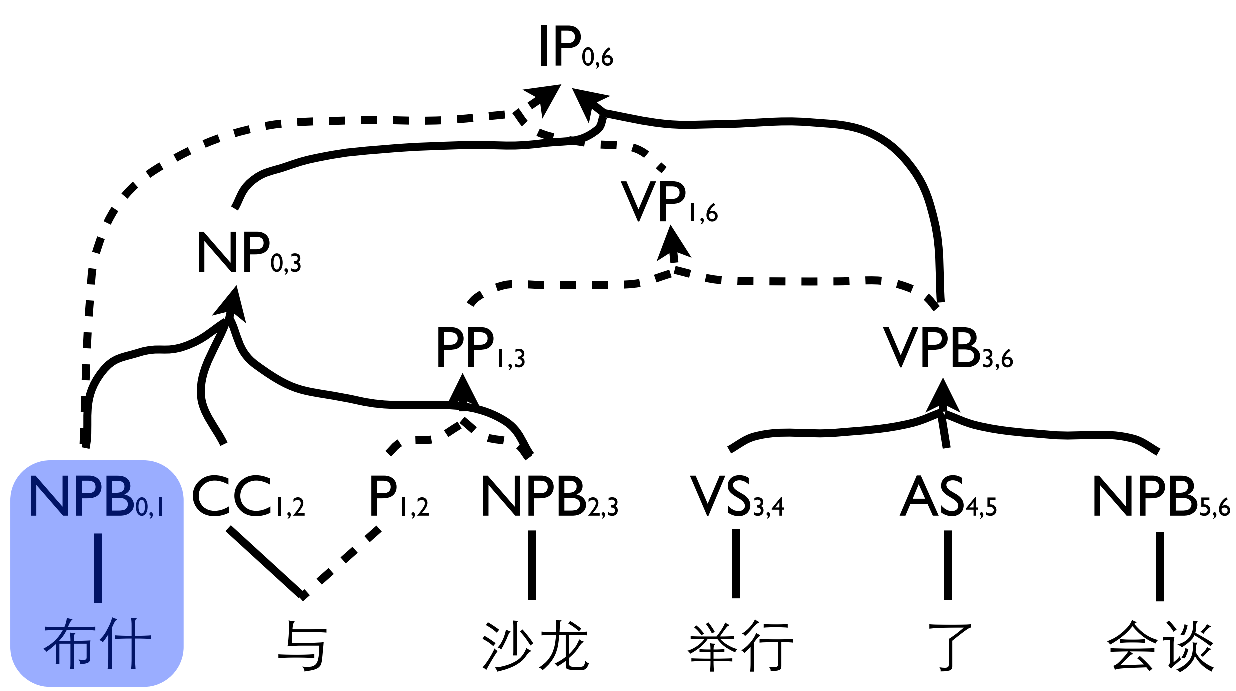


NPB$_{0,1}$

(Mi et al., 2008)

Forest-based Decoding

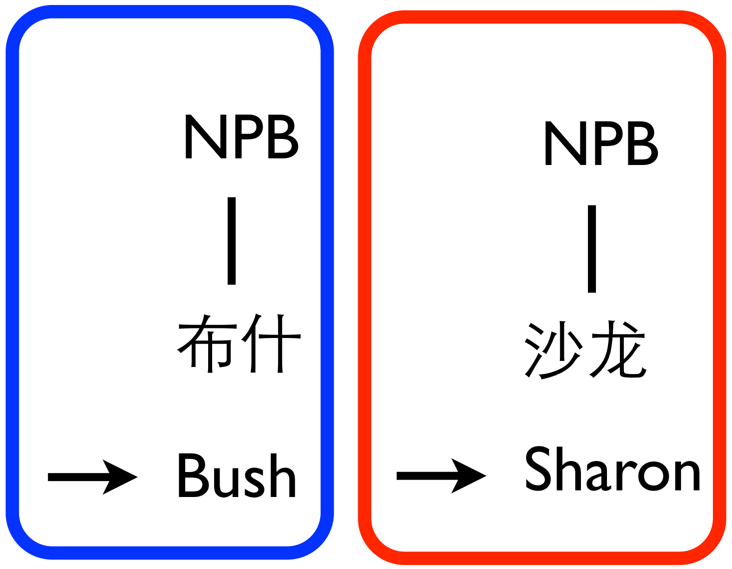


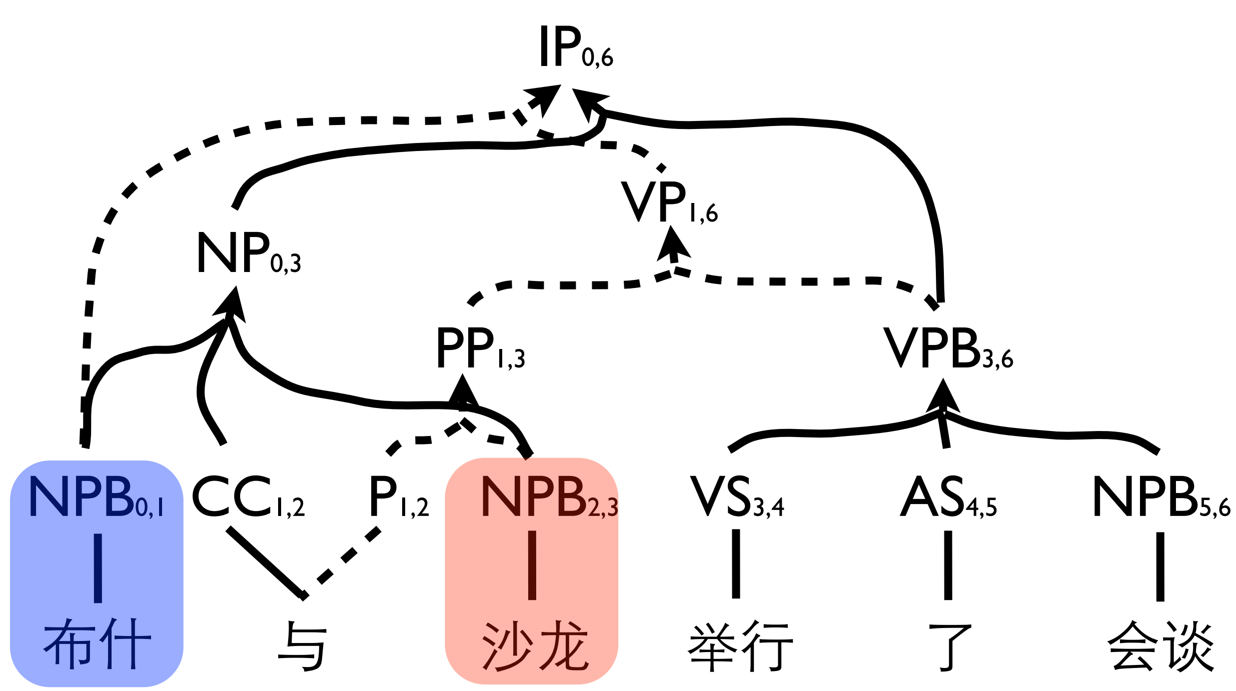


NPB0,1

(Mi et al., 2008)

Forest-based Decoding

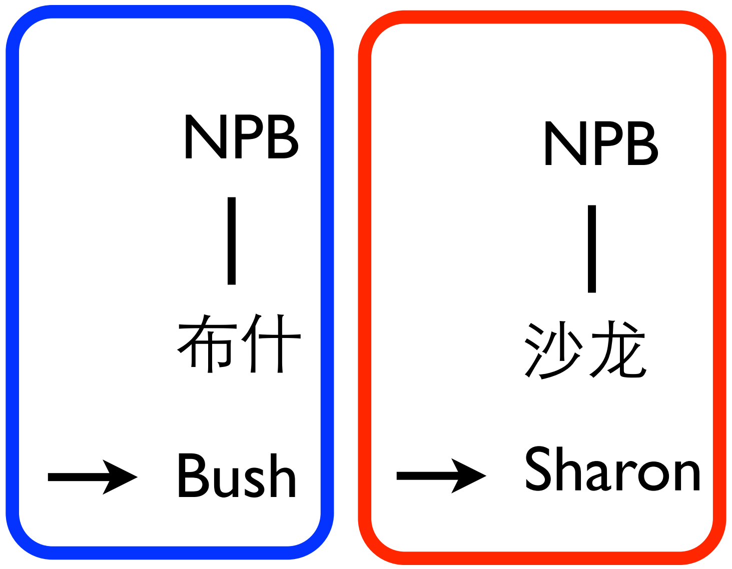


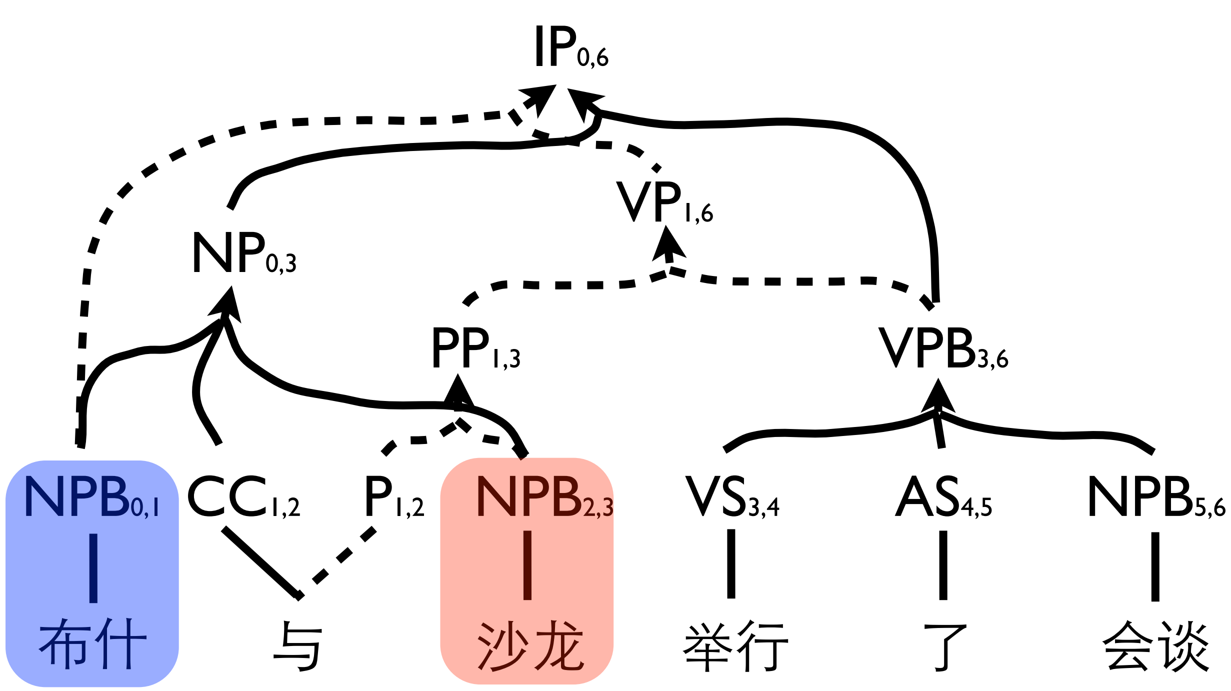


$NPB_{0,1}$

(Mi et al., 2008)

Forest-based Decoding

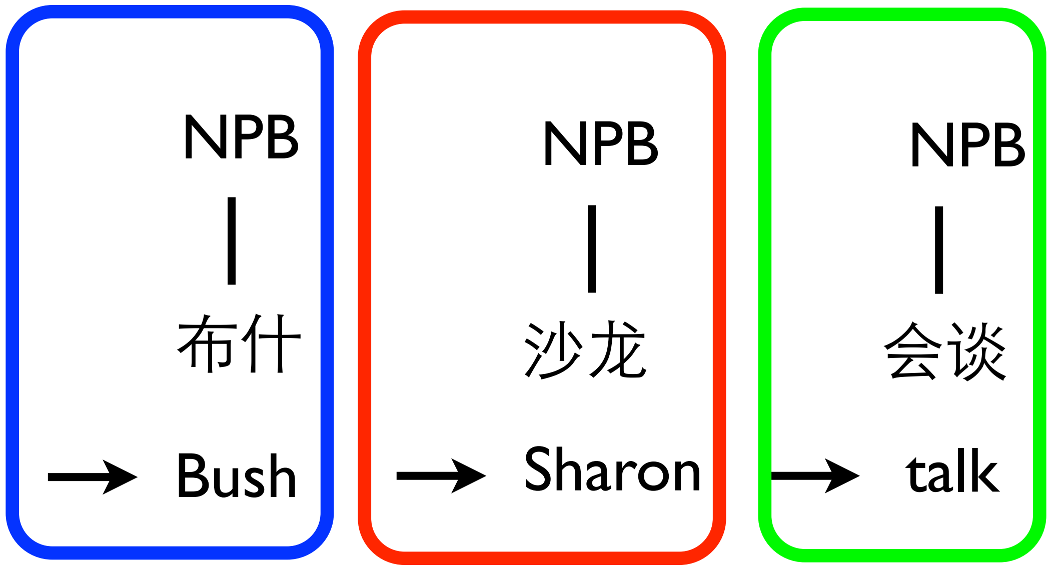


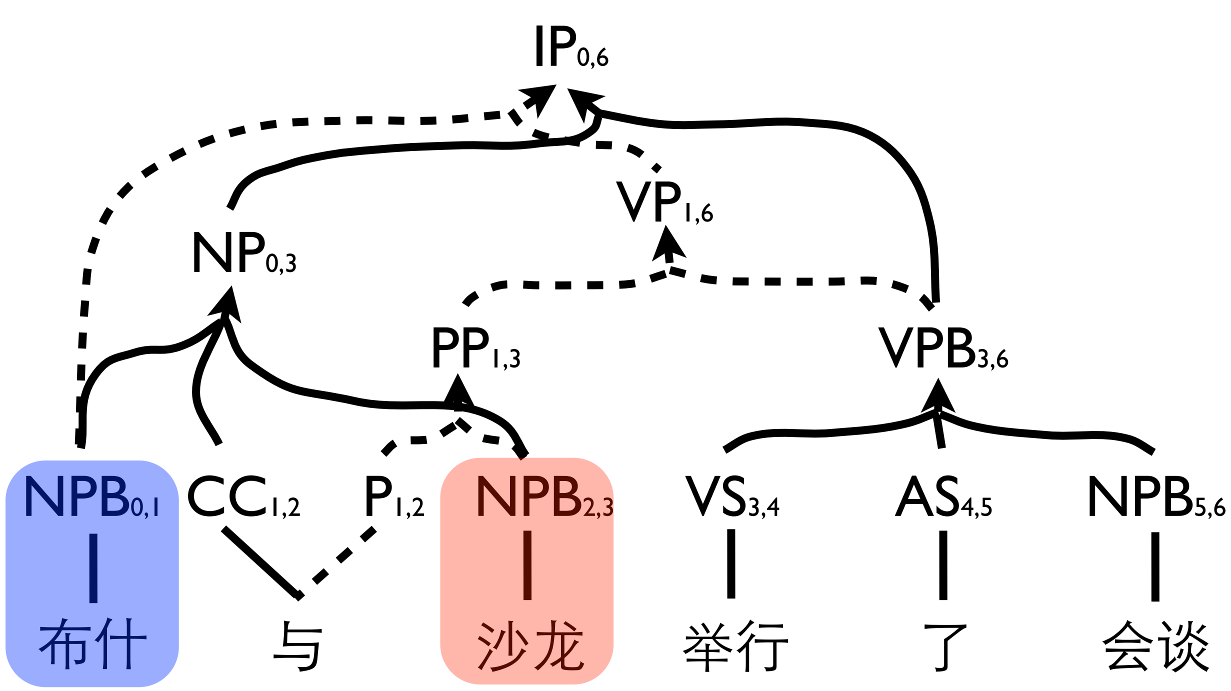


$NPB_{0,1}$ $NPB_{2,3}$

(Mi et al., 2008)

Forest-based Decoding

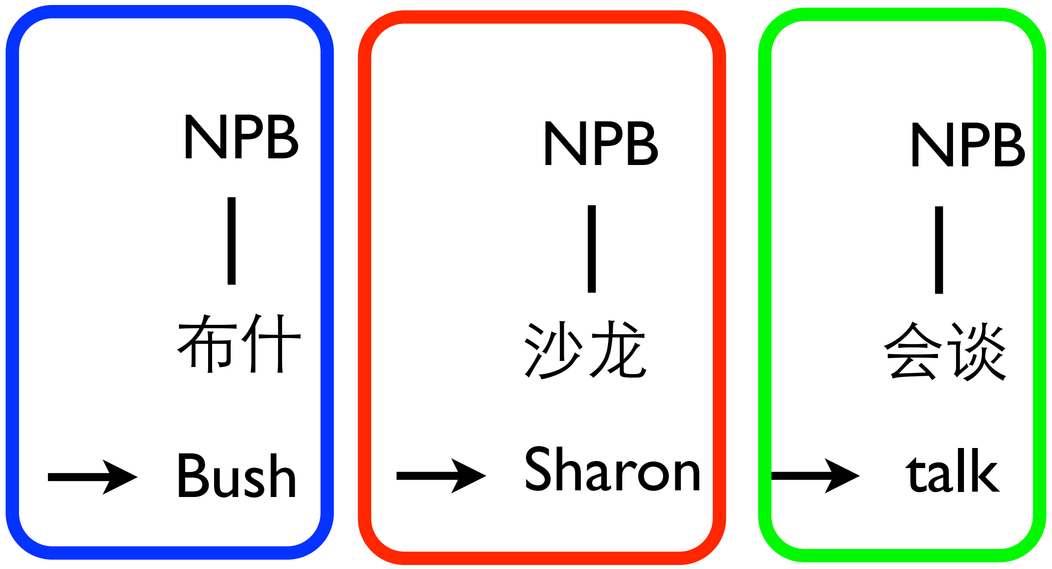


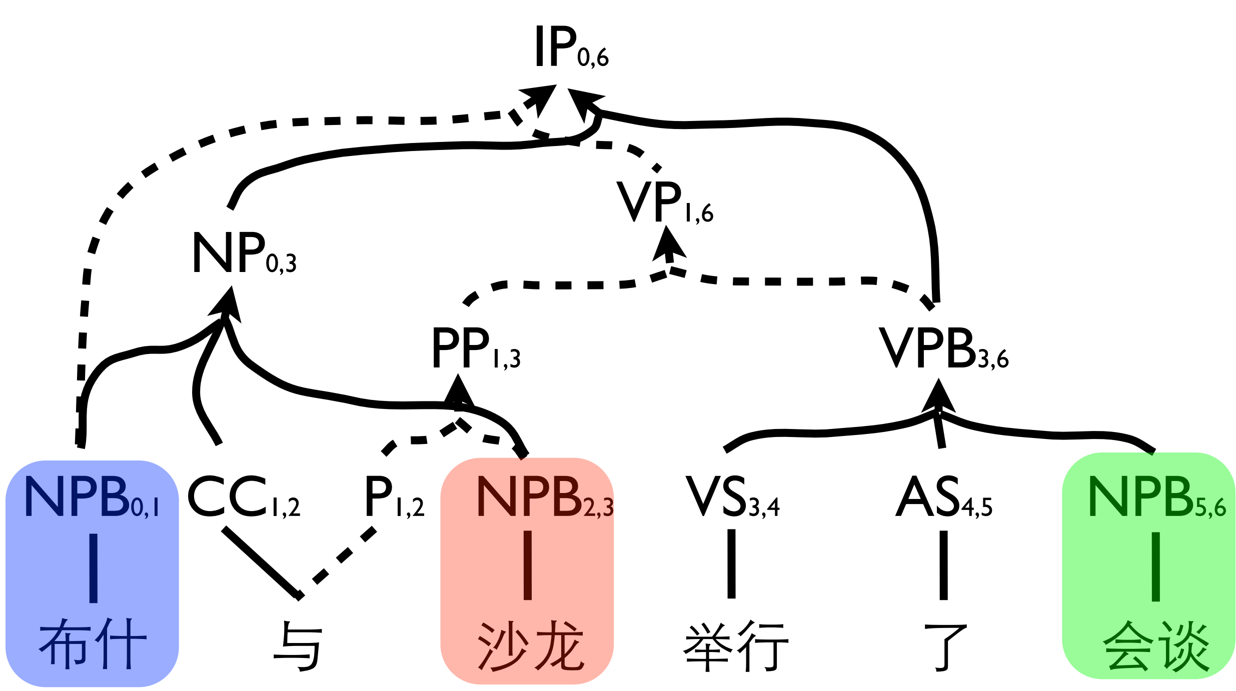


$NPB_{0,1}$ $NPB_{2,3}$

(Mi et al., 2008)

Forest-based Decoding

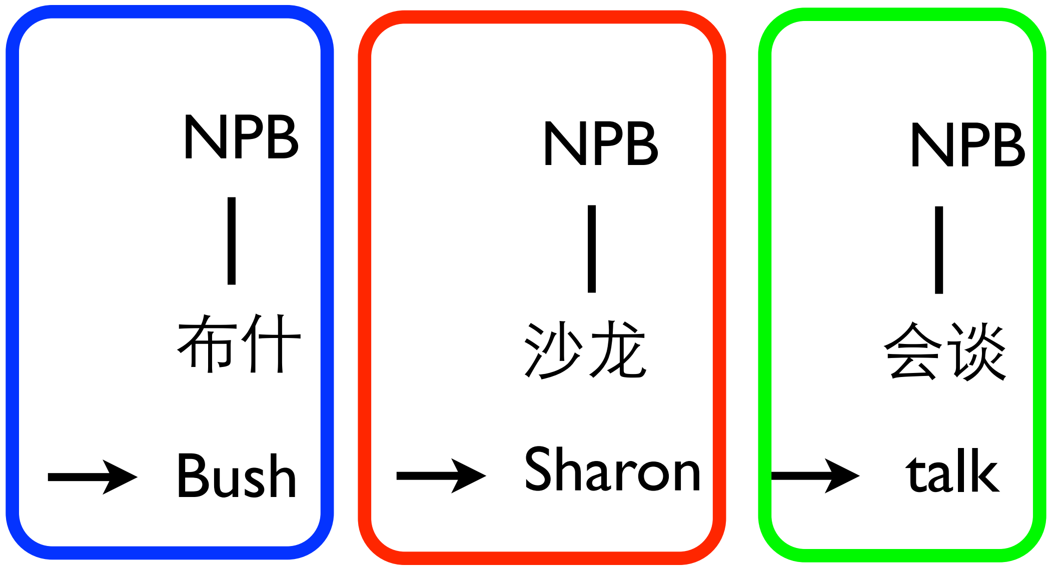


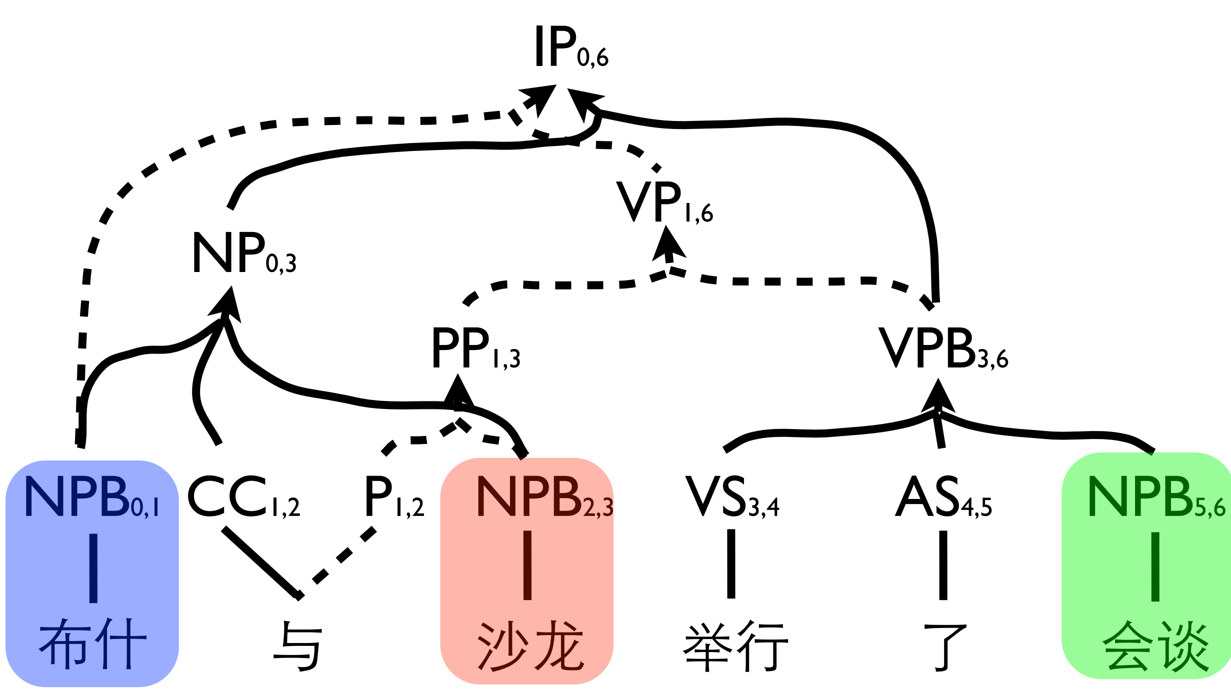


NPB0,1 NPB2,3

(Mi et al., 2008)

Forest-based Decoding

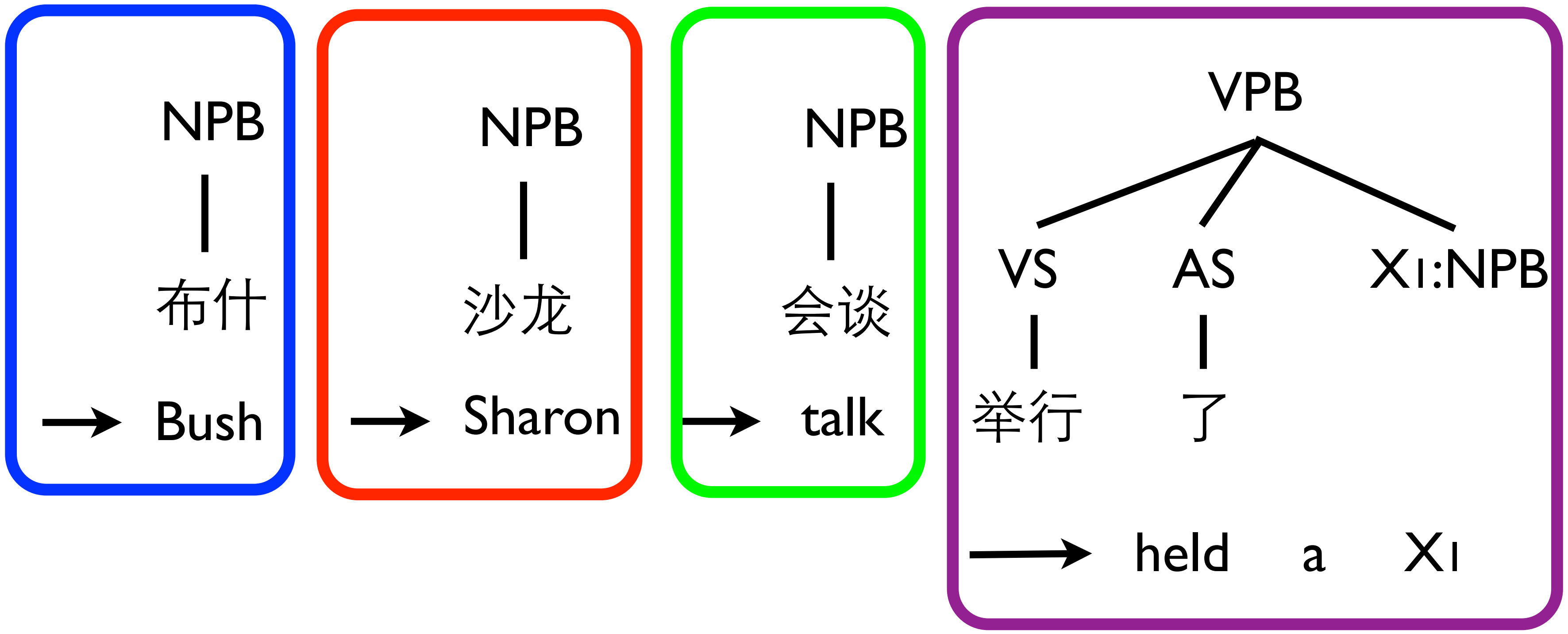


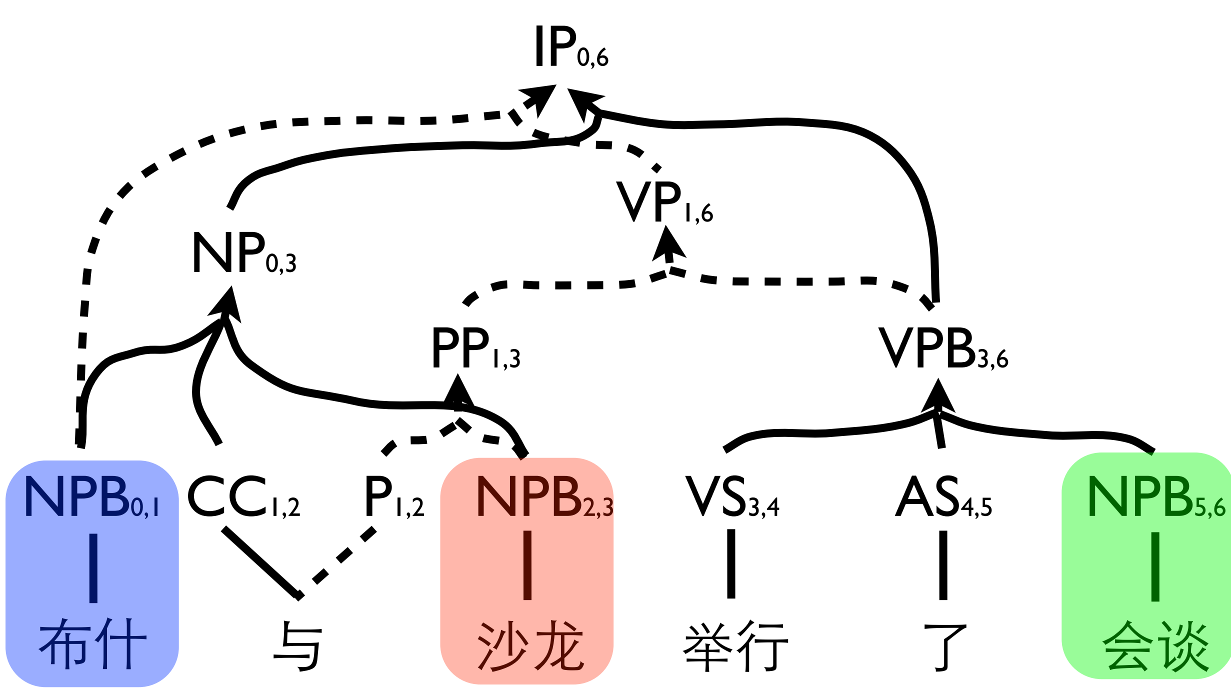


NPB0,1 NPB2,3 NPB5,6

(Mi et al., 2008)

Forest-based Decoding

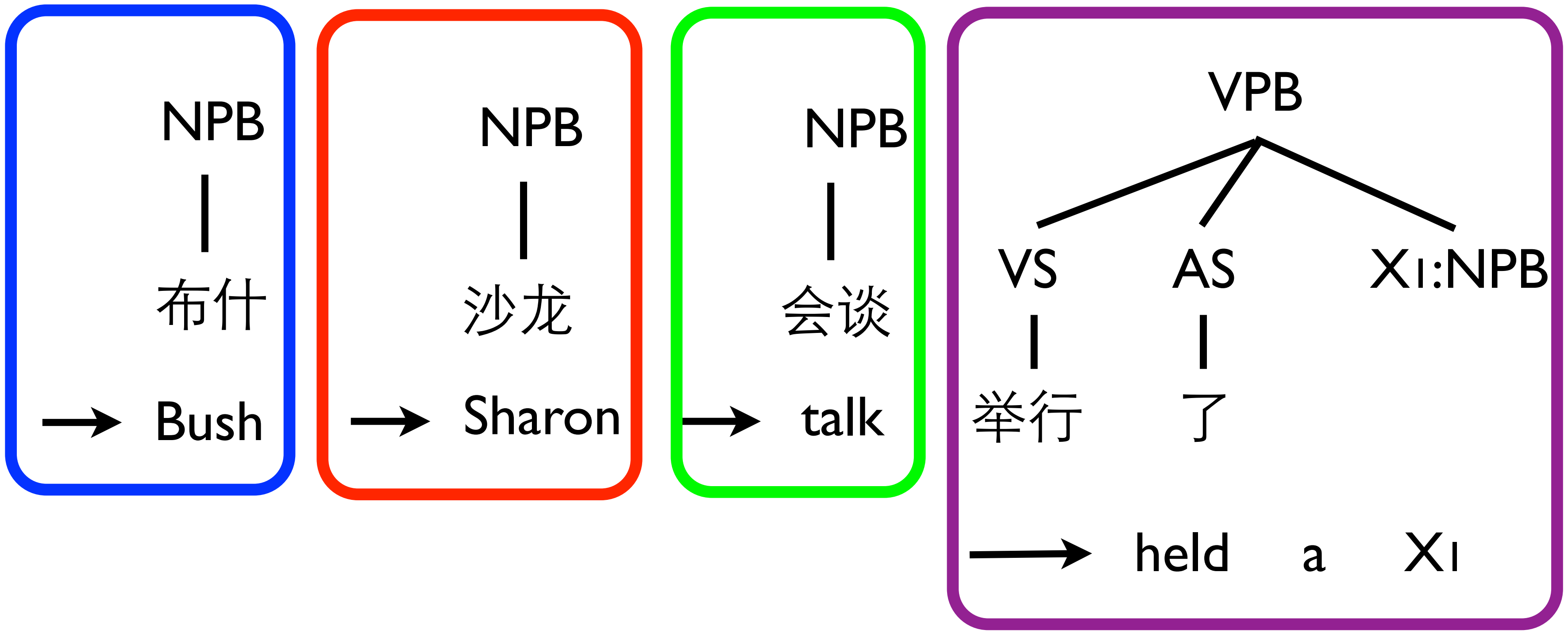


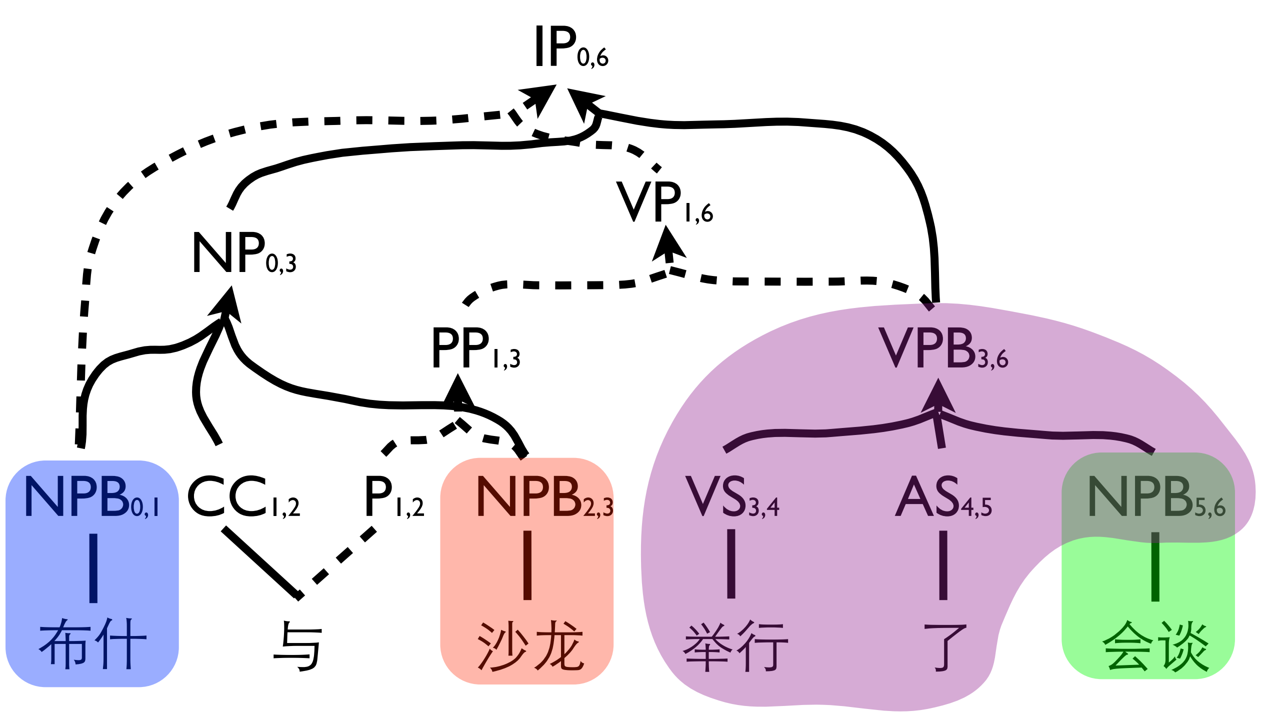


$NPB_{0,1}$ $NPB_{2,3}$ $NPB_{5,6}$

(Mi et al., 2008)

Forest-based Decoding

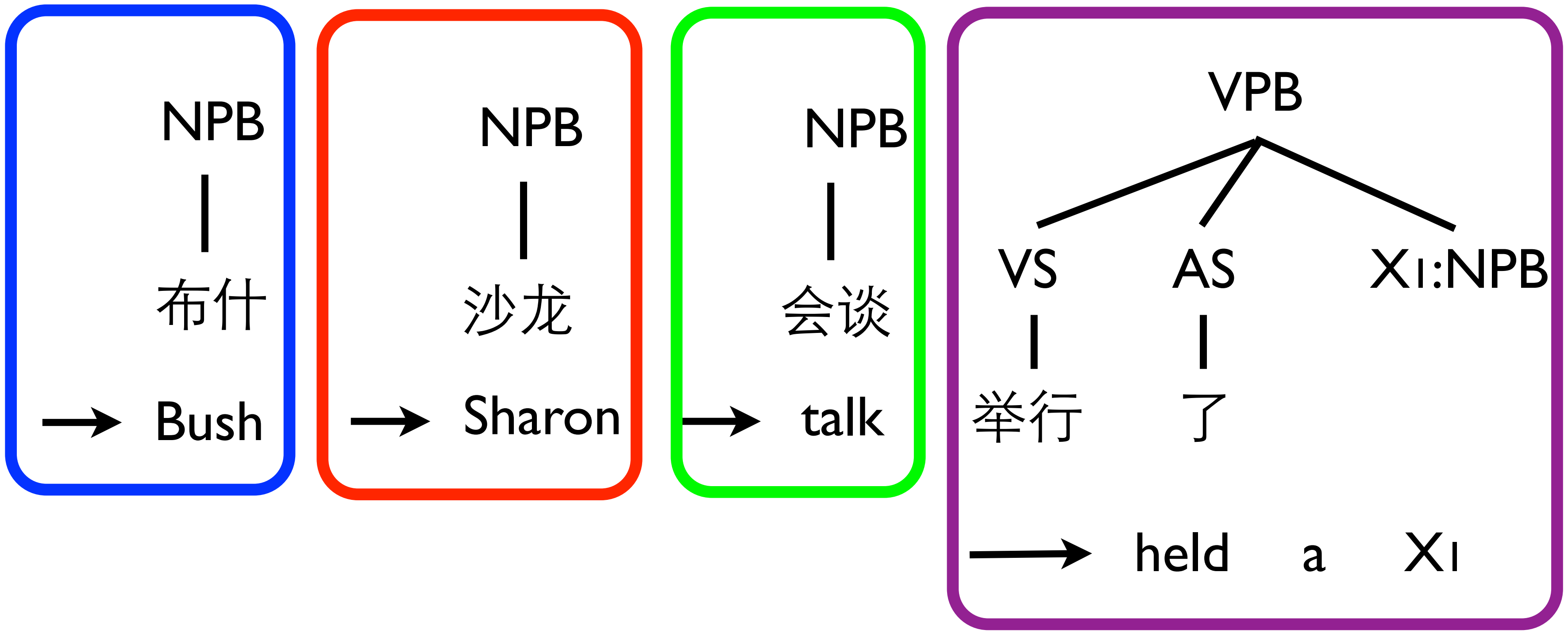


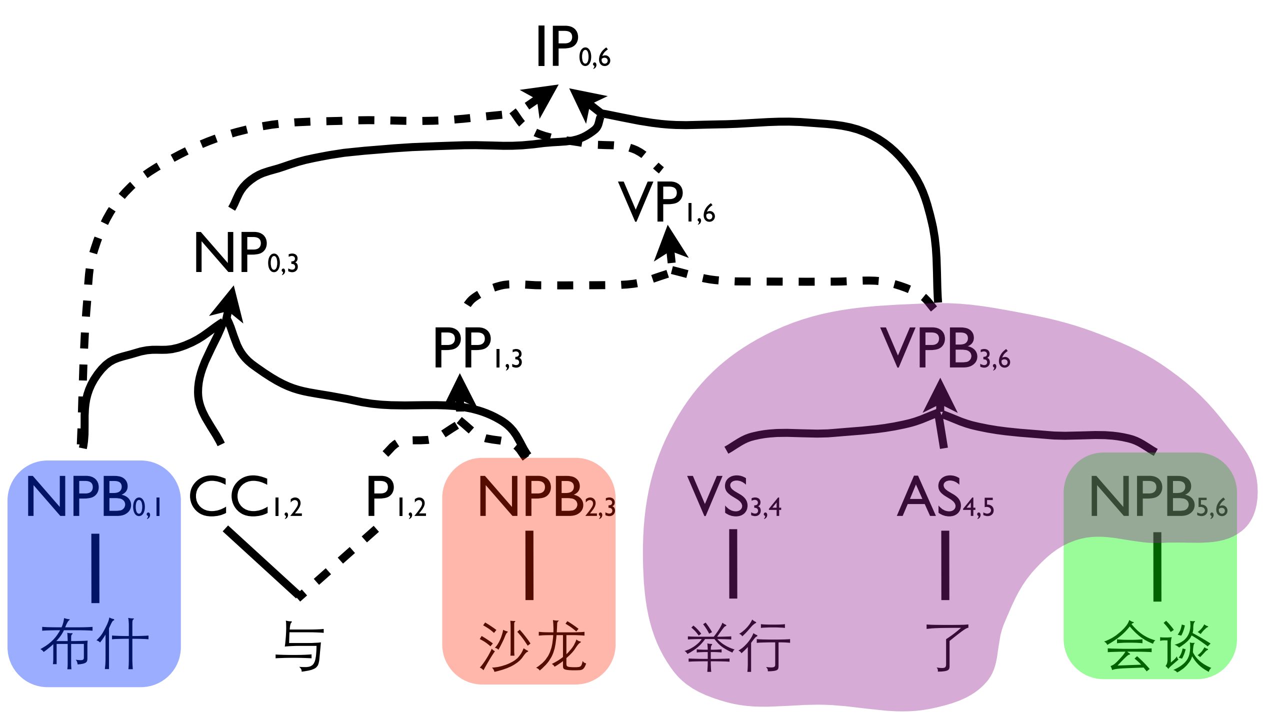


NPB0,1 NPB2,3 NPB5,6

(Mi et al., 2008)

Forest-based Decoding



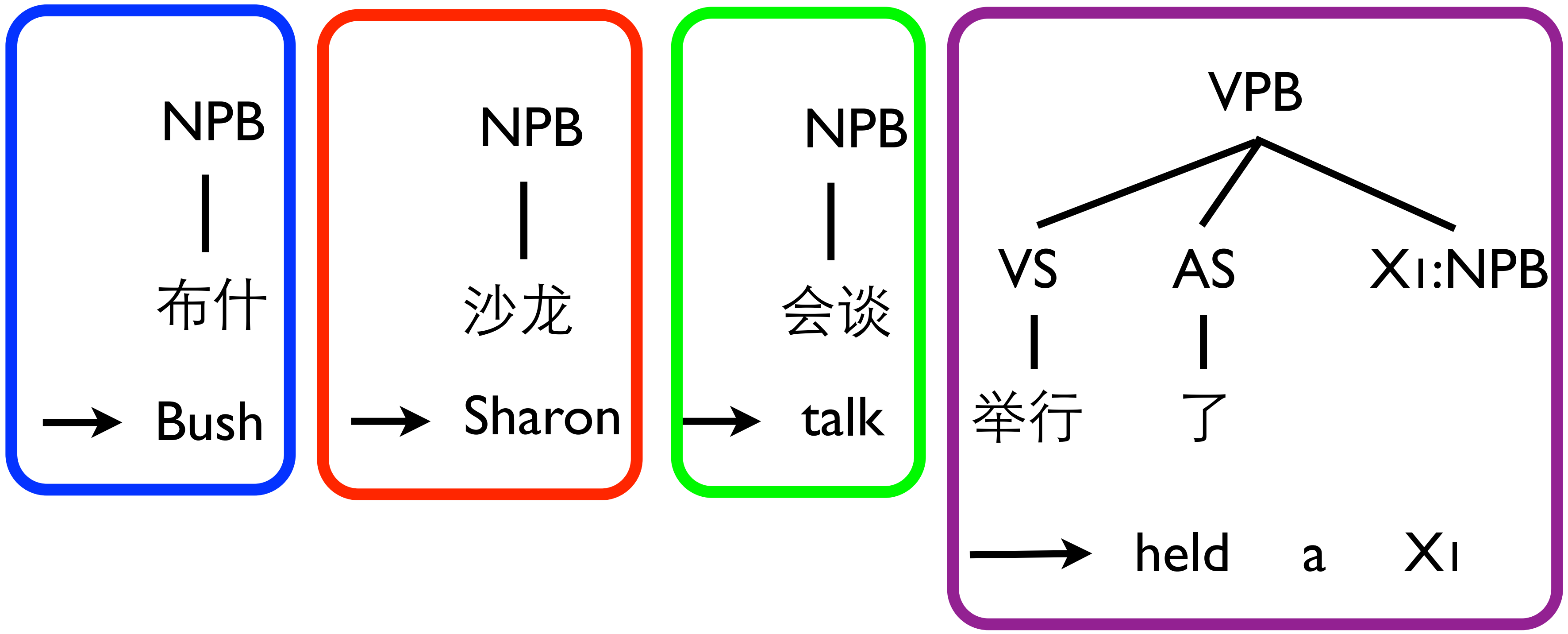


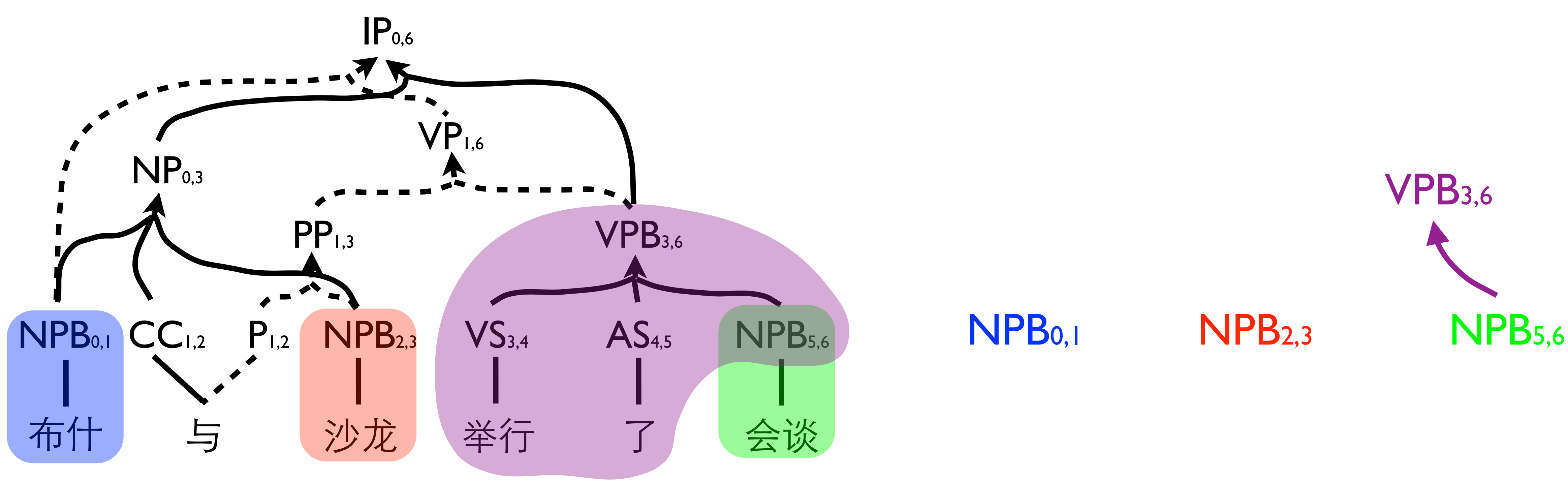
VPB3,6

NPB0,1 NPB2,3 NPB5,6

(Mi et al., 2008)

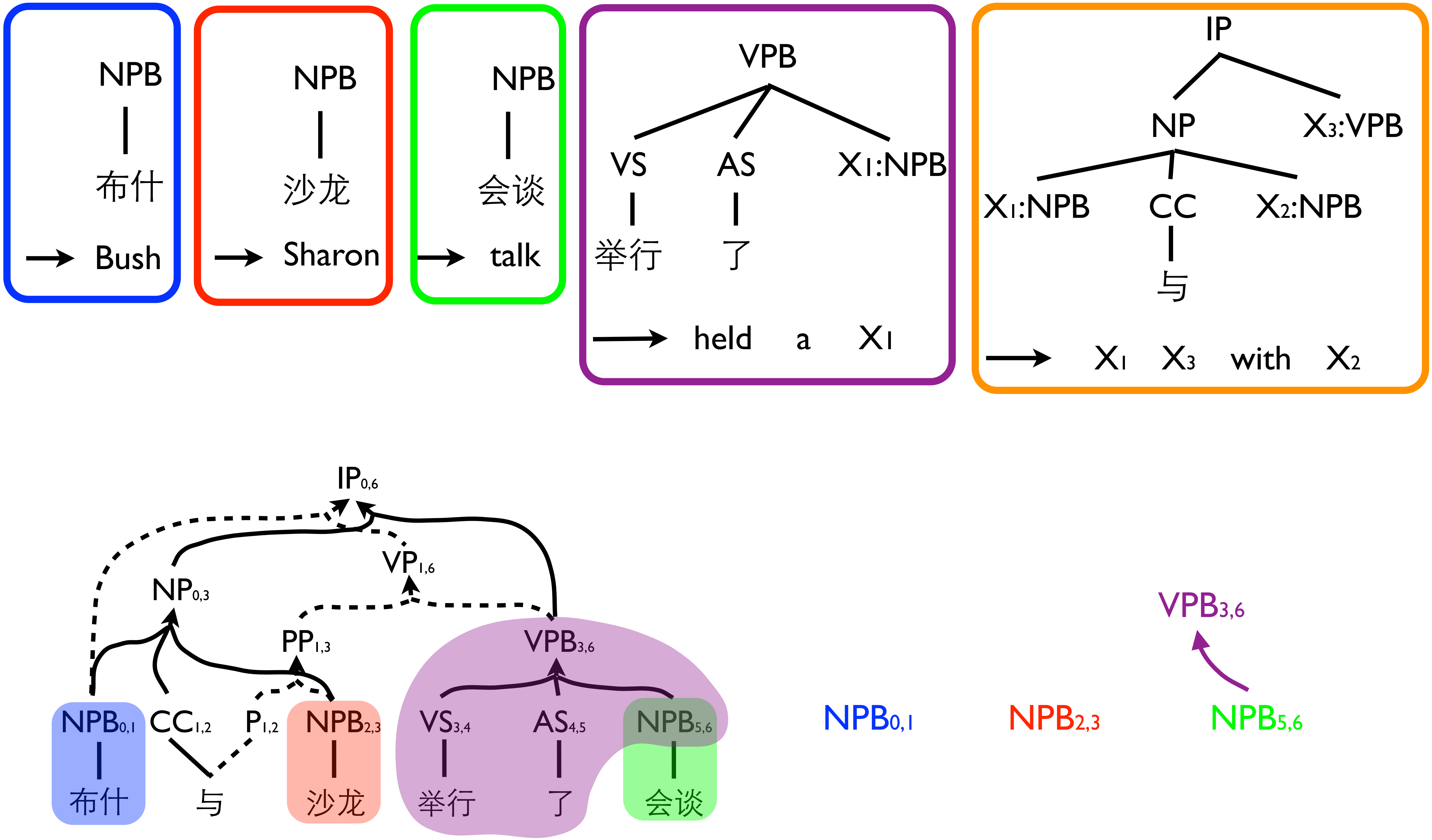
Forest-based Decoding





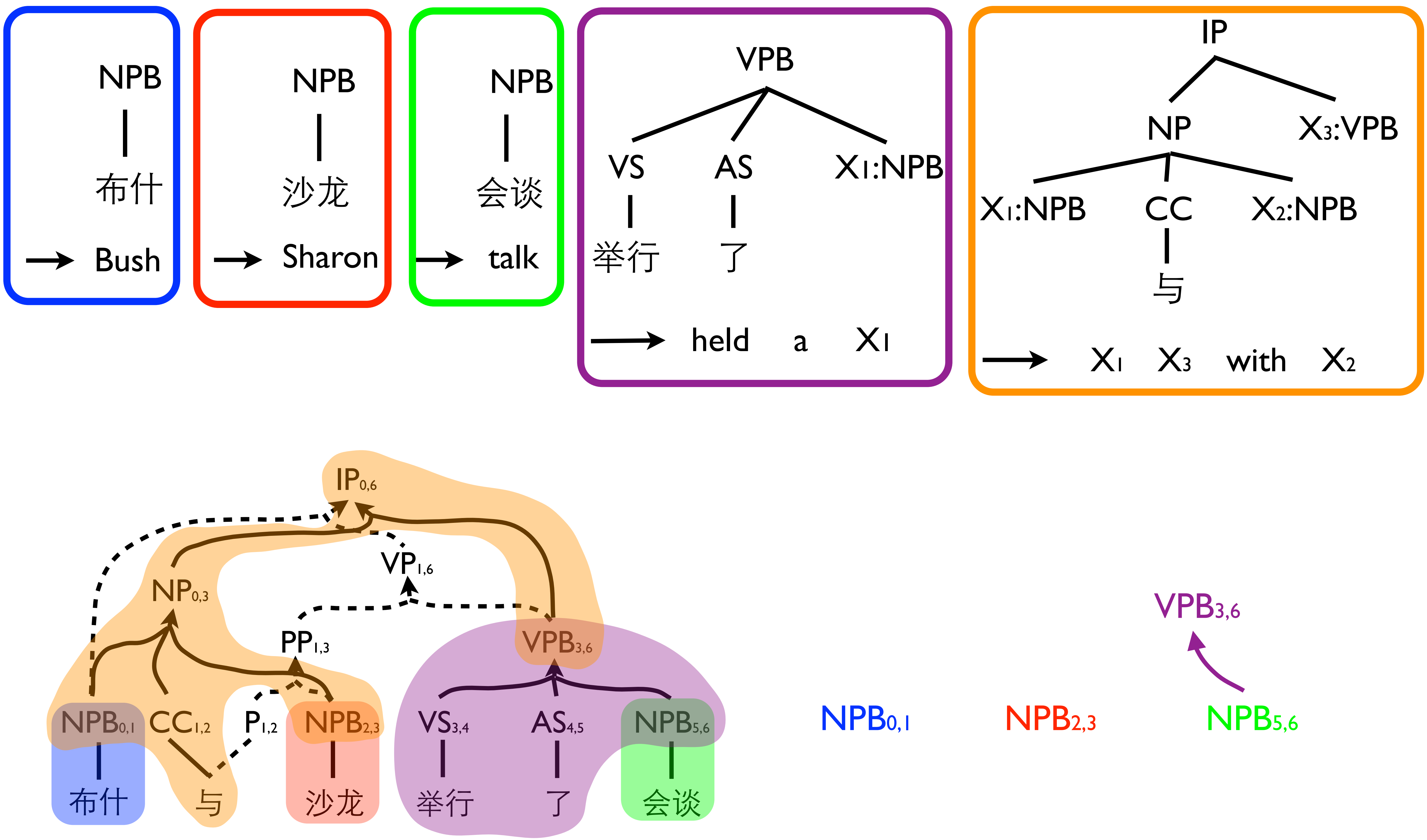
(Mi et al., 2008)

Forest-based Decoding



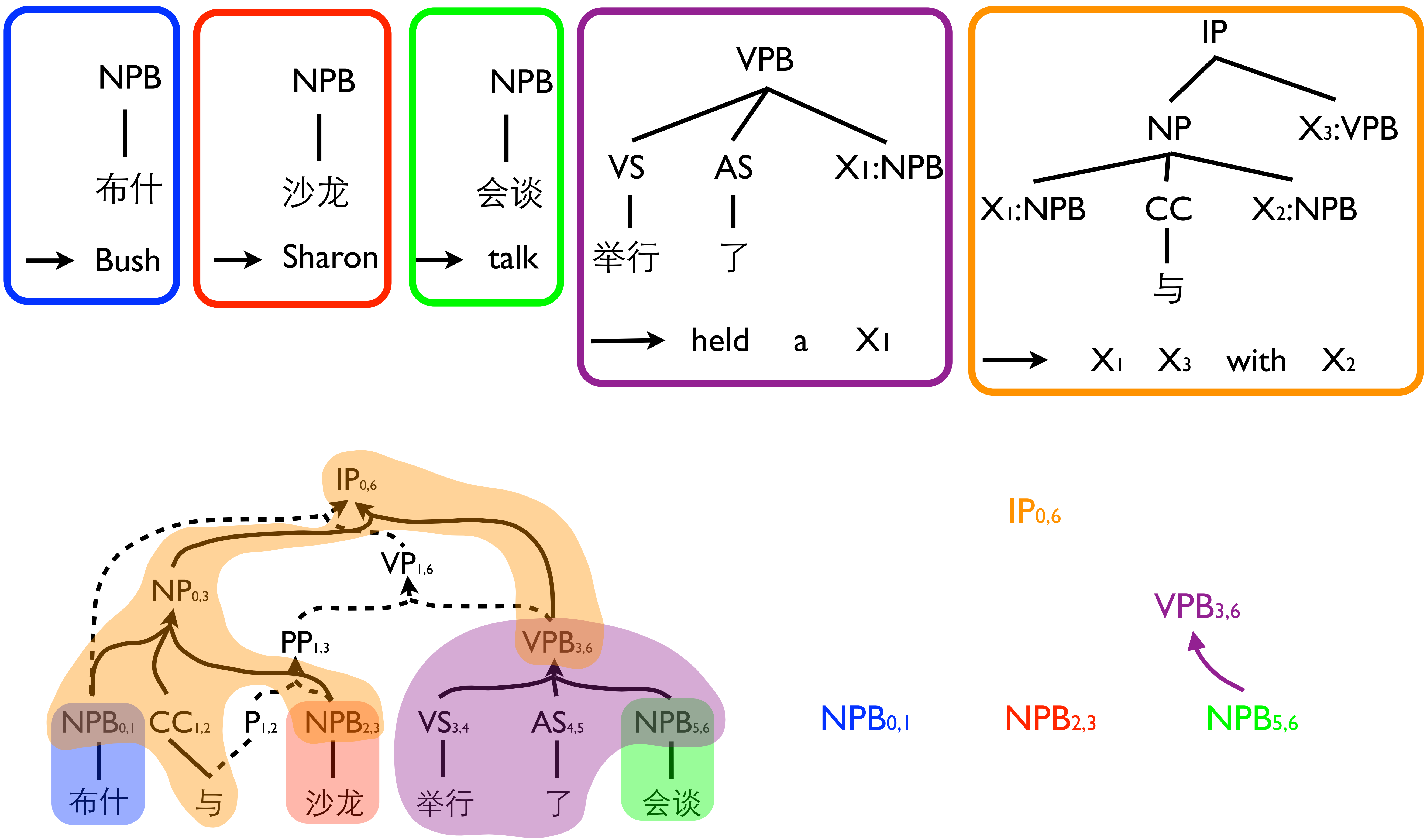
(Mi et al., 2008)

Forest-based Decoding



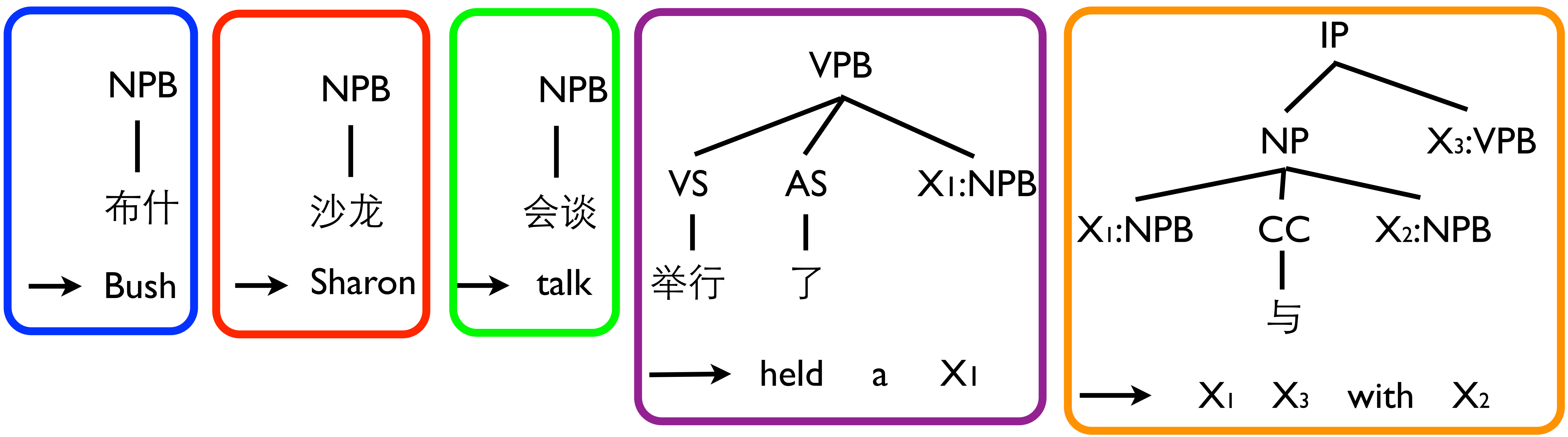
(Mi et al., 2008)

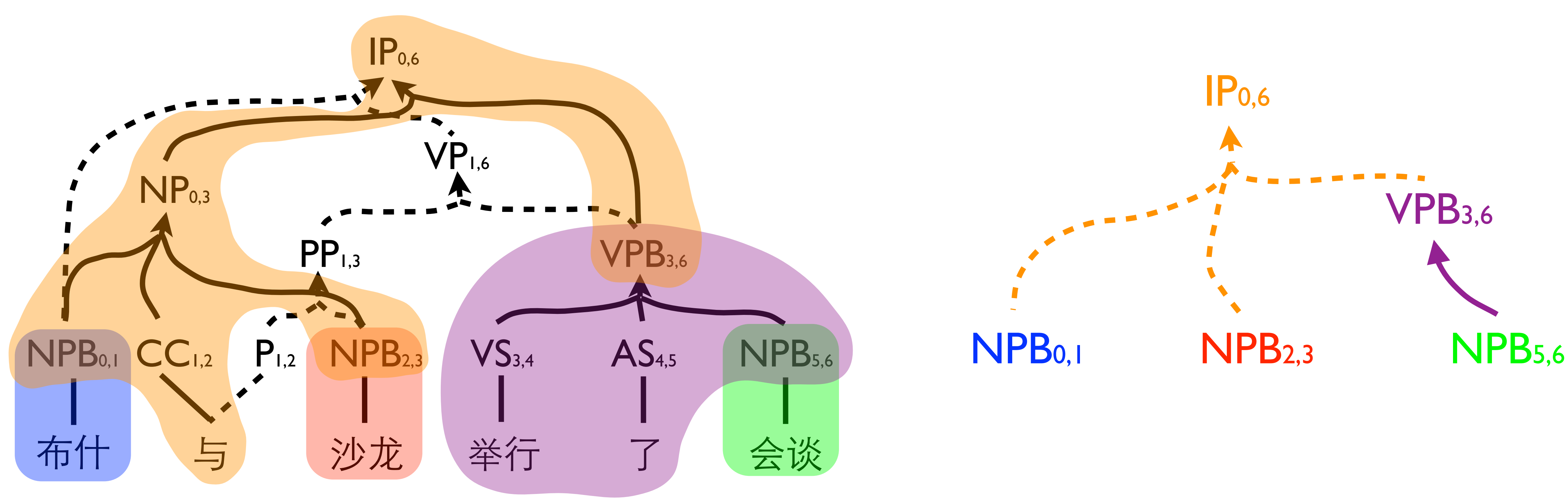
Forest-based Decoding



(Mi et al., 2008)

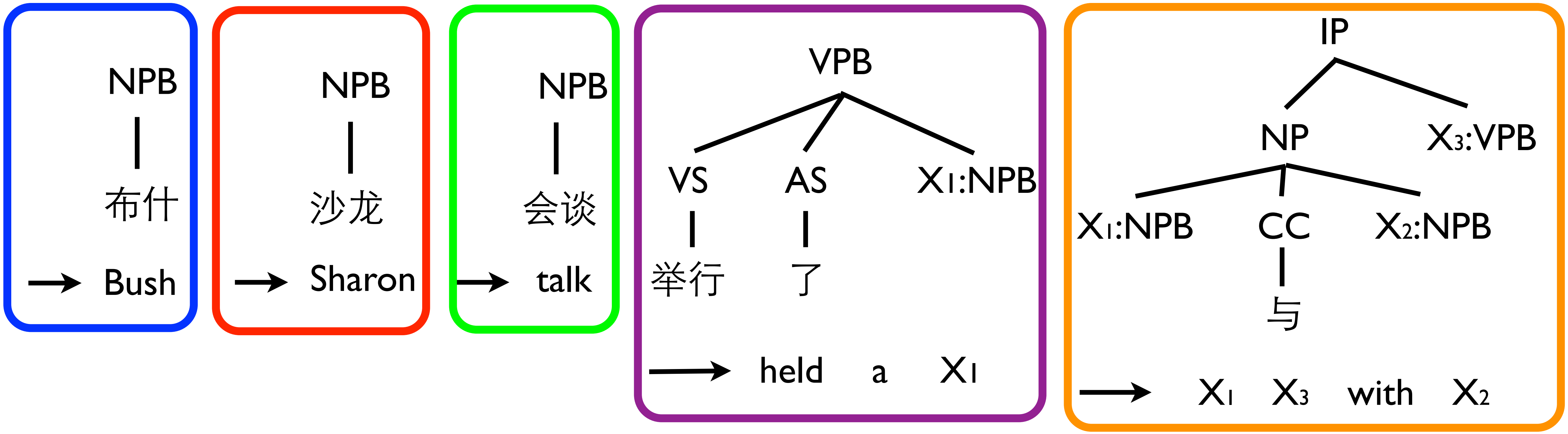
Forest-based Decoding

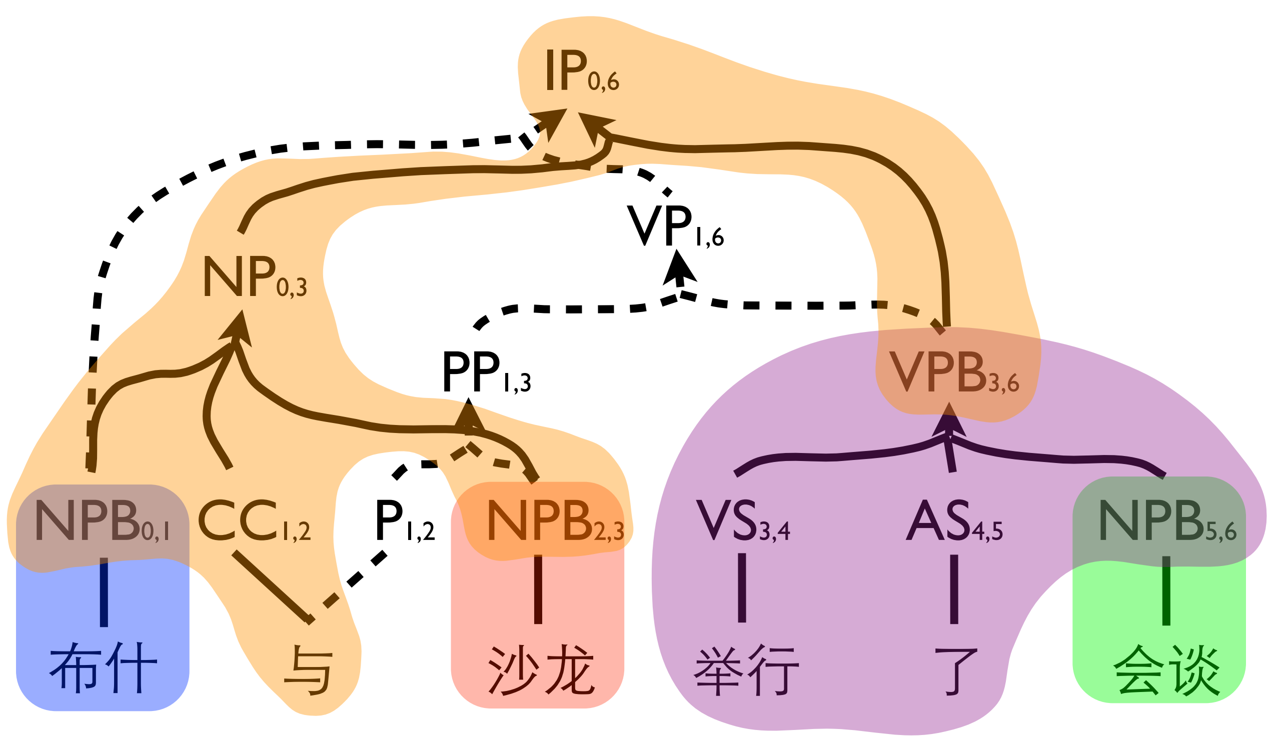


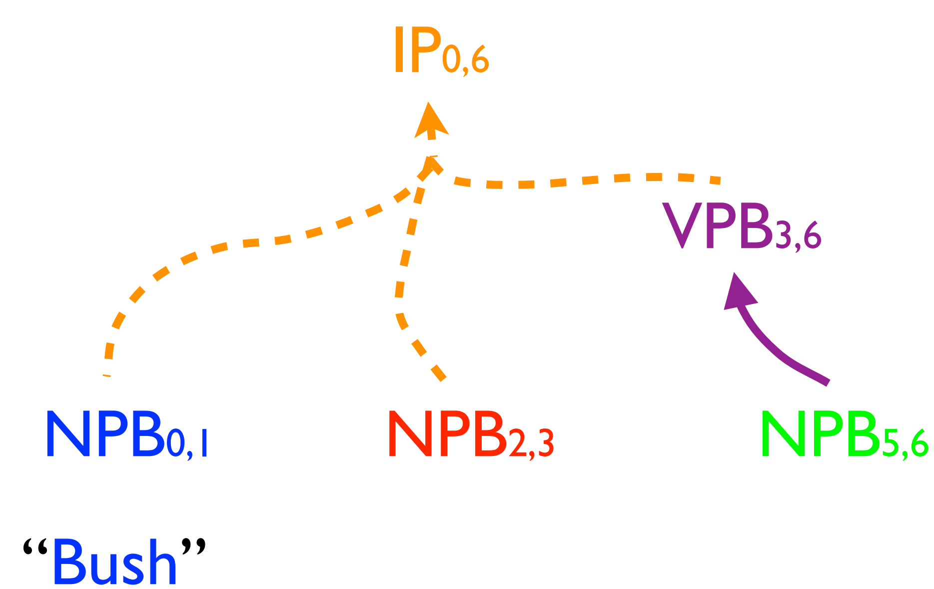


(Mi et al., 2008)

Forest-based Decoding

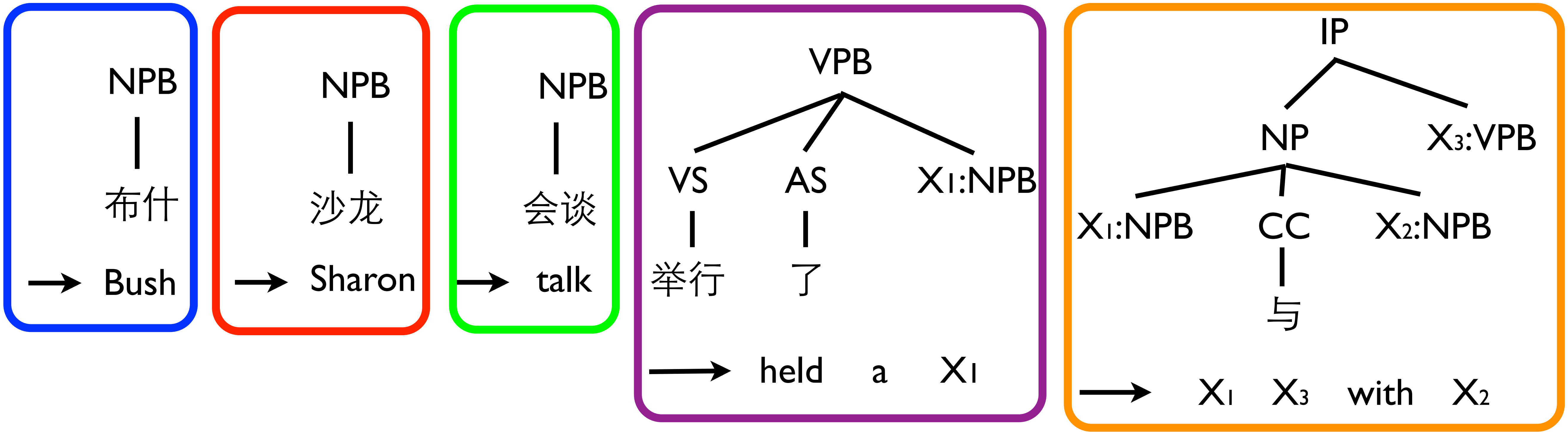


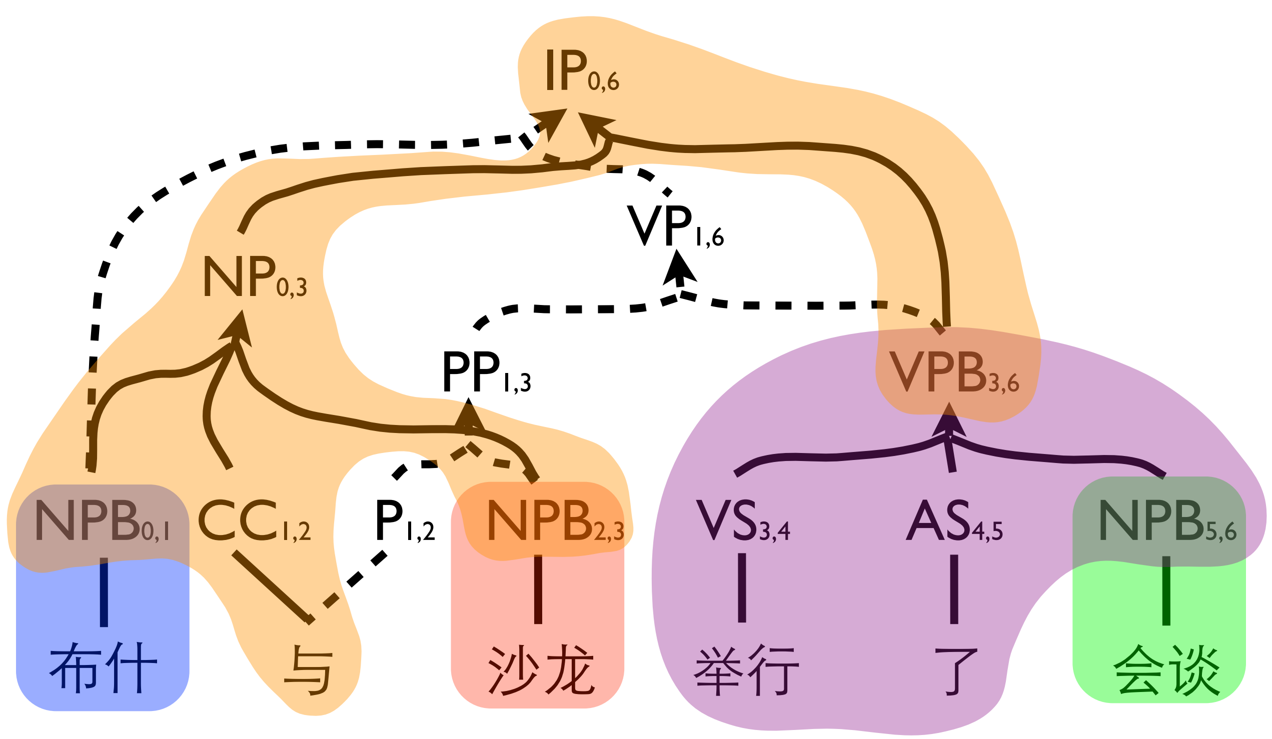


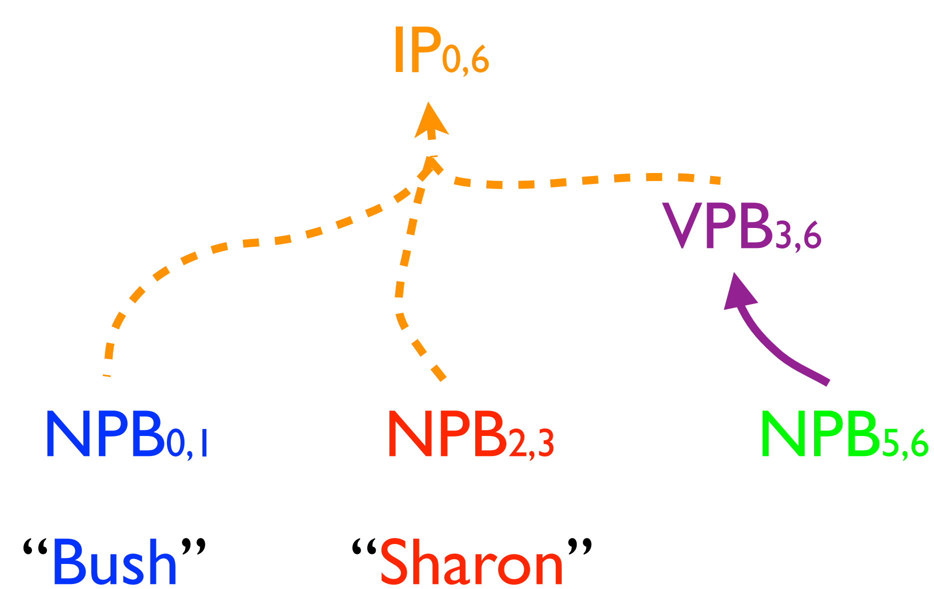


(Mi et al., 2008)

Forest-based Decoding

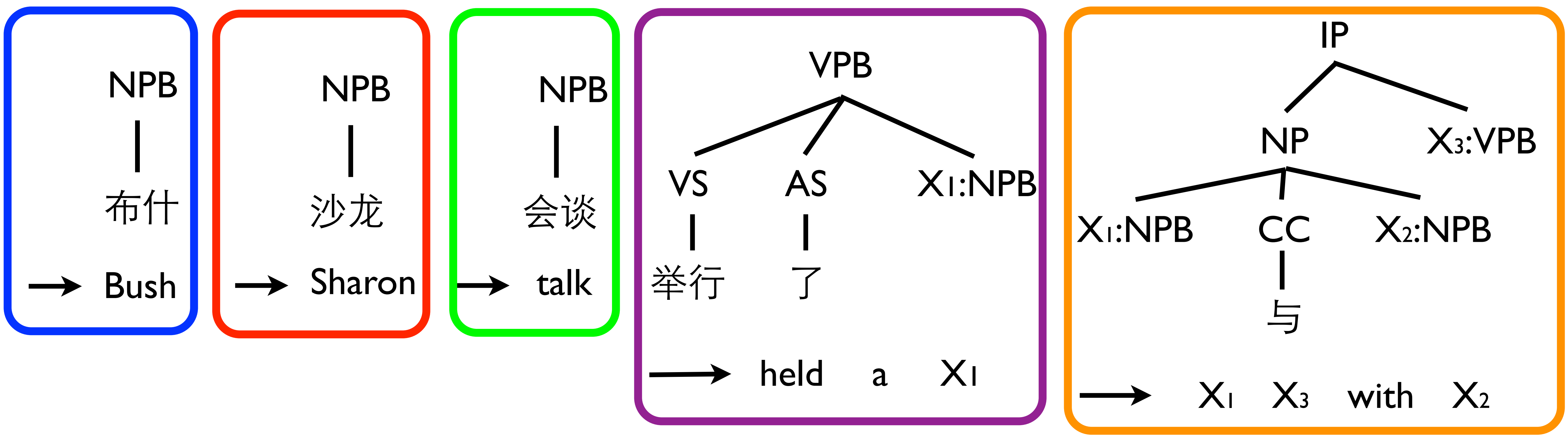


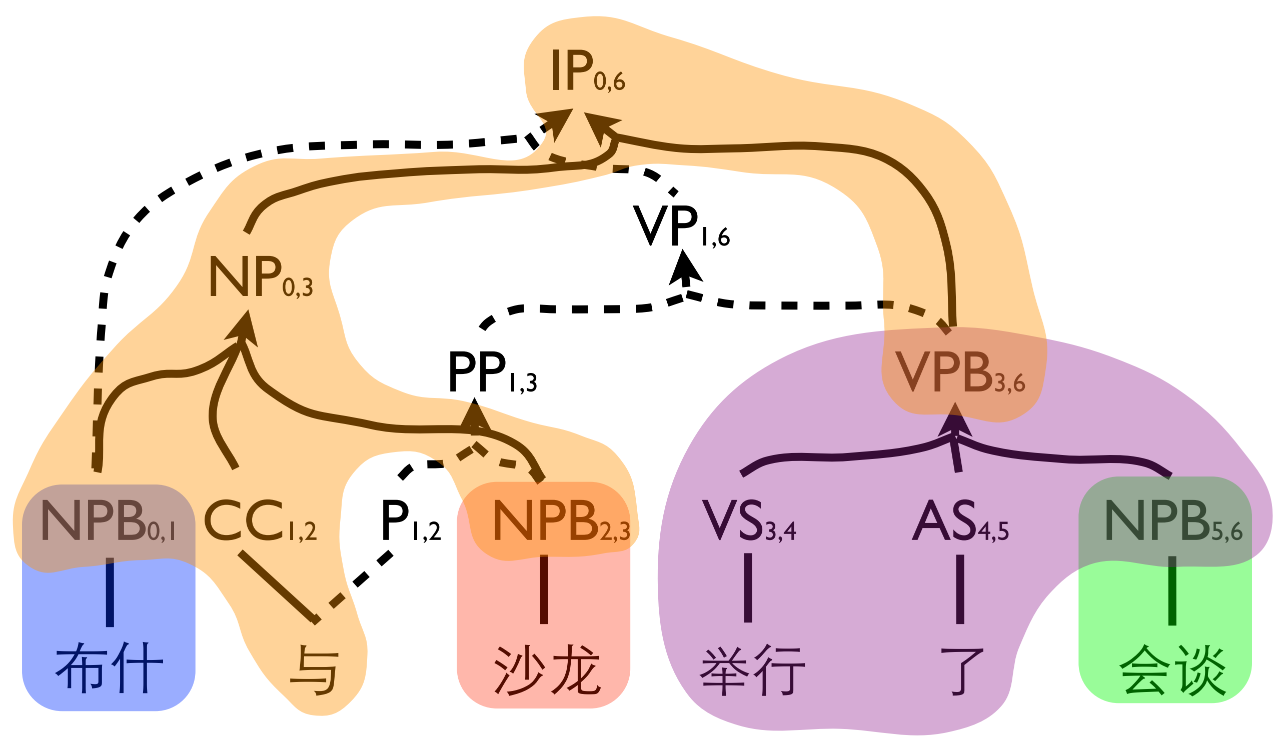


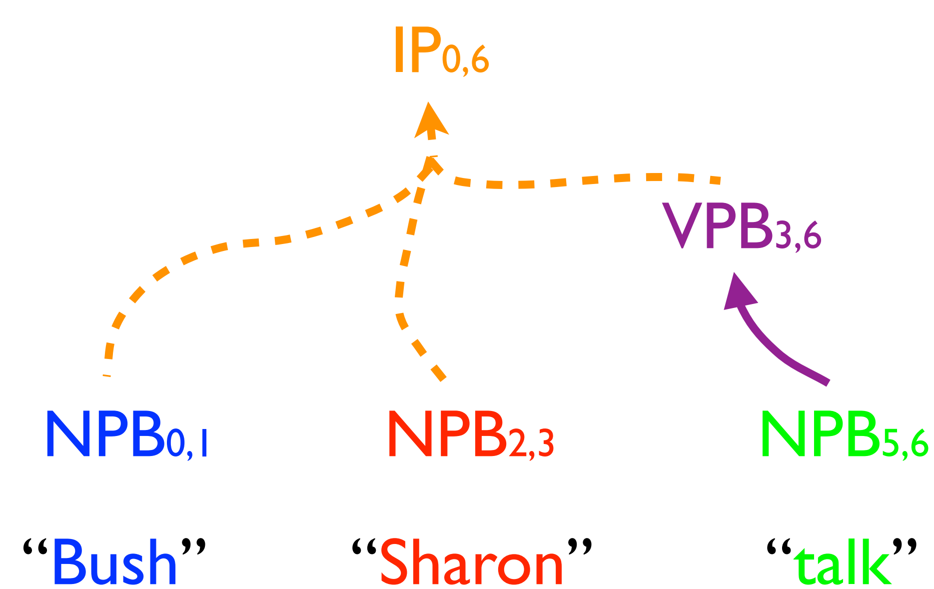


(Mi et al., 2008)

Forest-based Decoding

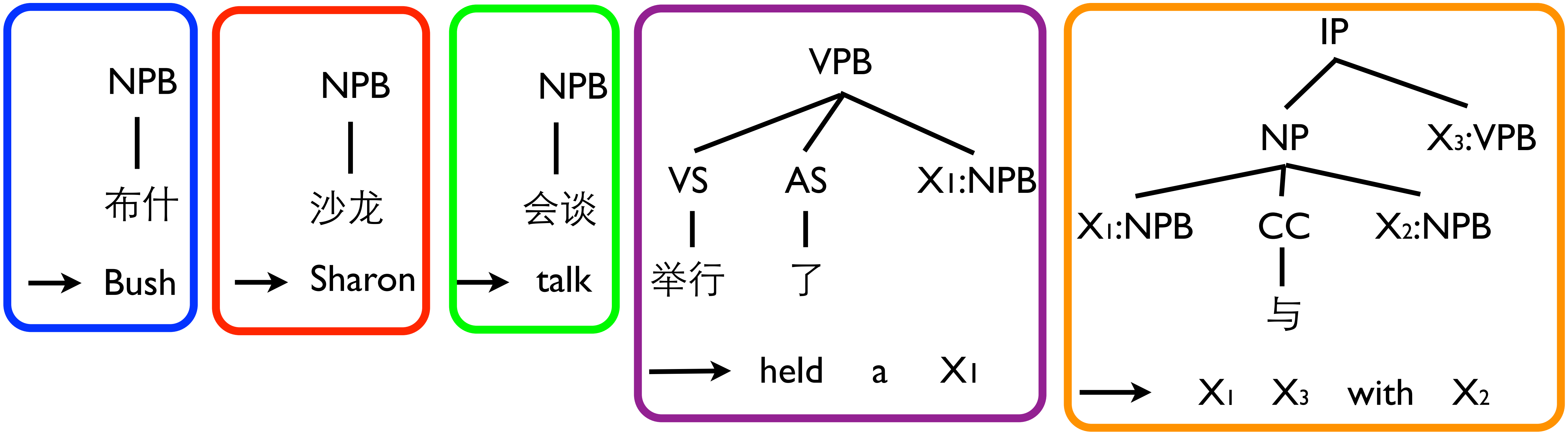


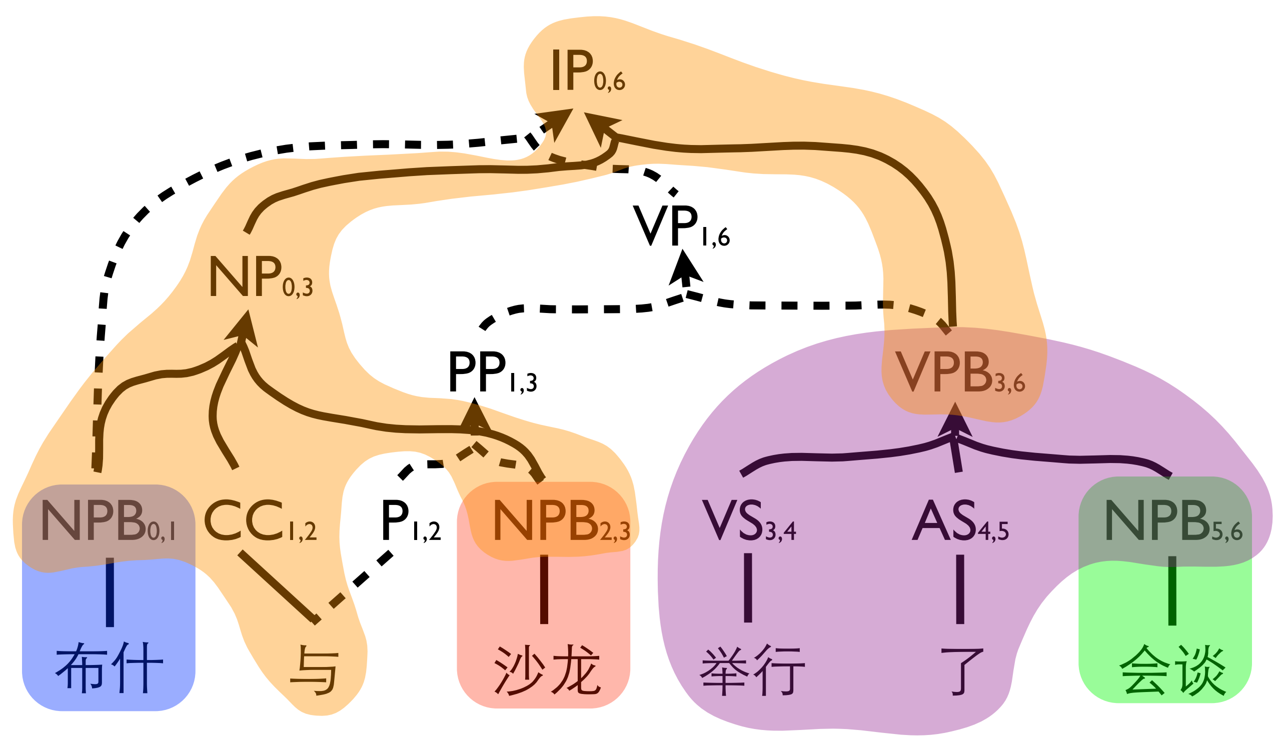


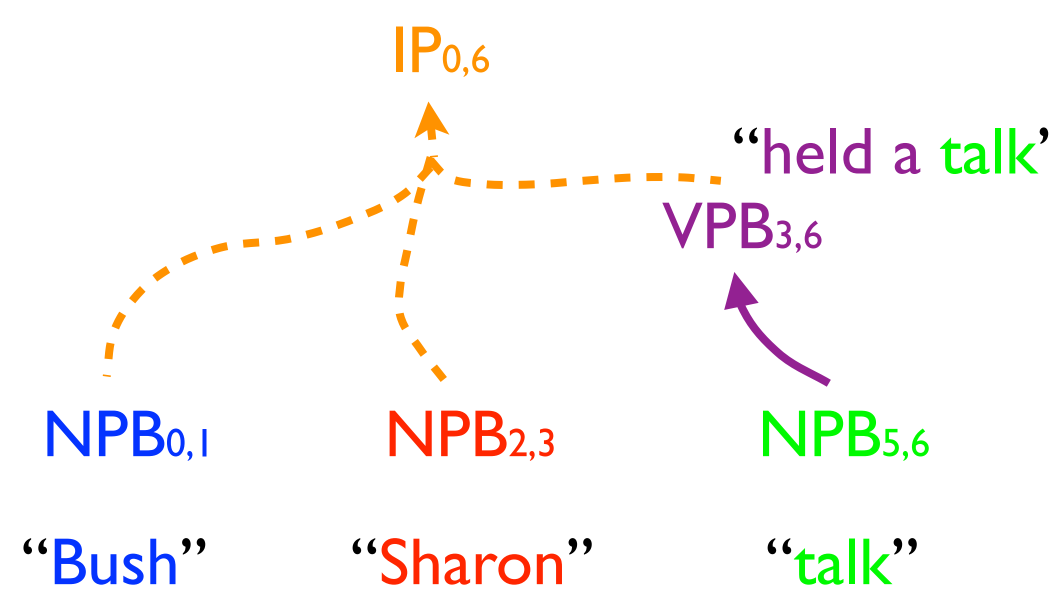


(Mi et al., 2008)

Forest-based Decoding

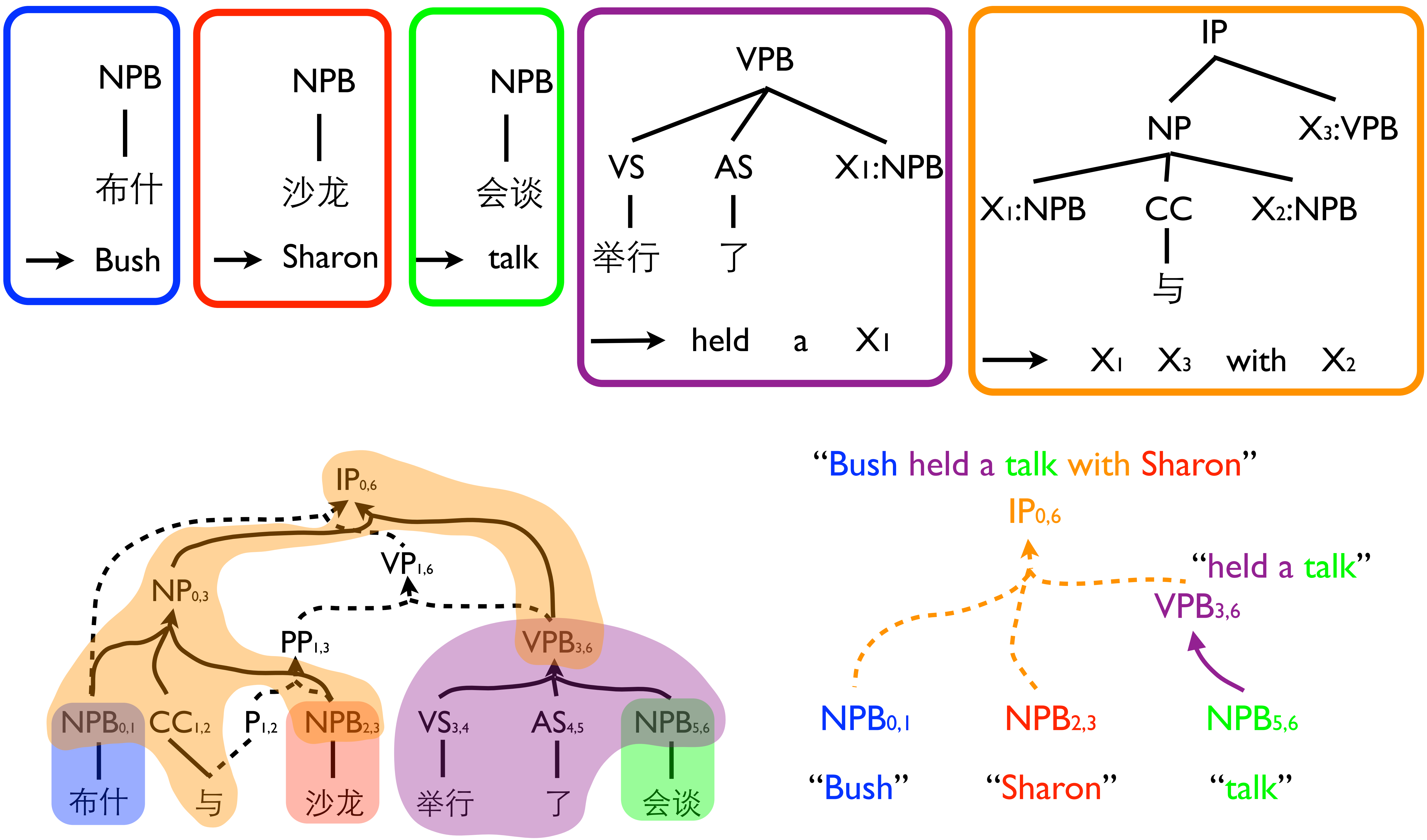






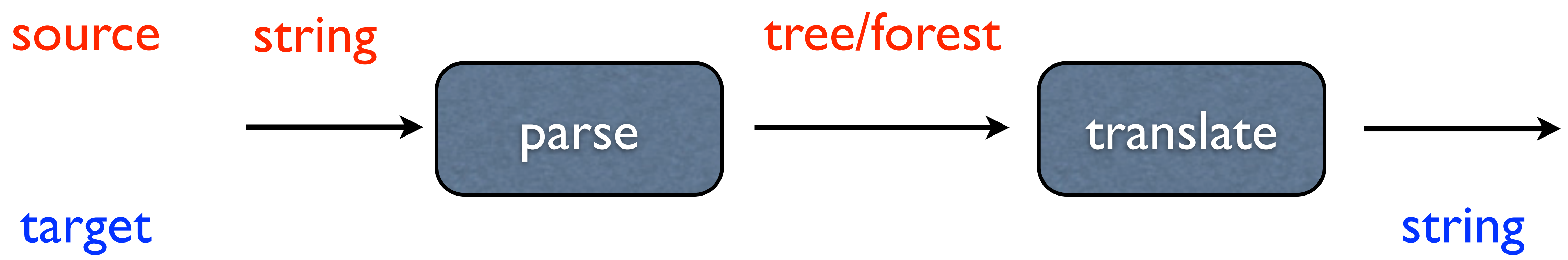
(Mi et al., 2008)

Forest-based Decoding

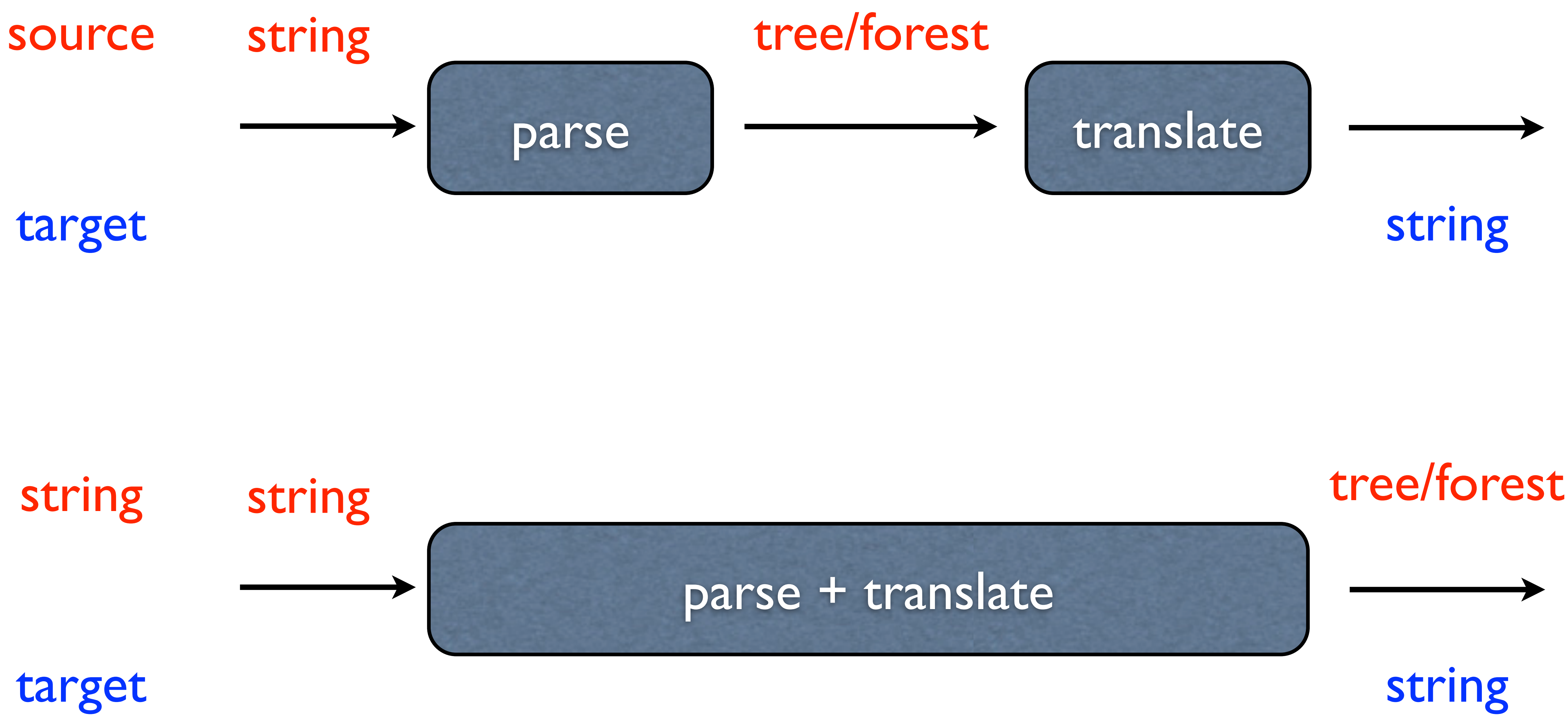


(Mi et al., 2008)

Forest-based Decoding



Forest-based Decoding



Joint Parsing and Translation

布什　与　沙龙　举行　了　会谈

(Liu and Liu, 2010)

Joint Parsing and Translation

NPB

布什

→ Bush

布什 与 沙龙 举行 了 会谈

(Liu and Liu, 2010)

Joint Parsing and Translation

NPB
布什
→ Bush

NPB
布什 与 沙龙 举行 了 会谈

(Liu and Liu, 2010)

Joint Parsing and Translation

NPB
|
布什
⟶ Bush

NPB
|
布什 与 沙龙 举行 了 会谈

Bush

(Liu and Liu, 2010)

Joint Parsing and Translation

NPB

布什 与 沙龙 举行 了 会谈

Bush

(Liu and Liu, 2010)

Joint Parsing and Translation

NPB

沙龙

→ Sharon

NPB

布什 与 沙龙 举行 了 会谈

Bush

(Liu and Liu, 2010)

Joint Parsing and Translation

NPB
沙龙
→ Sharon

NPB　　NPB
布什　与　沙龙　举行　了　会谈

Bush

(Liu and Liu, 2010)

Joint Parsing and Translation

NPB

沙龙

→ Sharon

NPB NPB

布什 与 沙龙 举行 了 会谈

Bush Sharon

(Liu and Liu, 2010)

Joint Parsing and Translation

NPB NPB

布什 与 沙龙 举行 了 会谈

Bush Sharon

(Liu and Liu, 2010)

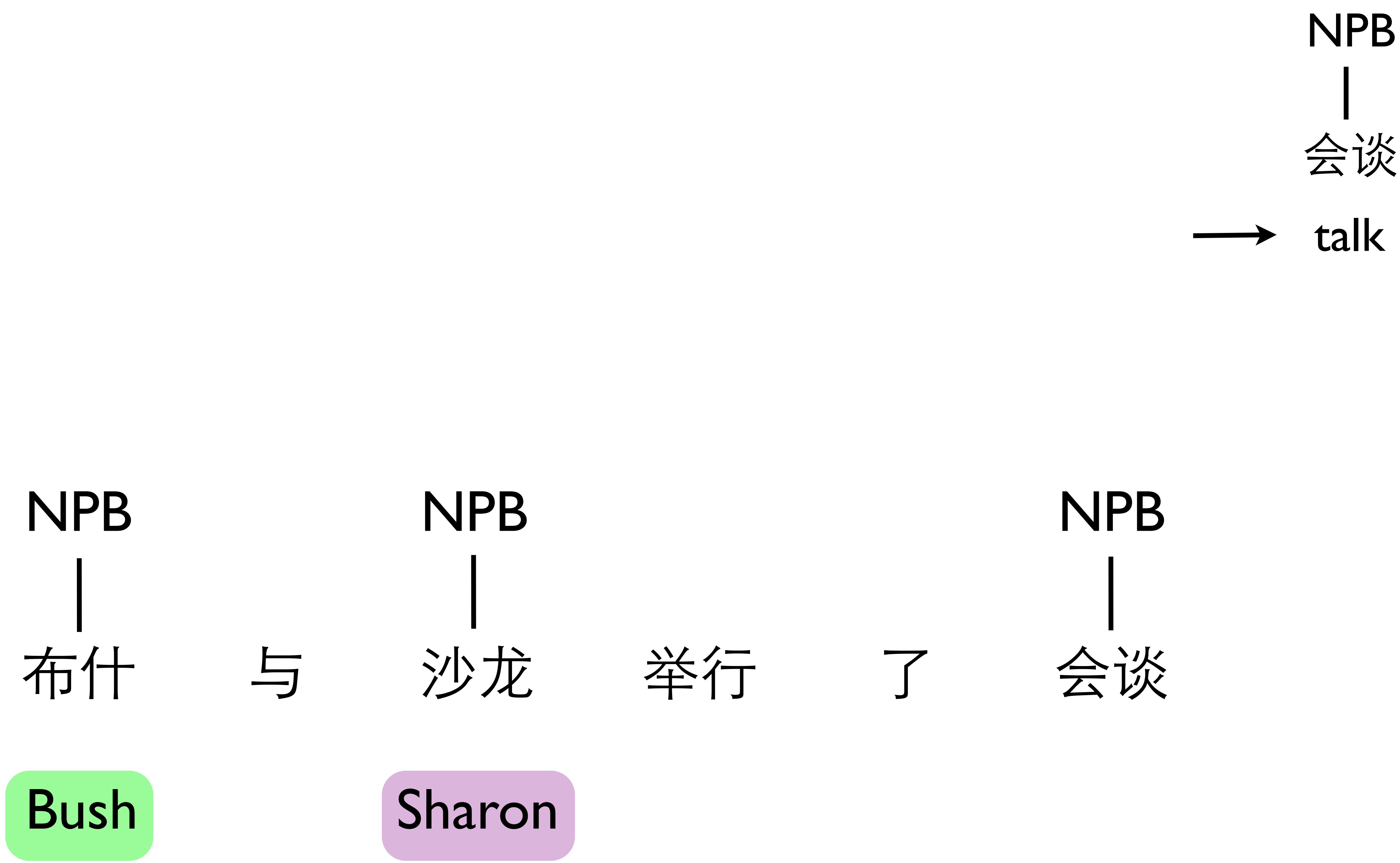
Joint Parsing and Translation

NPB
会谈
→ talk

NPB　　NPB
布什　与　沙龙　举行　了　会谈
Bush　　Sharon

(Liu and Liu, 2010)

Joint Parsing and Translation



(Liu and Liu, 2010)

Joint Parsing and Translation

NPB
会谈
→ talk

NPB		NPB			NPB
布什	与	沙龙	举行	了	会谈
Bush		Sharon			talk

(Liu and Liu, 2010)

Joint Parsing and Translation

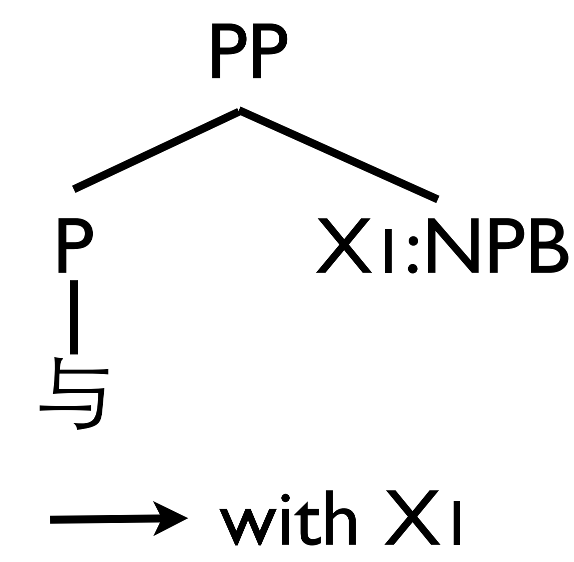
NPB NPB NPB

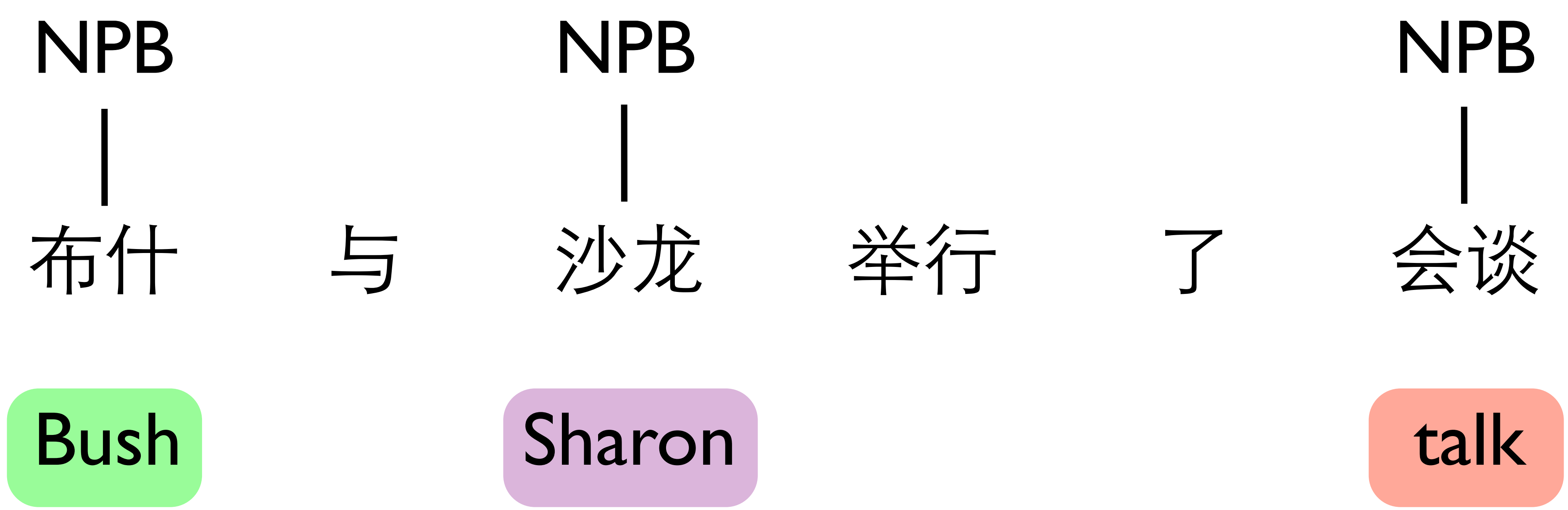
布什 与 沙龙 举行 了 会谈

Bush Sharon talk

(Liu and Liu, 2010)

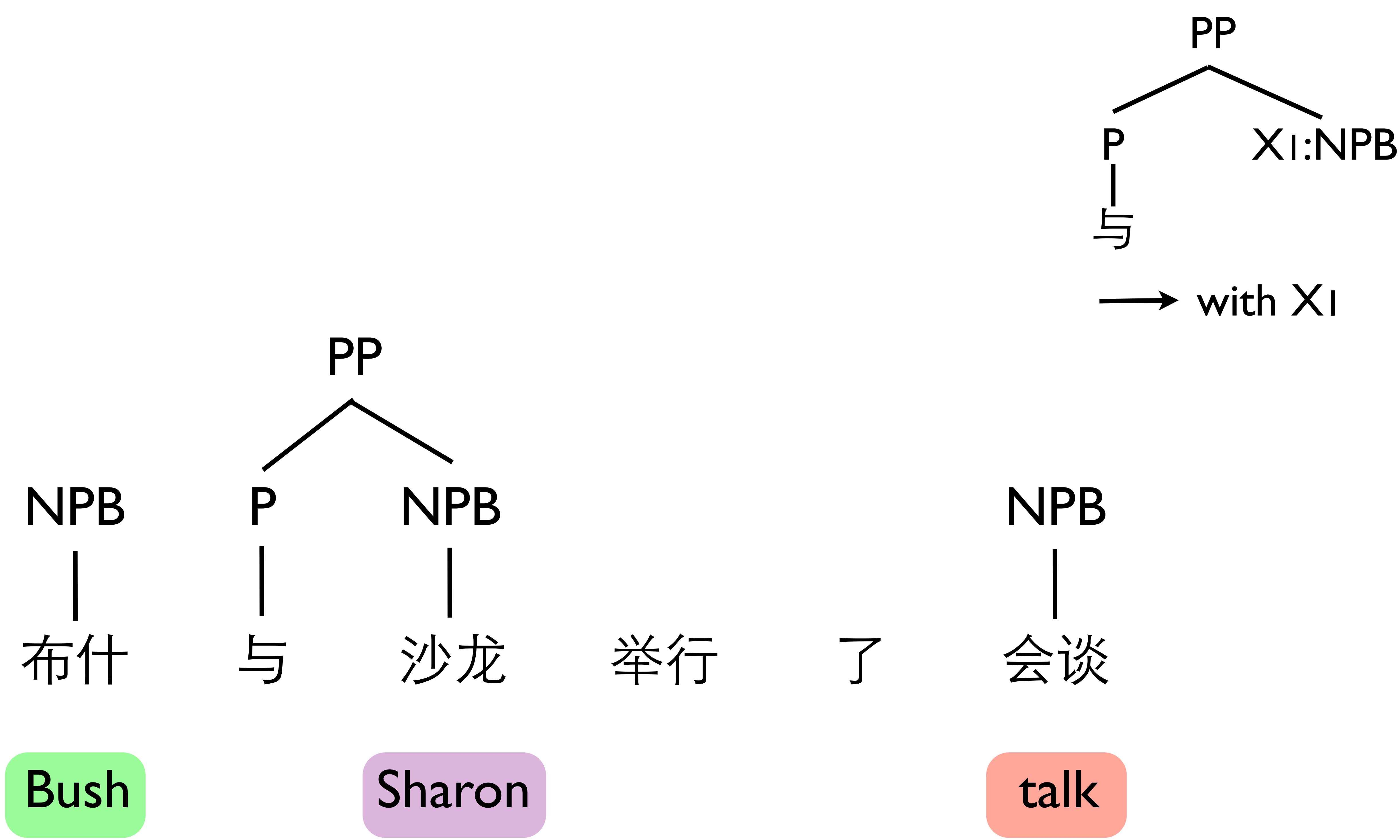
Joint Parsing and Translation





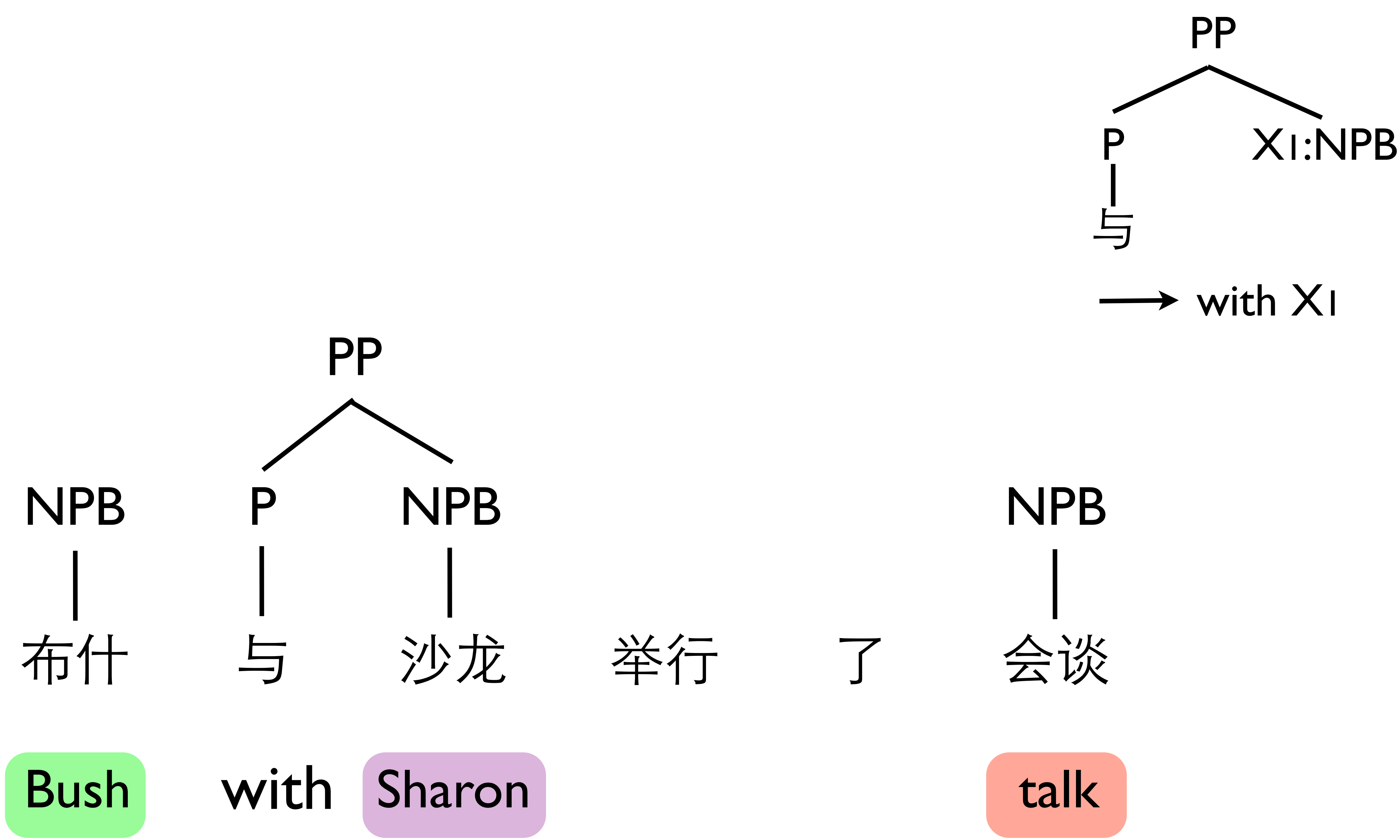
(Liu and Liu, 2010)

Joint Parsing and Translation



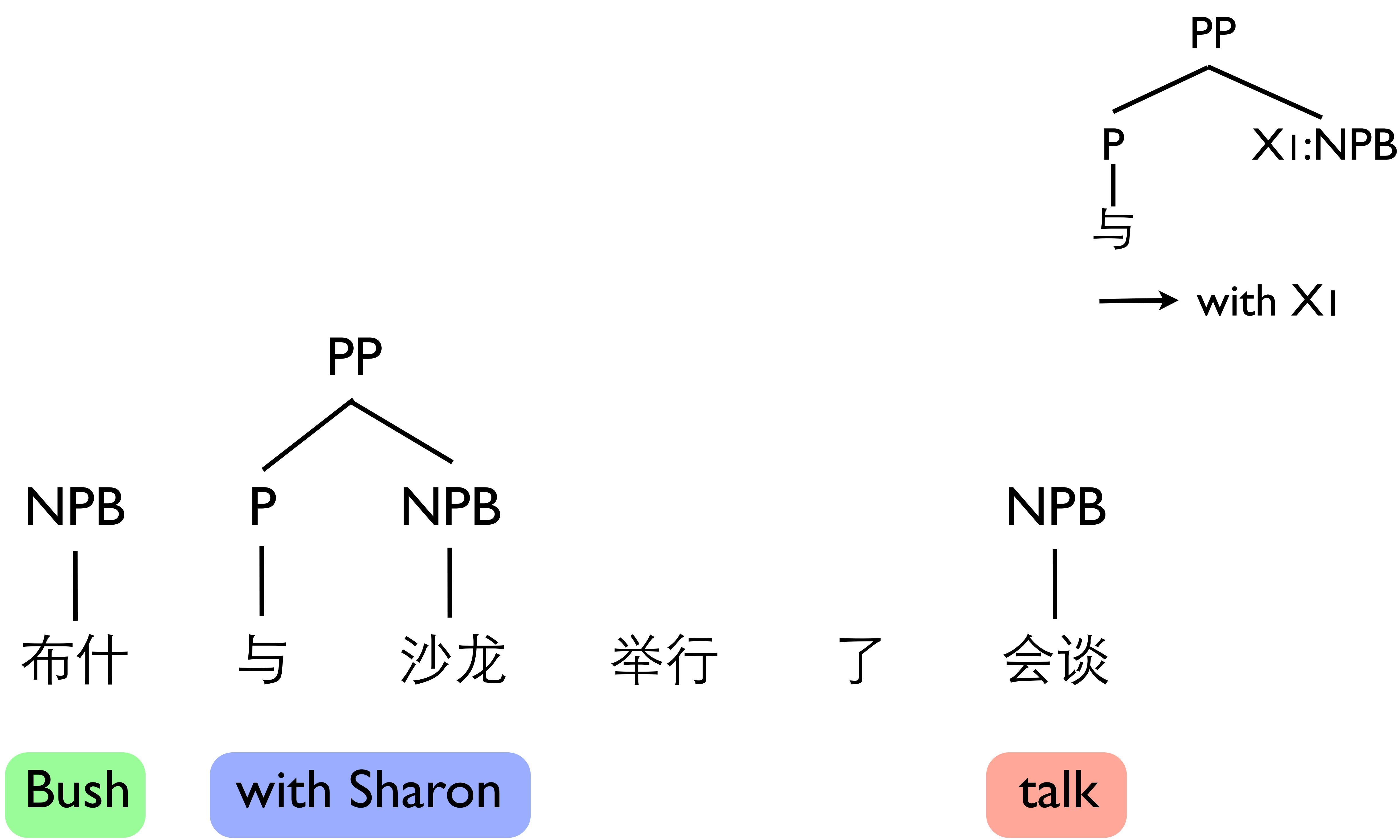
(Liu and Liu, 2010)

Joint Parsing and Translation



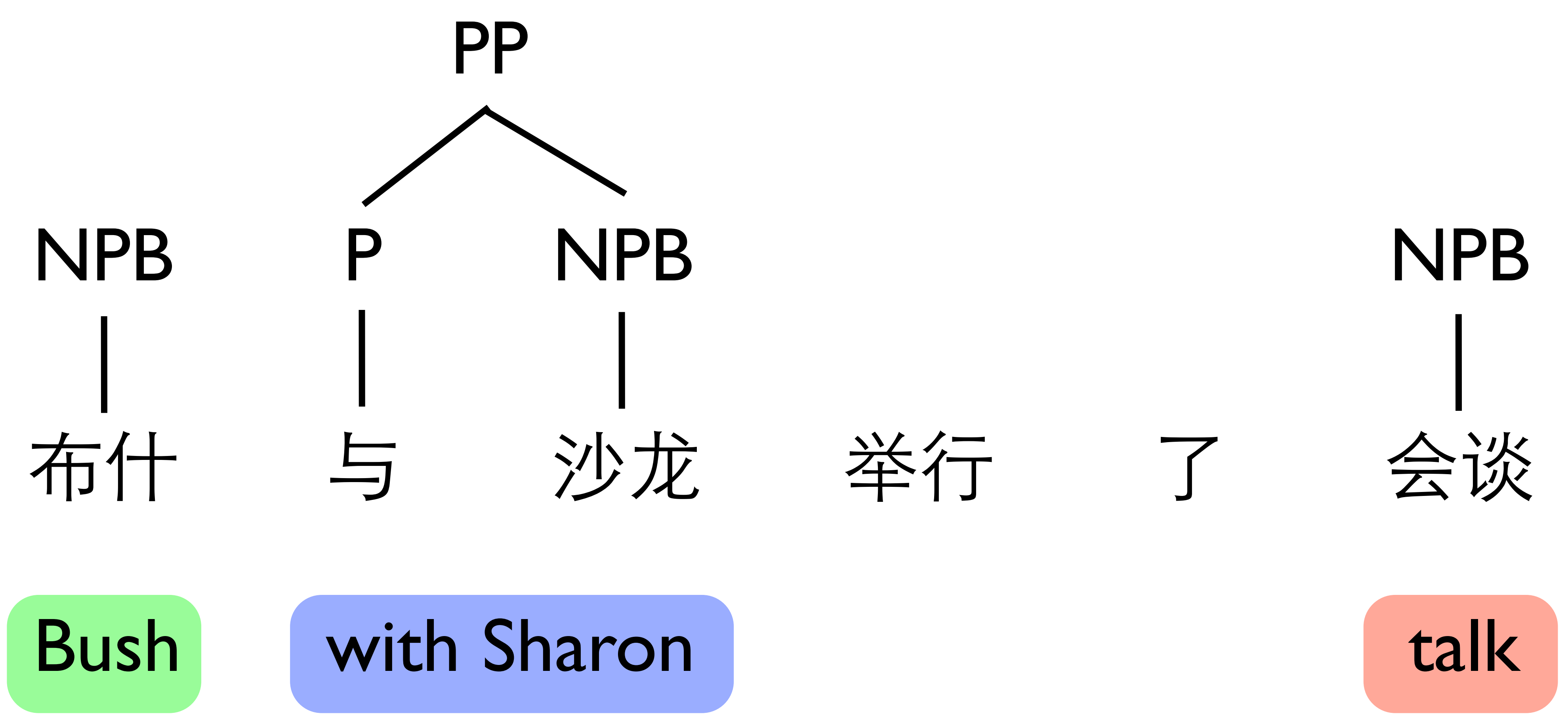
(Liu and Liu, 2010)

Joint Parsing and Translation



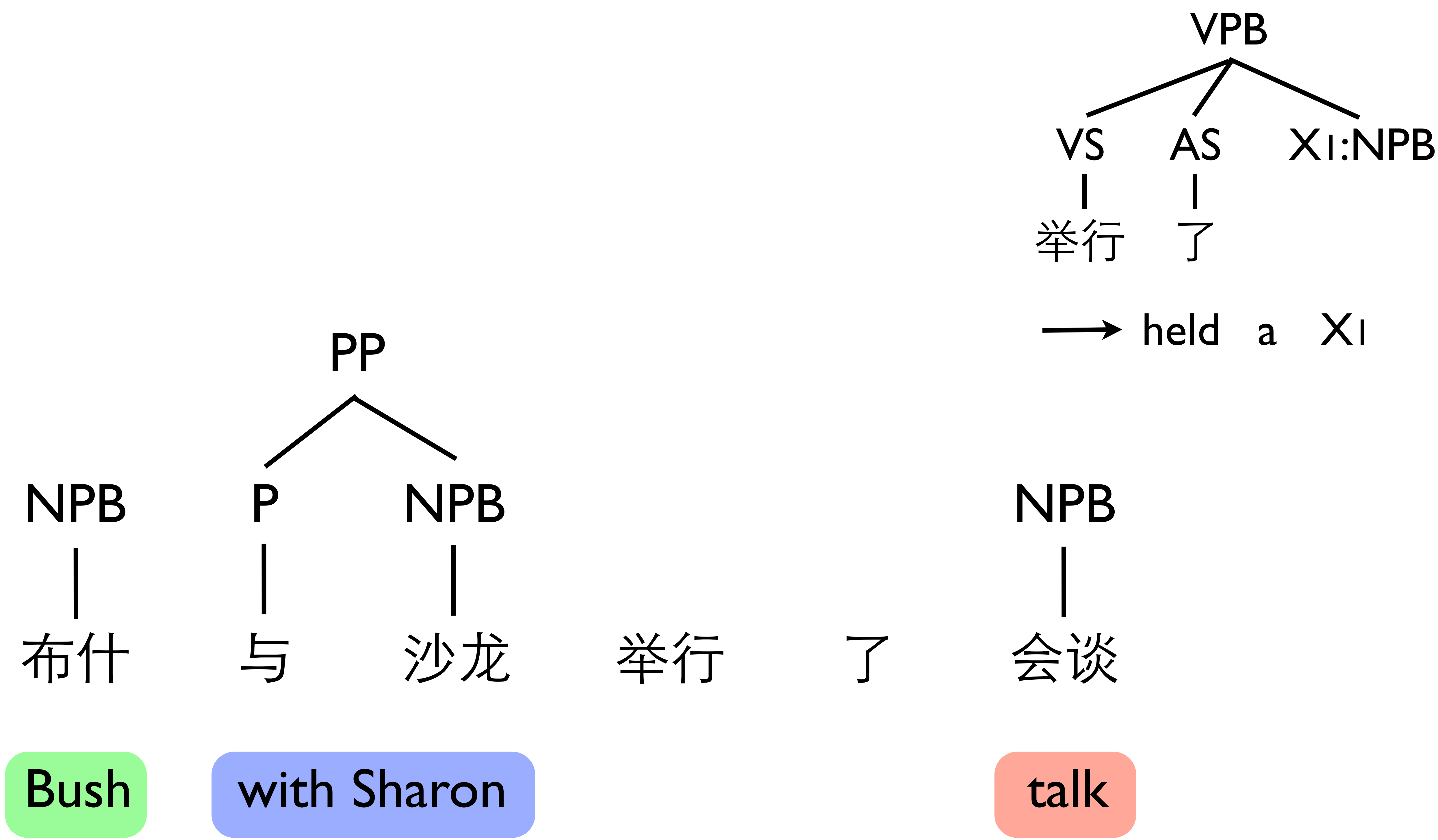
(Liu and Liu, 2010)

Joint Parsing and Translation



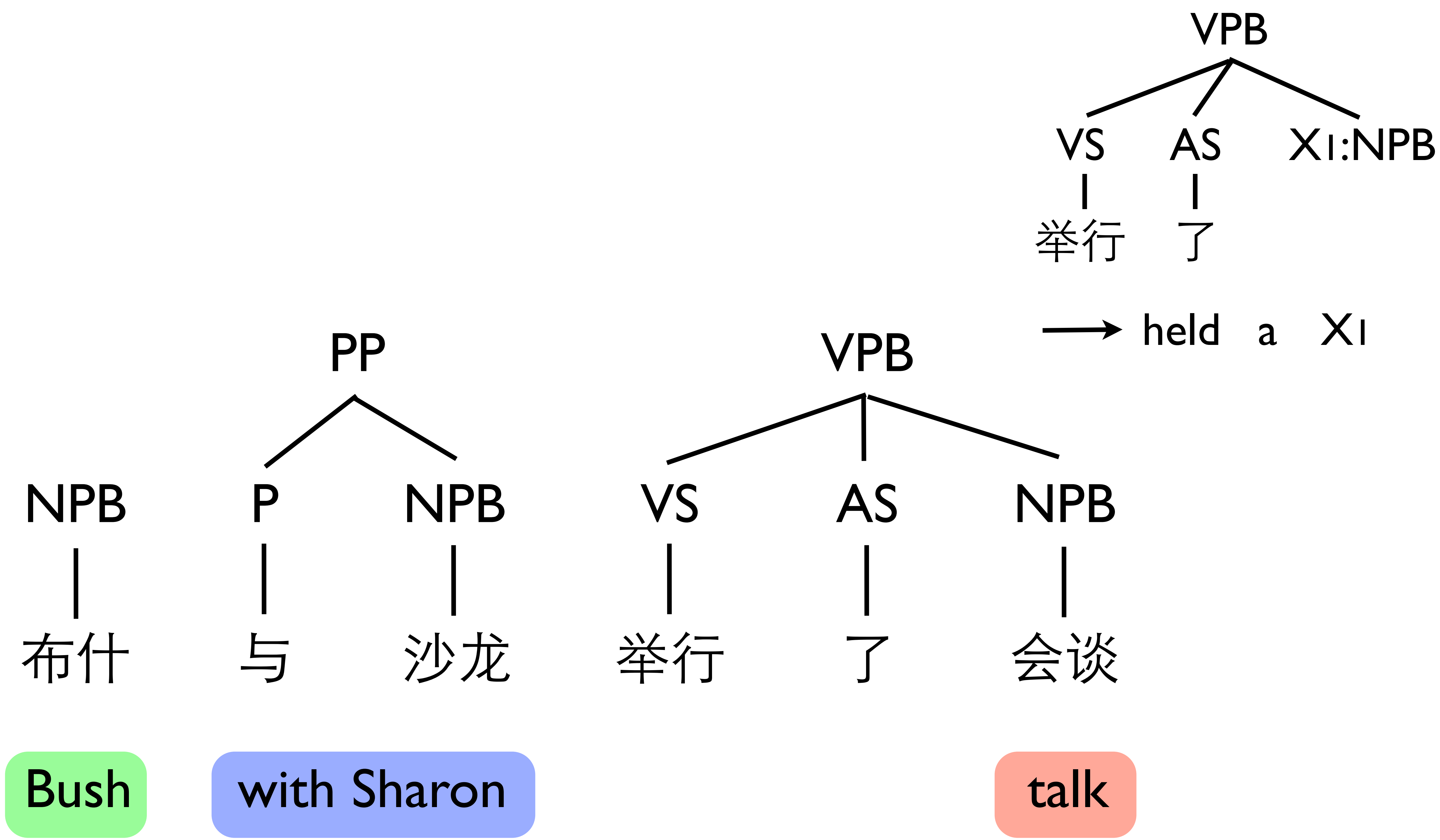
(Liu and Liu, 2010)

Joint Parsing and Translation



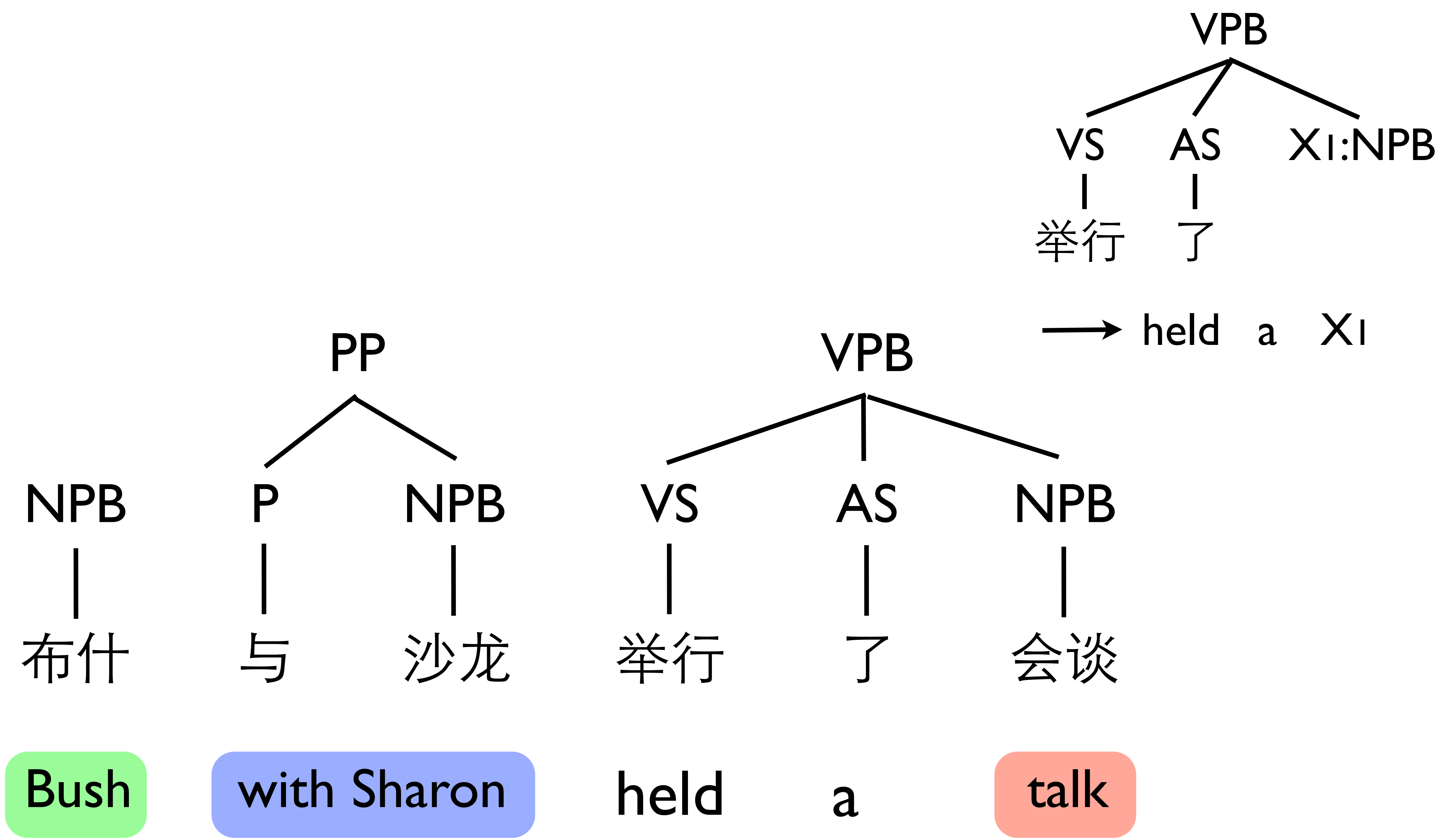
(Liu and Liu, 2010)

Joint Parsing and Translation



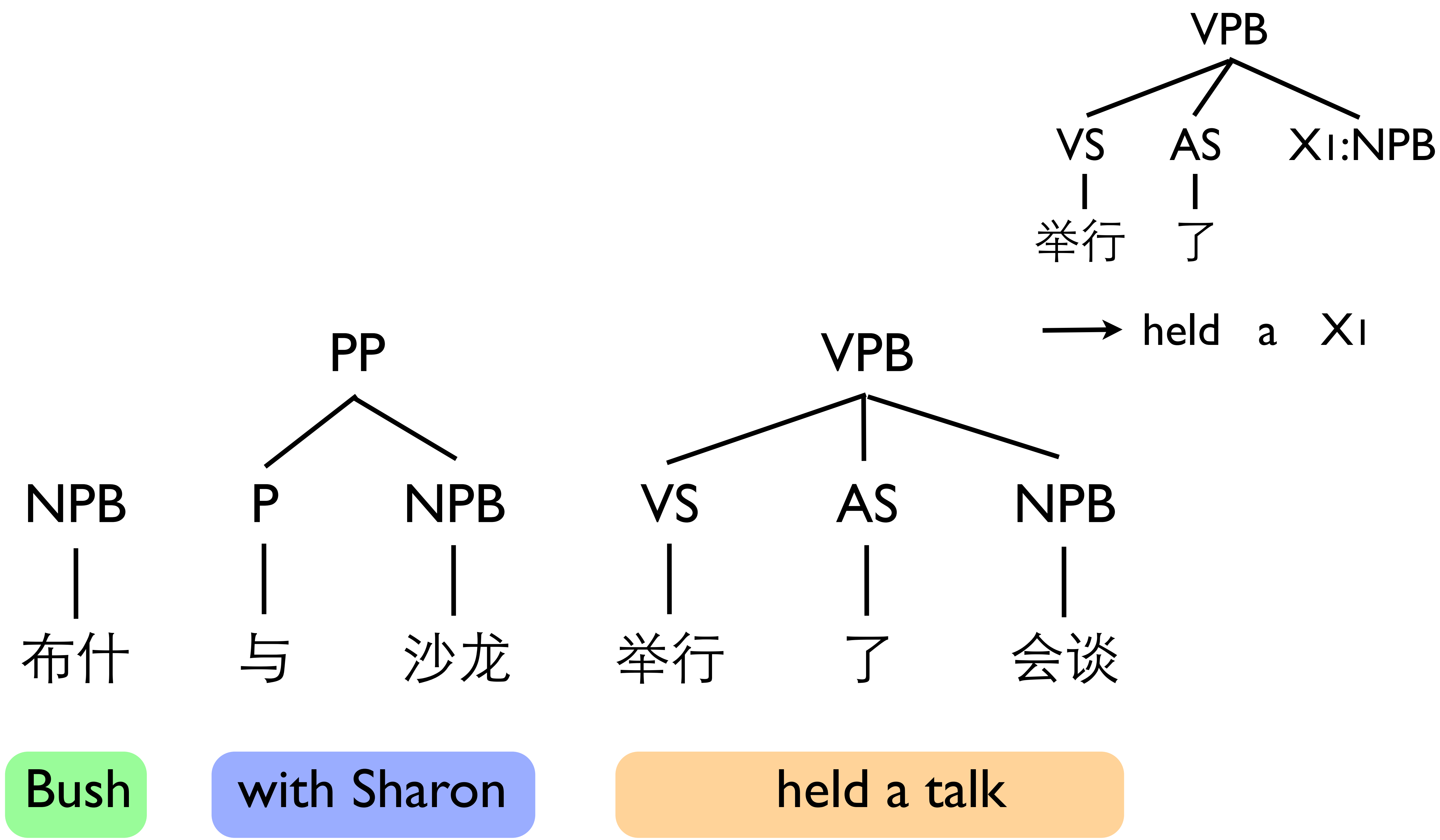
(Liu and Liu, 2010)

Joint Parsing and Translation



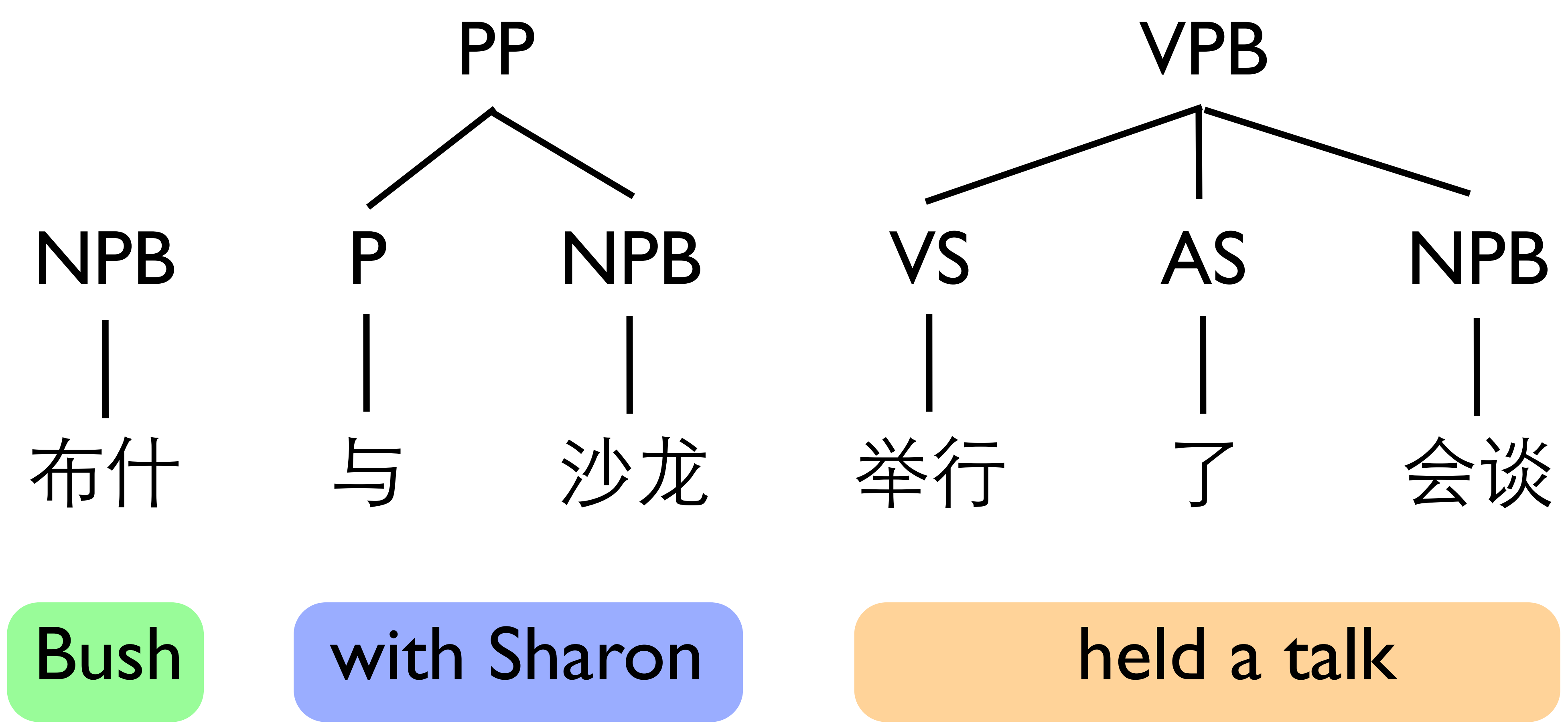
(Liu and Liu, 2010)

Joint Parsing and Translation



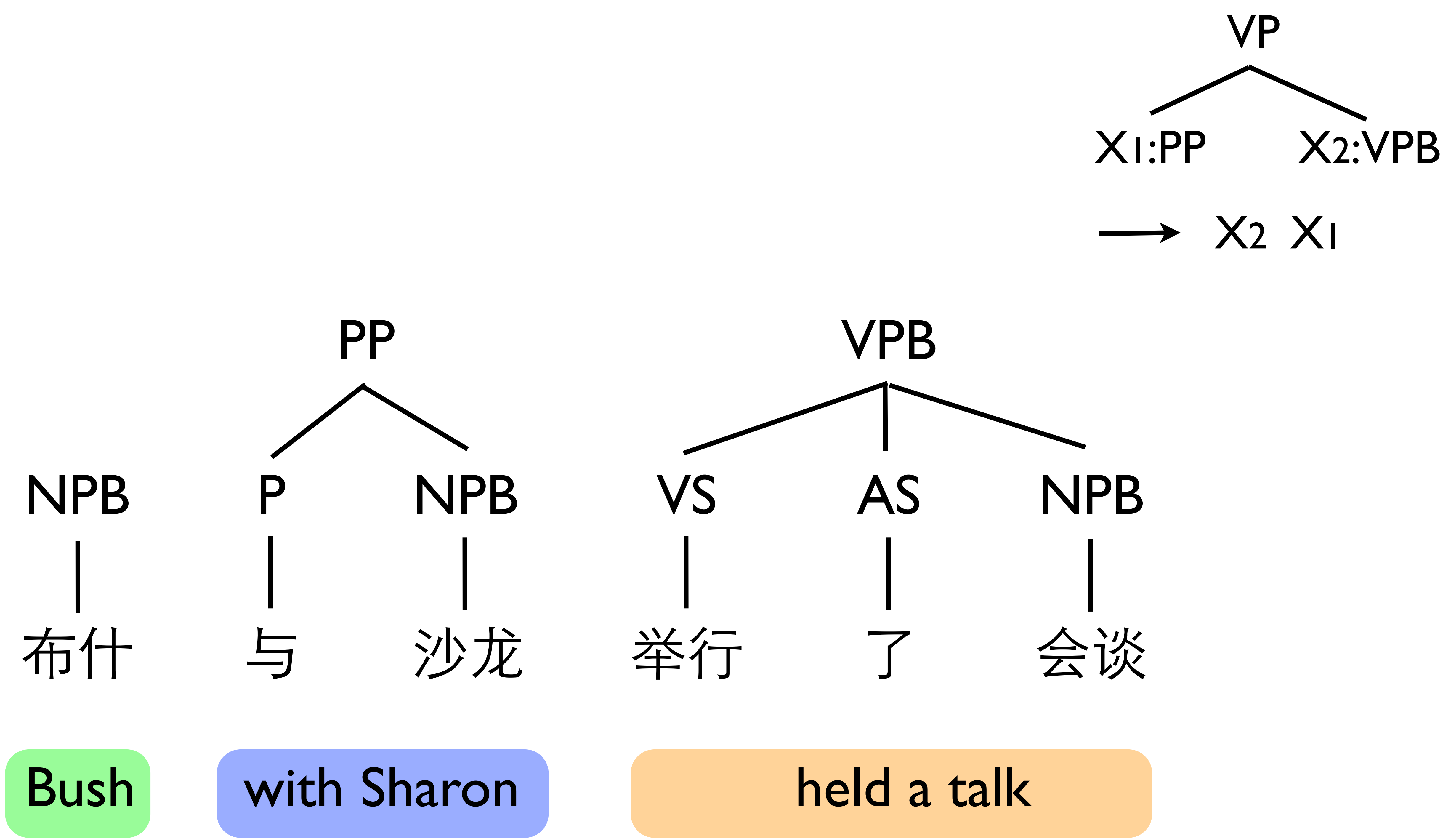
(Liu and Liu, 2010)

Joint Parsing and Translation



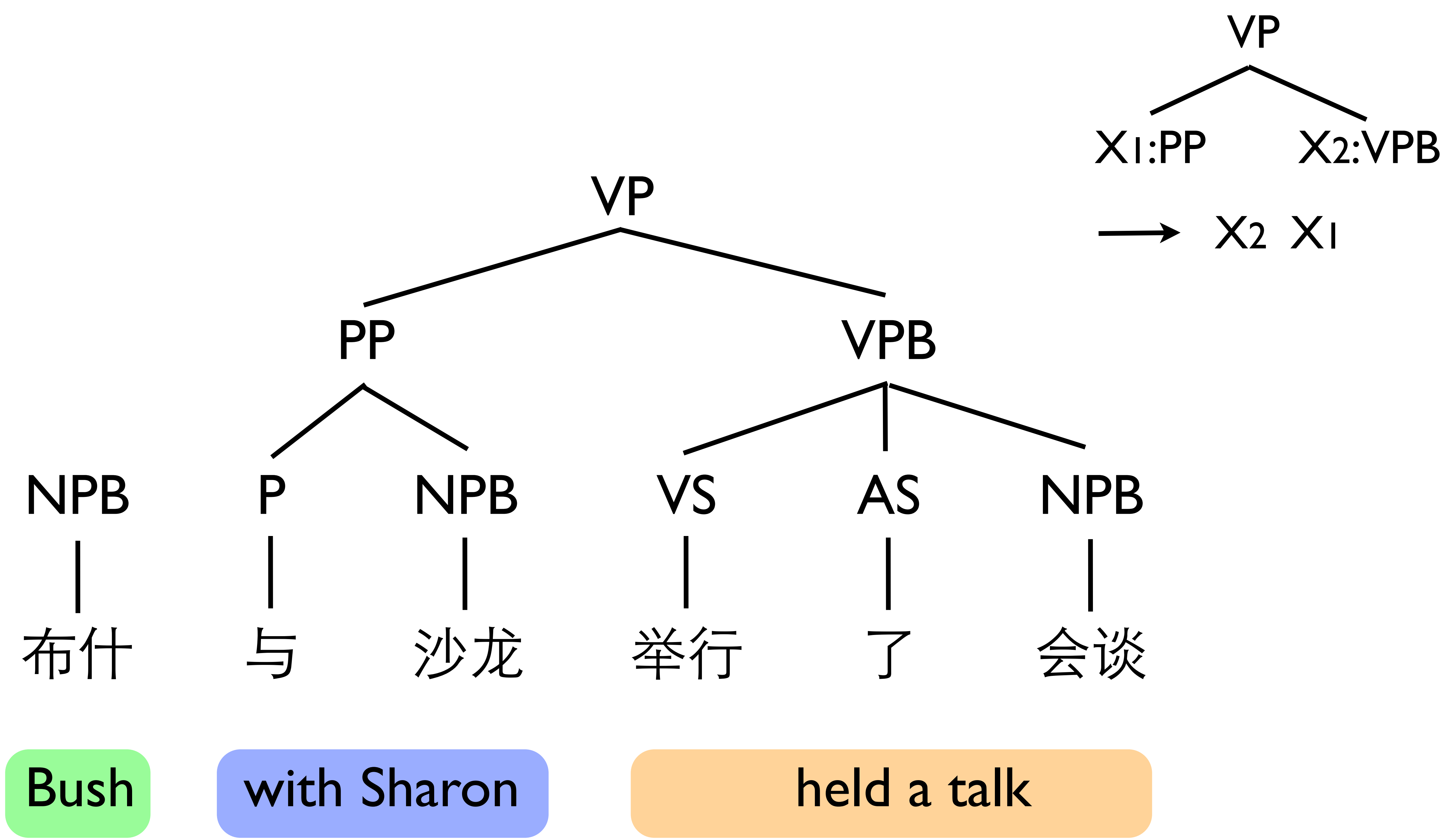
(Liu and Liu, 2010)

Joint Parsing and Translation



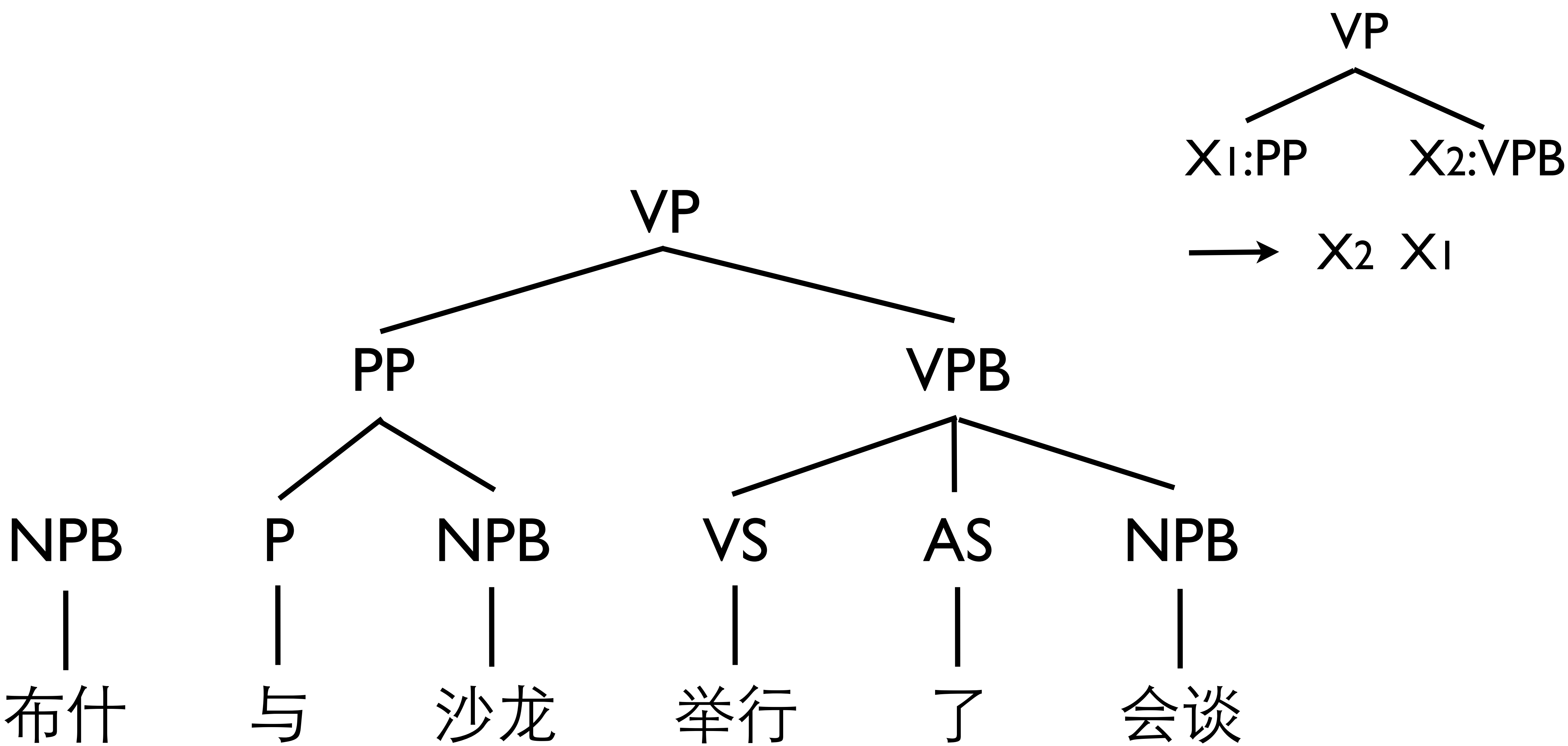
(Liu and Liu, 2010)

Joint Parsing and Translation



(Liu and Liu, 2010)

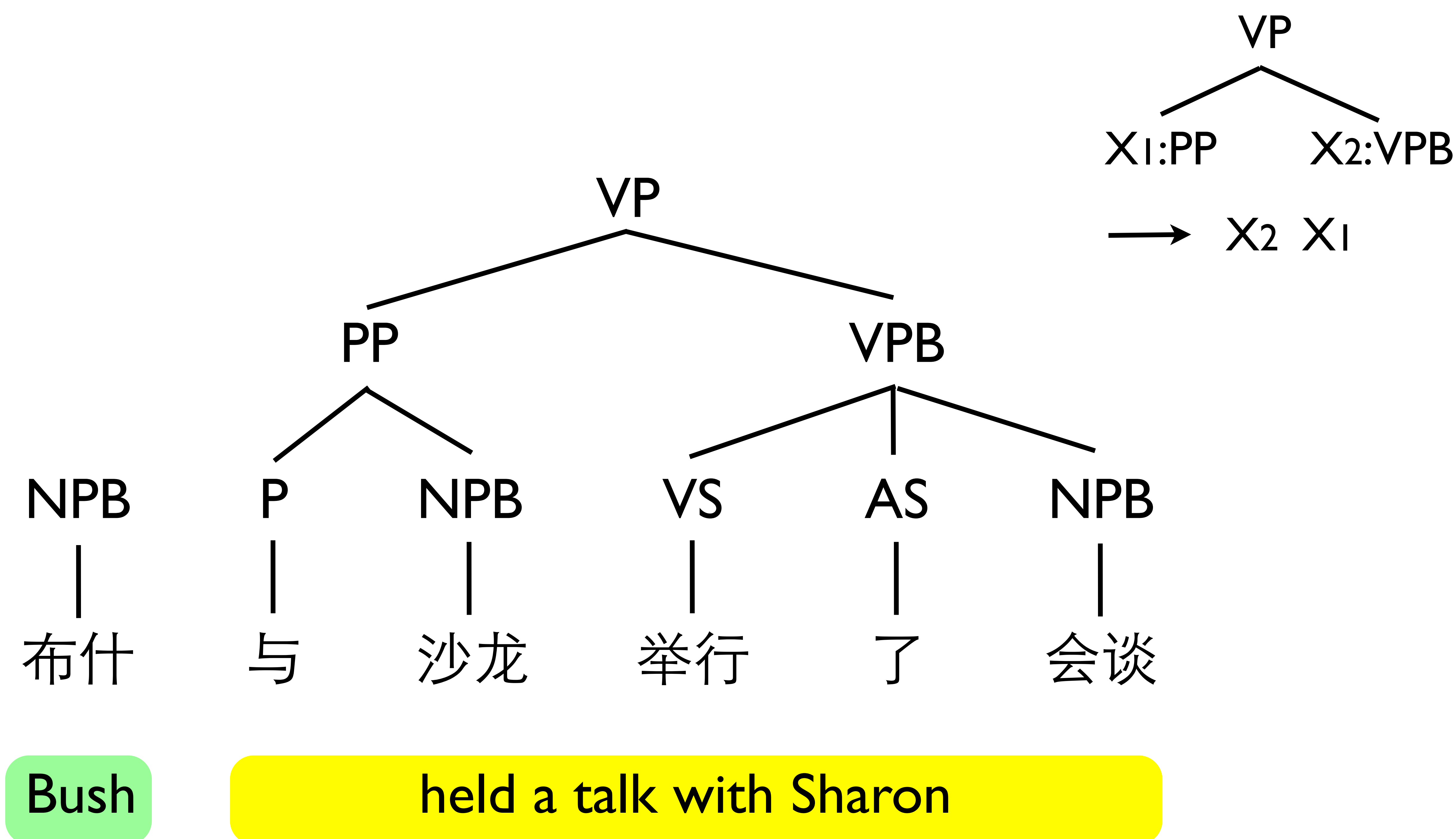
Joint Parsing and Translation



Bush

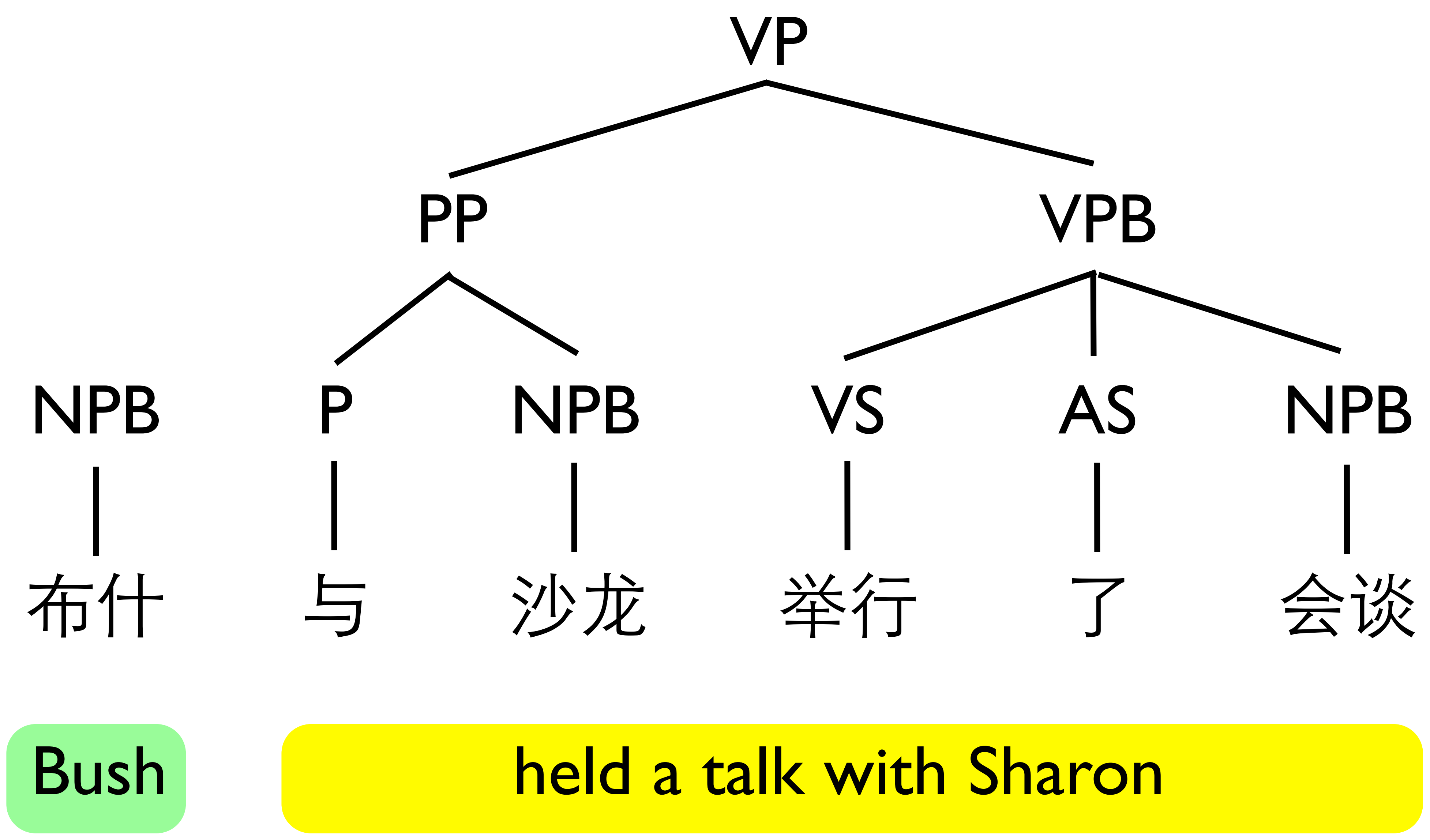
(Liu and Liu, 2010)

Joint Parsing and Translation



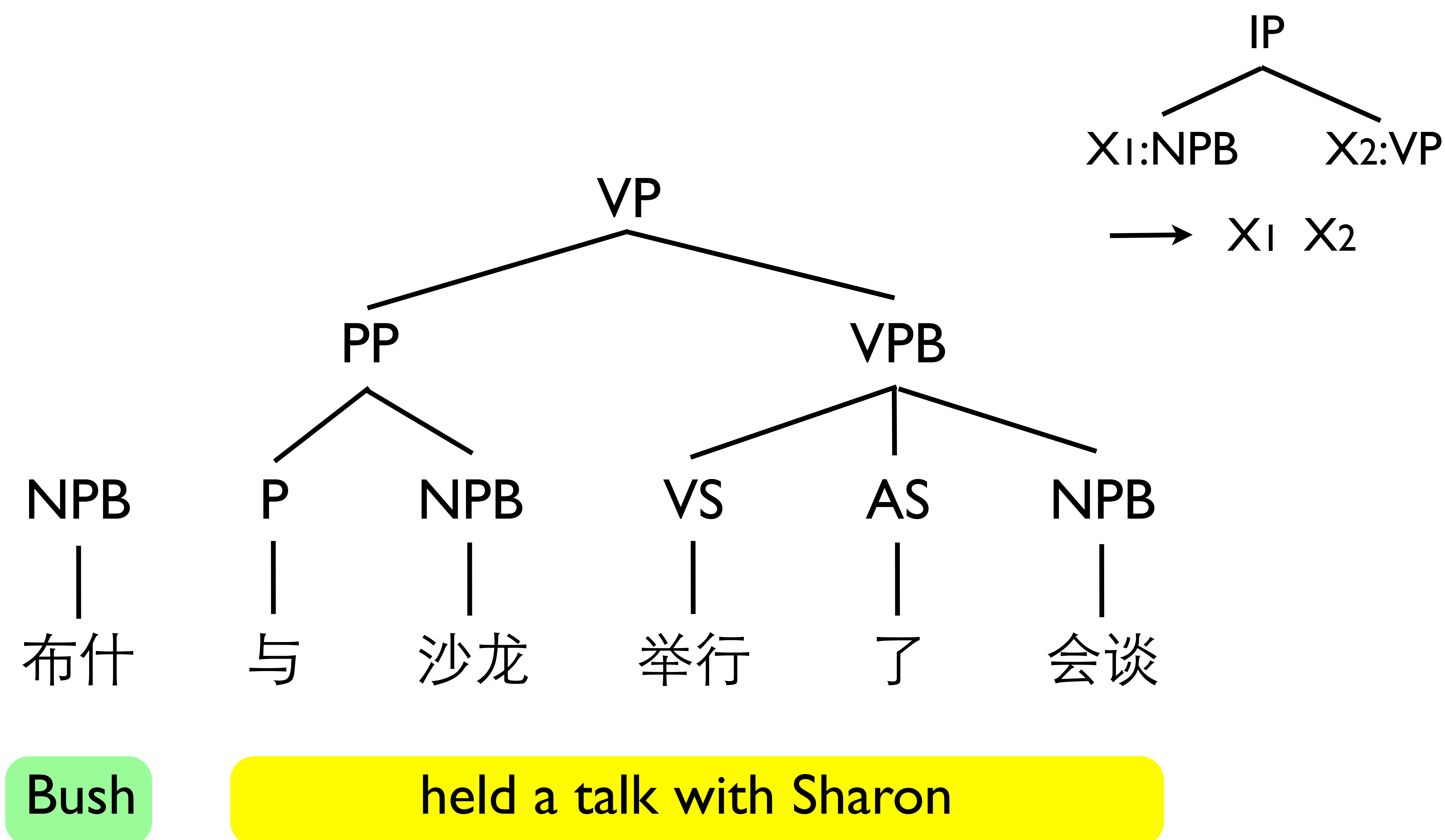
(Liu and Liu, 2010)

Joint Parsing and Translation



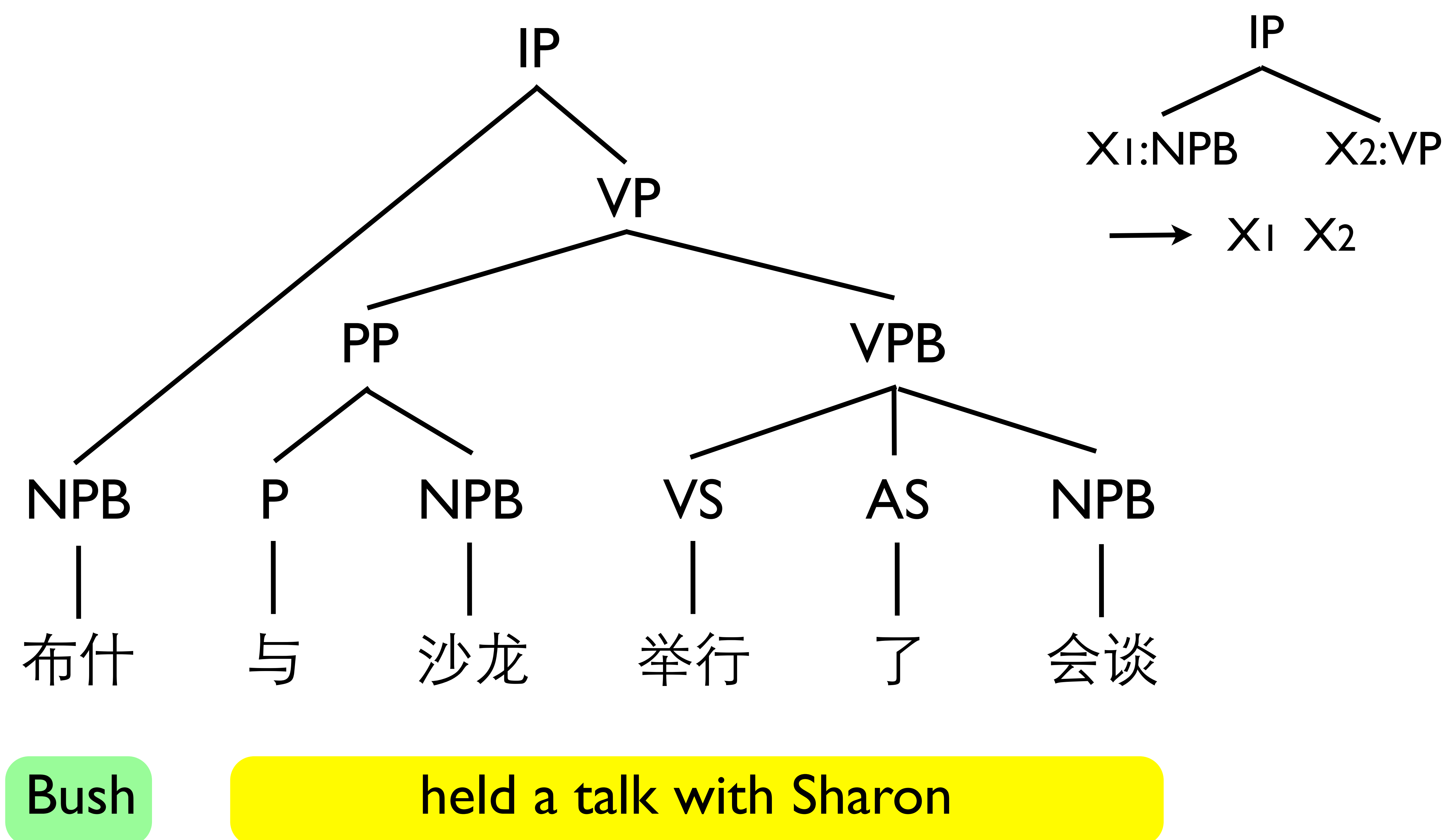
(Liu and Liu, 2010)

Joint Parsing and Translation



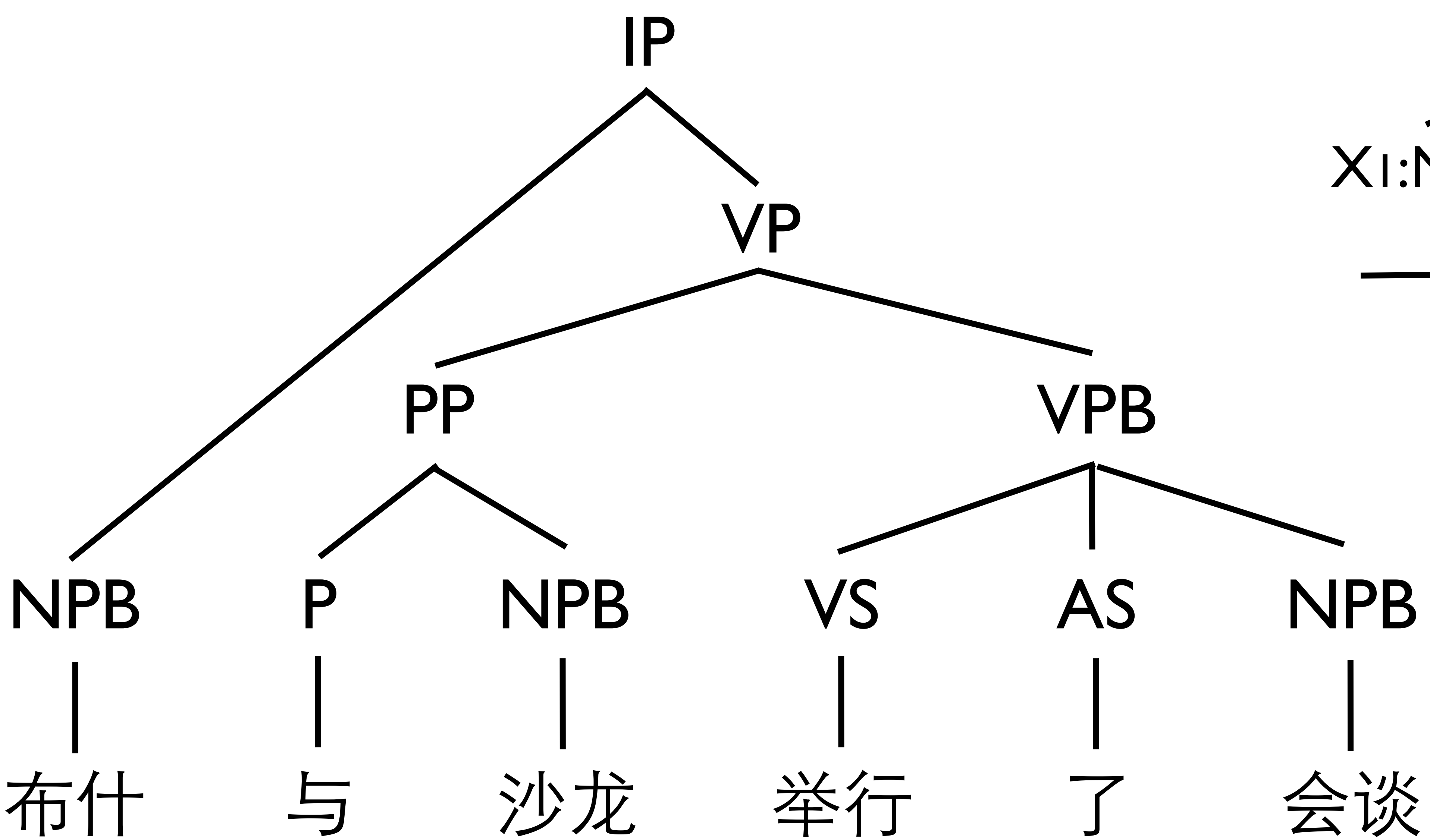
(Liu and Liu, 2010)

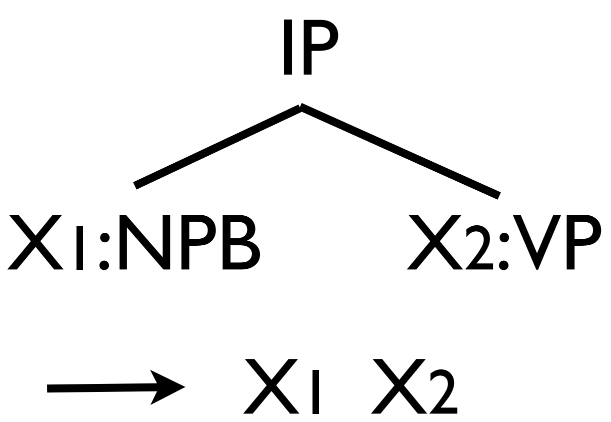
Joint Parsing and Translation



(Liu and Liu, 2010)

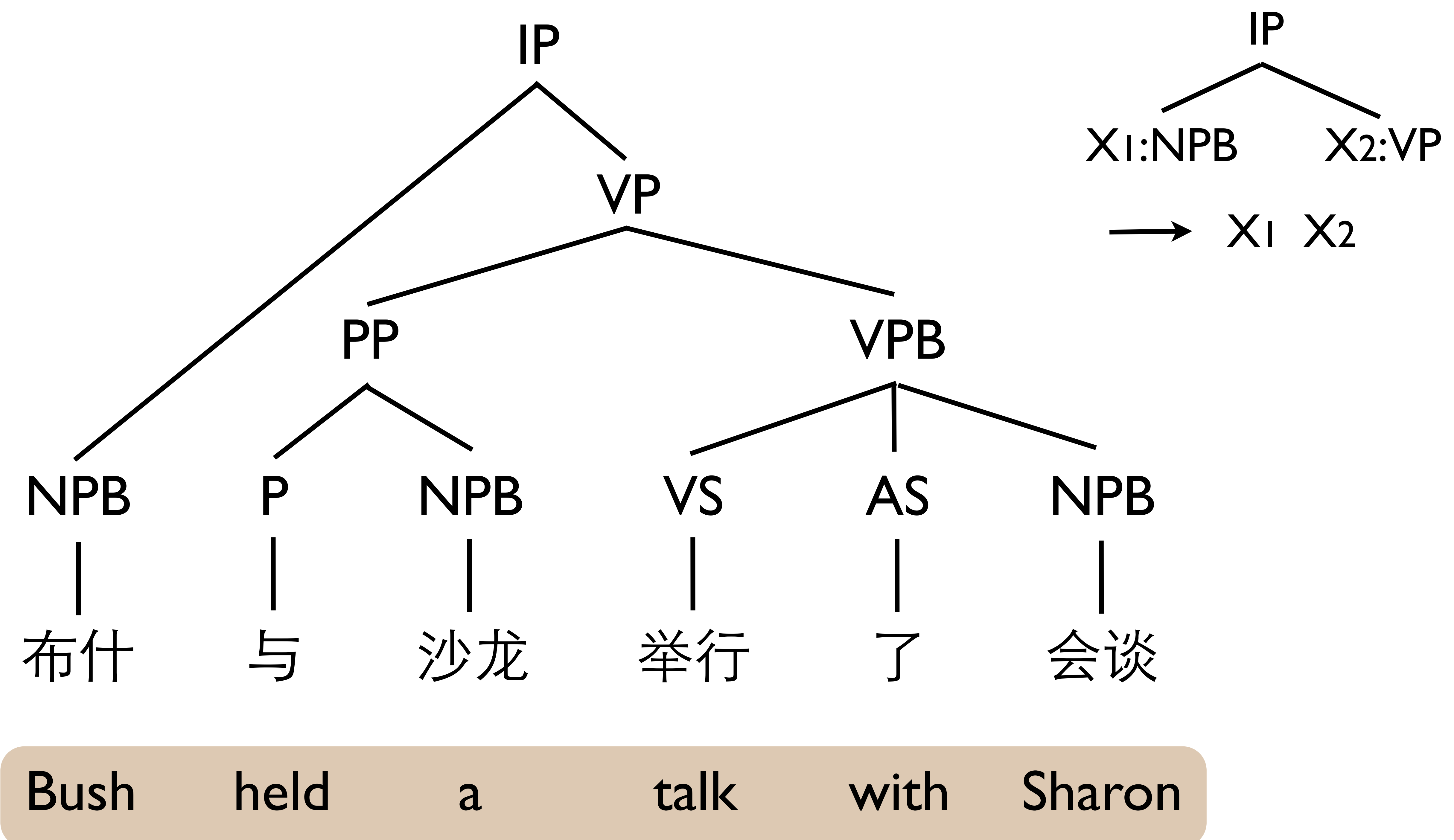
Joint Parsing and Translation





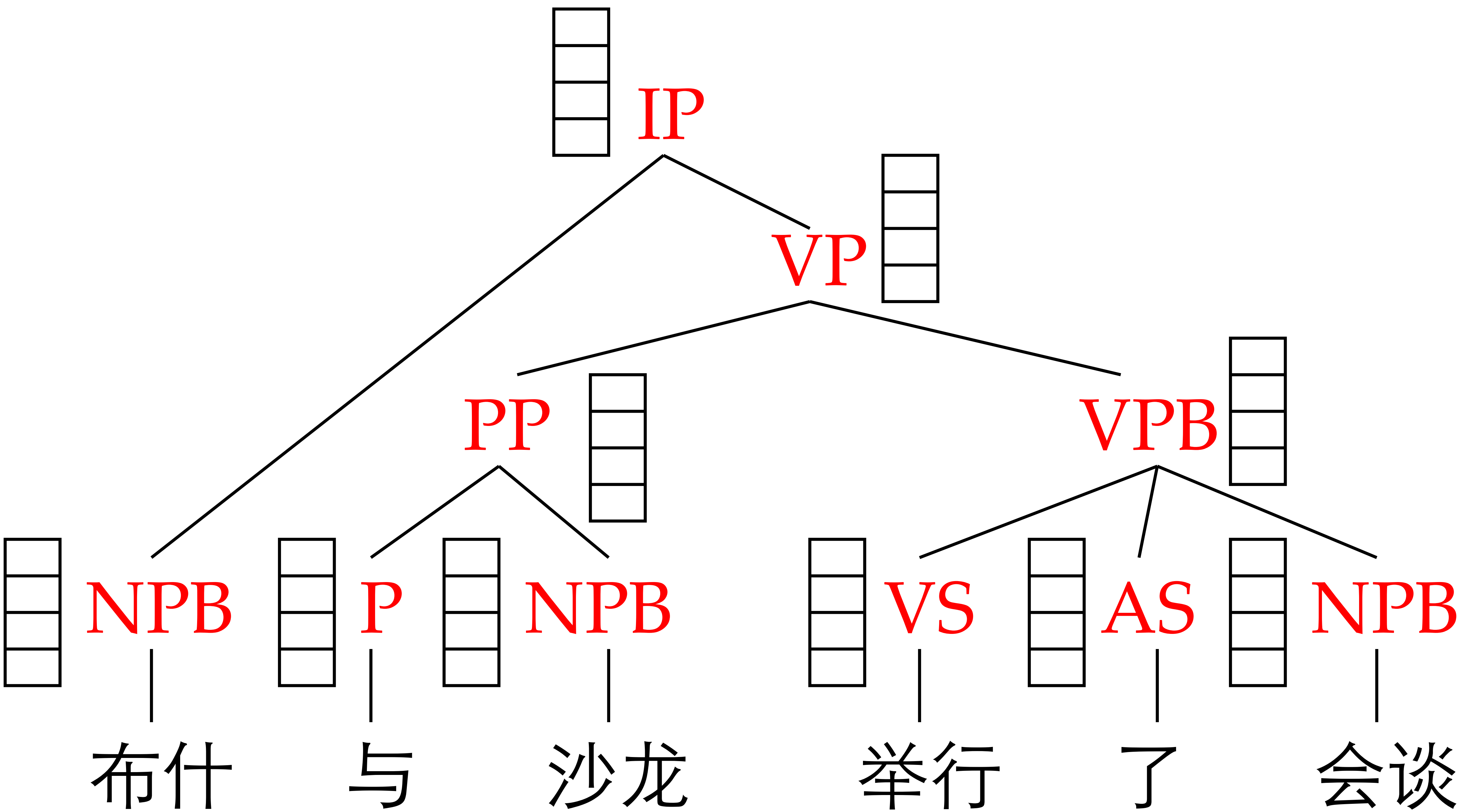
(Liu and Liu, 2010)

Joint Parsing and Translation



(Liu and Liu, 2010)

Stacks in Tree-to-String Translation

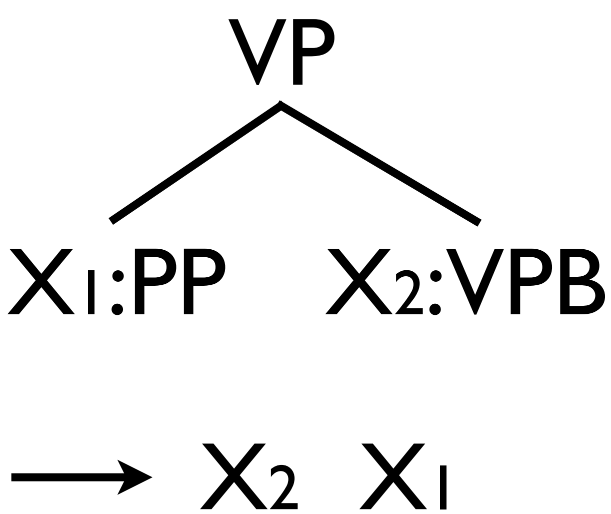


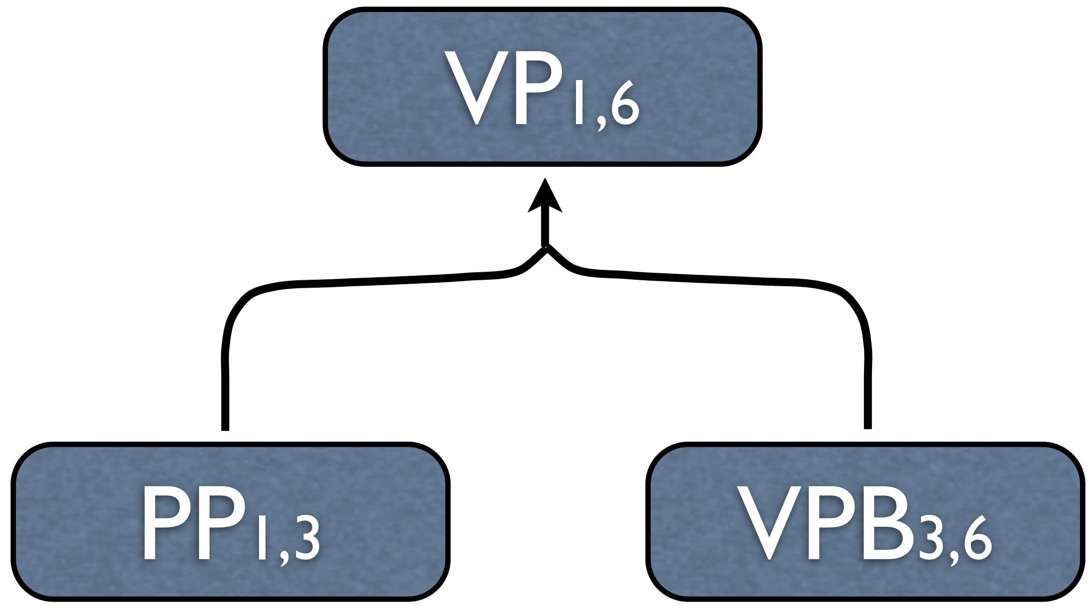
Exhaustive Search

Exhaustive Search

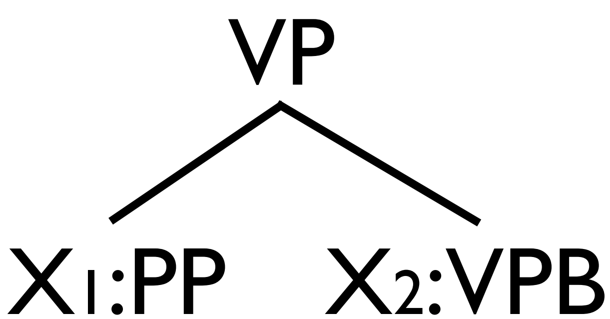
$$\mathrm{VP}(x_1{:}\mathrm{PP}\ \ x_2{:}\mathrm{VPB}) \longrightarrow x_2\ \ x_1$$

Exhaustive Search

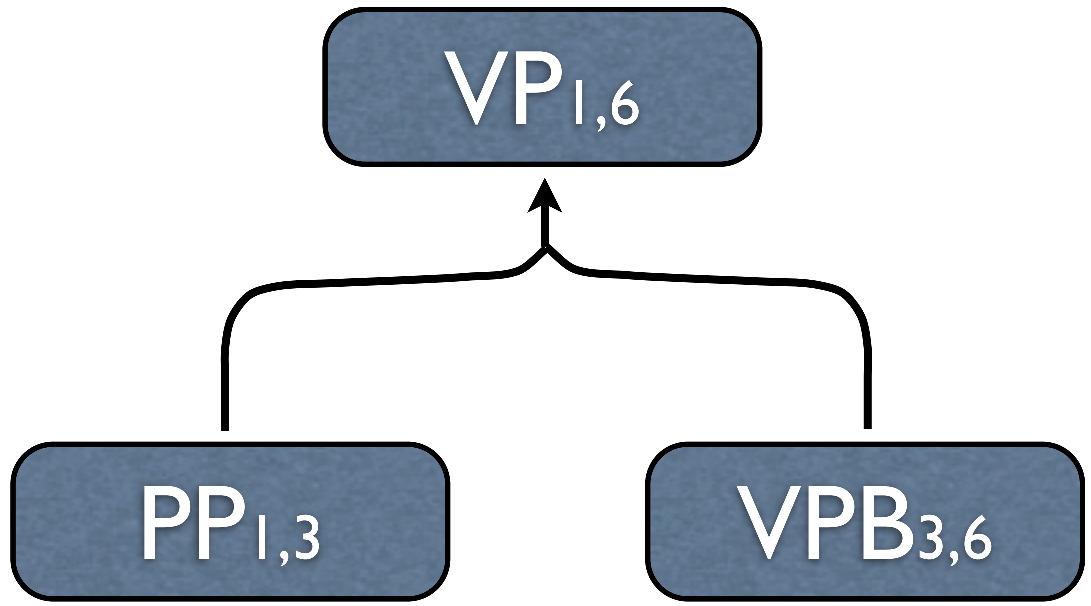




Exhaustive Search



$\longrightarrow X_2 \; X_1$



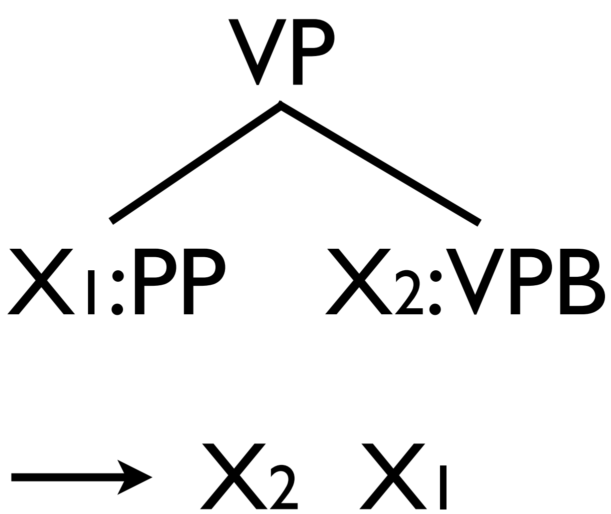
with Sharon

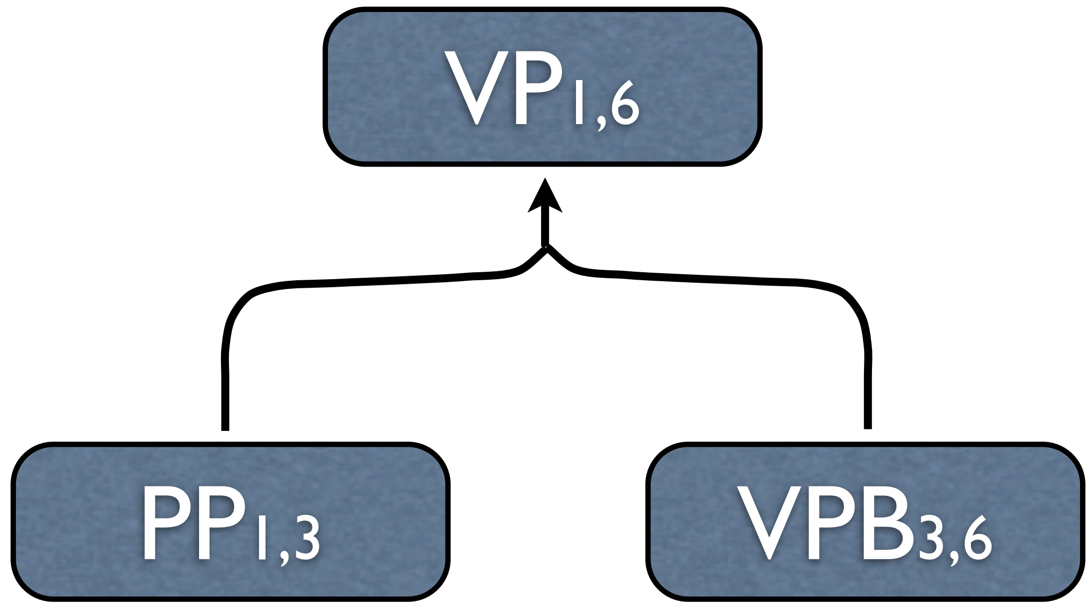
and Sharon

Sharon with

Sharon and

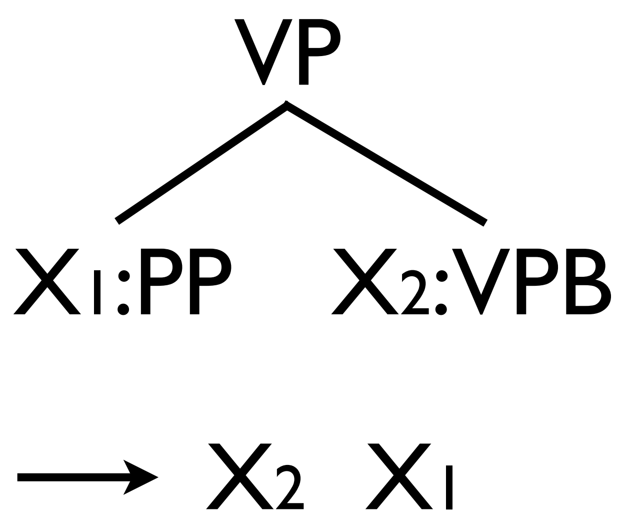
Exhaustive Search

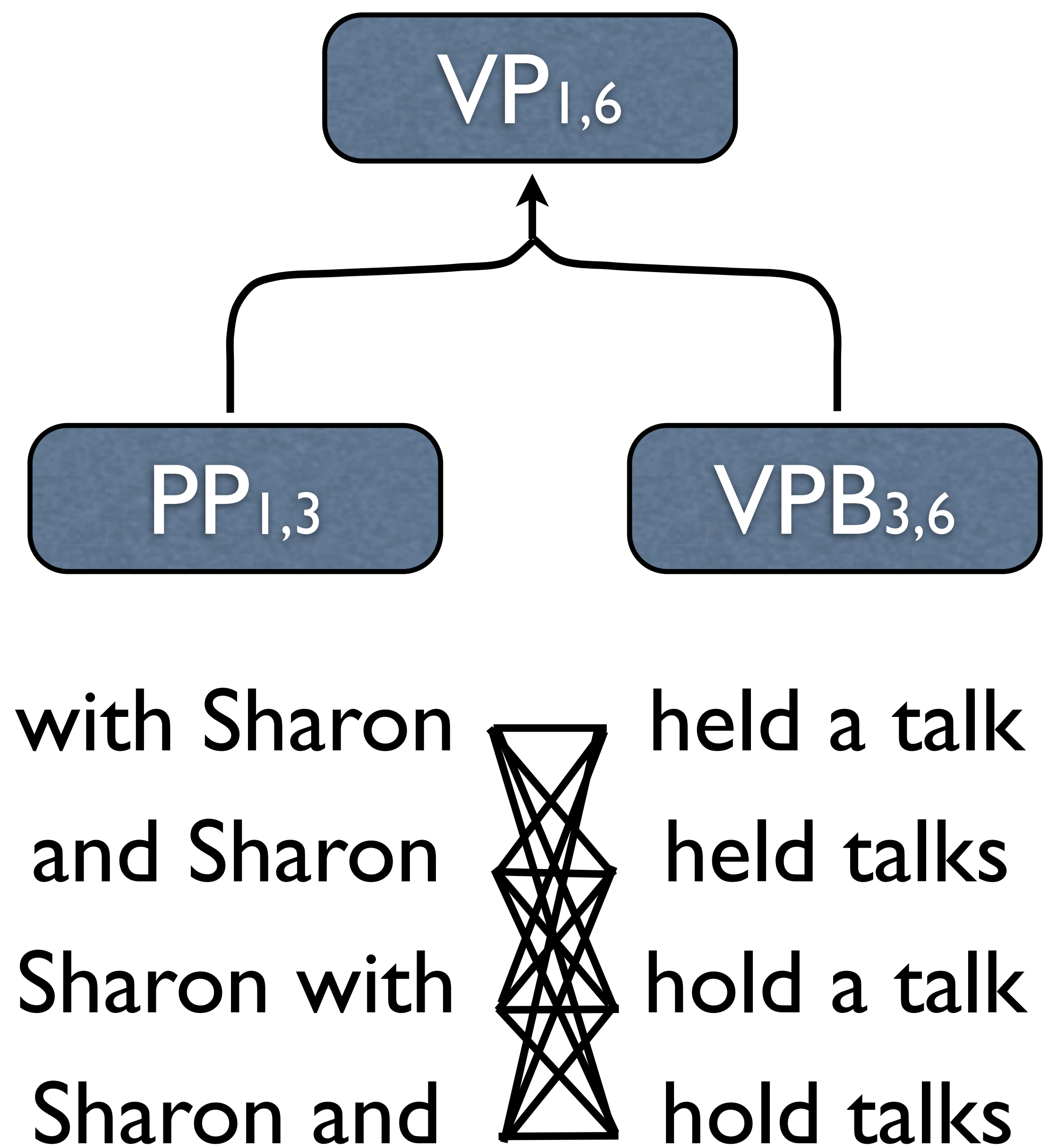




with Sharon	held a talk
and Sharon	held talks
Sharon with	hold a talk
Sharon and	hold talks

Exhaustive Search

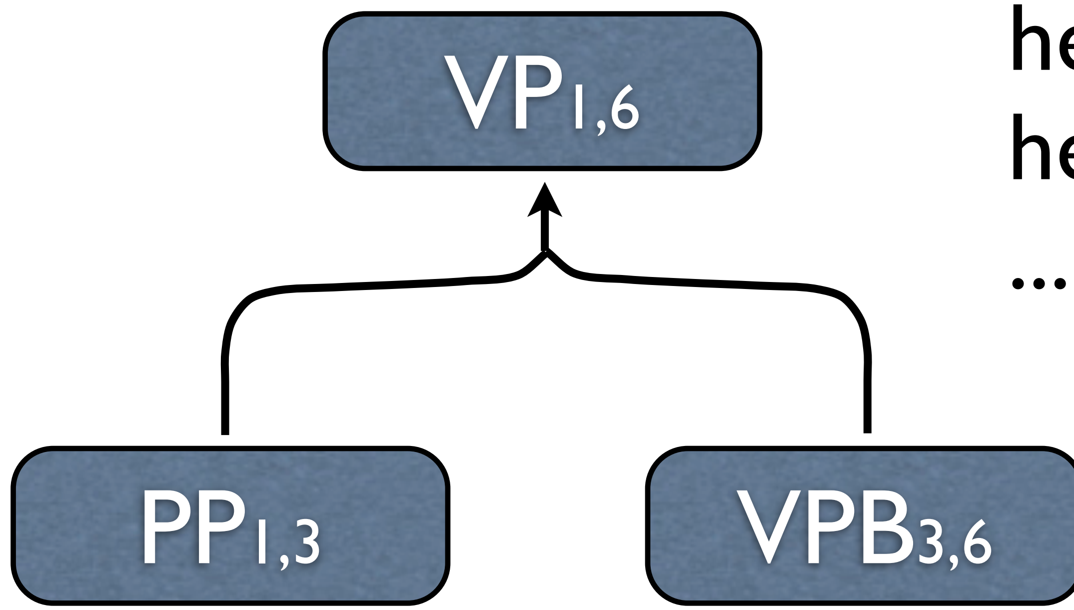




Exhaustive Search

VP → X1:PP X2:VPB

→ X2 X1



with Sharon	held a talk
and Sharon	held talks
Sharon with	hold a talk
Sharon and	hold talks

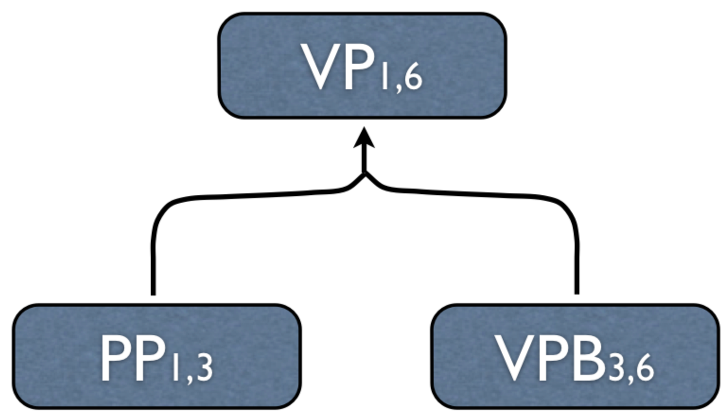
held a talk with Sharon
held a talk and Sharon
held talks with Sharon
held talks and Sharon
...

Calculating N-best List

$\mathbf{a}$	$\mathbf{b}$
1.0	1.3
2.9	1.5
3.4	4.2
5.6	5.1
6.0	6.3

What's the N-best list of $\mathbf{a}_i + \mathbf{b}_j$'s ?

Monotonicity

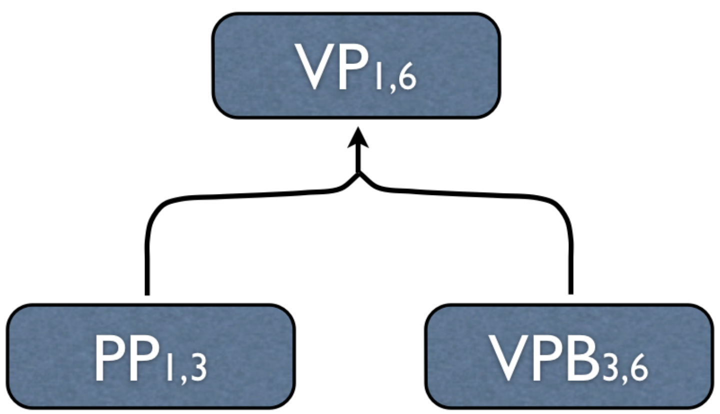


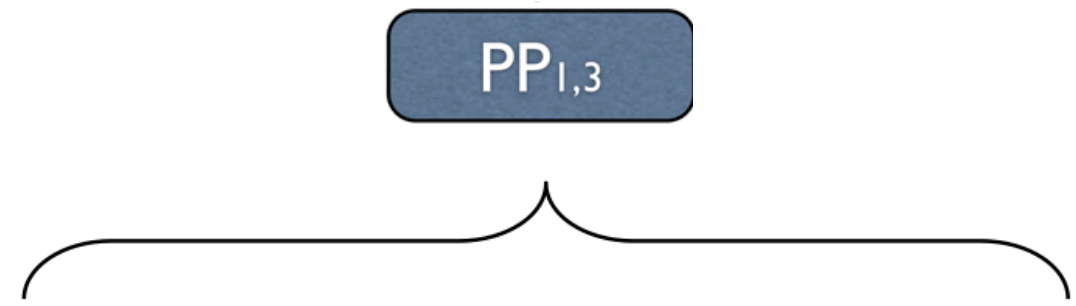
monotonic

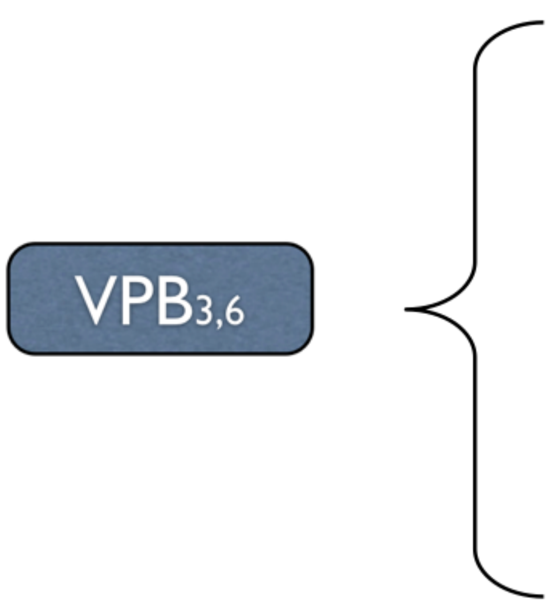
		PP$_{1,3}$: with * Sharon	and * Sharon	Sharon * with	Sharon * and
VPB$_{3,6}$		1.0	3.0	4.0	6.5
held * talk	1.0	2.0	4.0	5.0	7.5
held * talks	1.1	2.1	4.1	5.1	7.6
hold * talk	2.0	3.0	5.0	6.0	8.5
hold * talks	3.5	4.5	6.5	7.5	10.0

(Chiang 2007; Huang and Chiang, 2007)

Non-Monotonicity



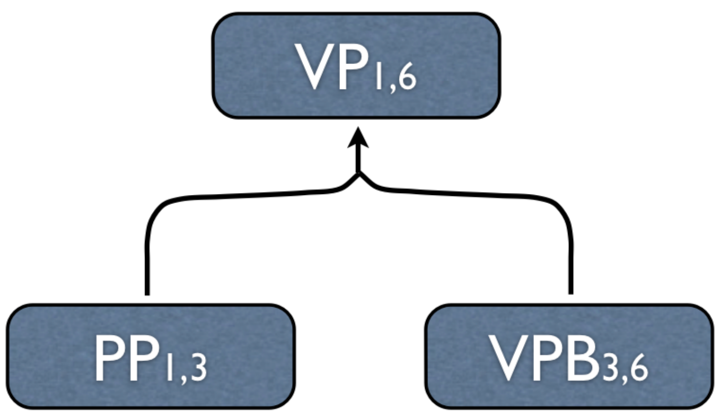




		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.0	4.0	5.0	7.5
held * talks	1.1	2.1	4.1	5.1	7.6
hold * talk	2.0	3.0	5.0	6.0	8.5
hold * talks	3.5	4.5	6.5	7.5	10.0

(Chiang 2007; Huang and Chiang, 2007)

Non-Monotonicity



$PP_{1,3}$

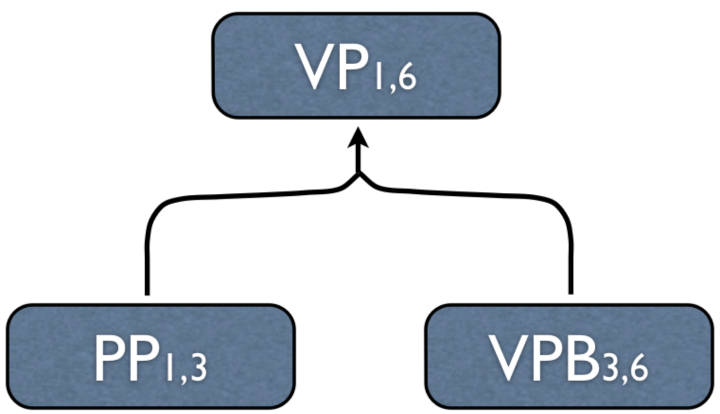
$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.0	4.0	5.0	7.5
held * talks	1.1	2.1	4.1	5.1	7.6
hold * talk	2.0	3.0	5.0	6.0	8.5
hold * talks	3.5	4.5	6.5	7.5	10.0

(Chiang 2007; Huang and Chiang, 2007)

Non-Monotonicity

$$\textcolor{red}{log(p(\text{with}|\text{talk})) - log(p(\text{with}))}$$



$PP_{1,3}$

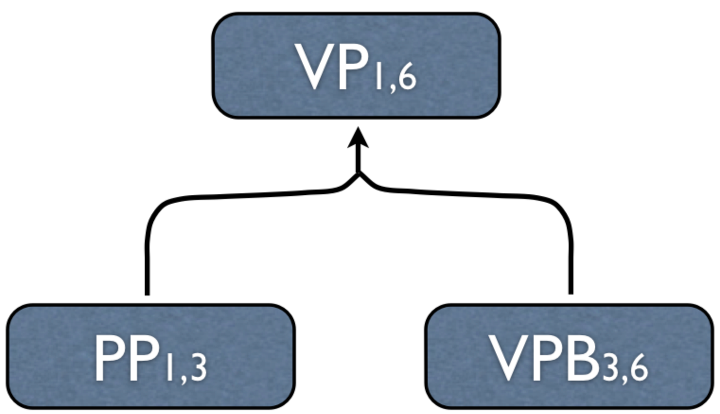
$VPB_{3,6}$

		$\boxed{\text{with}}$ * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * $\boxed{\text{talk}}$	1.0	2.0	4.0	5.0	7.5
held * talks	1.1	2.1	4.1	5.1	7.6
hold * talk	2.0	3.0	5.0	6.0	8.5
hold * talks	3.5	4.5	6.5	7.5	10.0

(Chiang 2007; Huang and Chiang, 2007)

Non-Monotonicity

$$log(p(\text{with}|\text{talk})) - log(p(\text{with}))$$



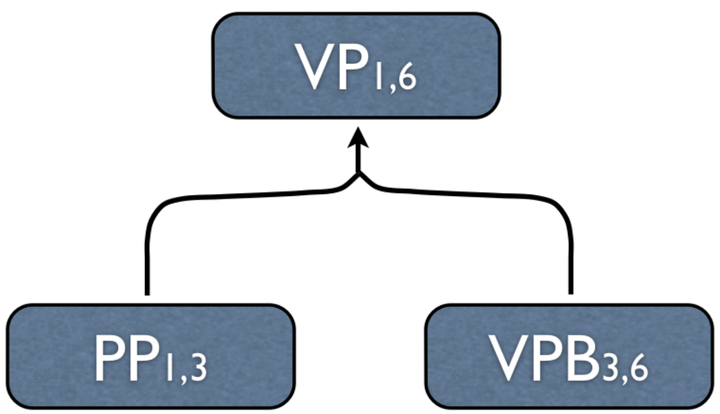
PP1,3

VPB3,6			$\boxed{\text{with}}$ * Sharon	and * Sharon	Sharon * with	Sharon * and
			1.0	3.0	4.0	6.5
	held * $\boxed{\text{talk}}$	1.0	2.0+0.5	4.0+2.0	5.0+4.0	7.5+4.0
	held * talks	1.1	2.1+0.3	4.1+1.5	5.1+3.5	7.6+3.0
	hold * talk	2.0	3.0+0.5	5.0+2.0	6.0+4.0	8.5+4.0
	hold * talks	3.5	4.5+0.3	6.5+1.5	7.5+3.5	10+3.5

(Chiang 2007; Huang and Chiang, 2007)

Non-Monotonicity

$$log(p(\text{with}|\text{talk})) - log(p(\text{with}))$$



LM introduces non-monotonicity

$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.0+0.5	4.0+2.0	5.0+4.0	7.5+4.0
held * talks	1.1	2.1+0.3	4.1+1.5	5.1+3.5	7.6+3.0
hold * talk	2.0	3.0+0.5	5.0+2.0	6.0+4.0	8.5+4.0
hold * talks	3.5	4.5+0.3	6.5+1.5	7.5+3.5	10+3.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue

4-best

$PP_{1,3}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

$VPB_{3,6}$

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue

4-best

$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue: 2.5

4-best

PP$_{1,3}$

VPB$_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue

4-best 2.5

$PP_{1,3}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

$VPB_{3,6}$

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue

4-best 2.5

$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue: 2.4

4-best: 2.5

$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

PP1,3

queue 2.4 6.0

4-best 2.5

VPB3,6

VPB3,6		with * Sharon 1.0	and * Sharon 3.0	Sharon * with 4.0	Sharon * and 6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue: 6.0

4-best: 2.4 2.5

$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue | 6.0

4-best | 2.4 | 2.5

$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue: 6.0 3.5

4-best: 2.4 2.5

$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue: 6.0 | 3.5 | 5.6

4-best: 2.4 | 2.5

$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue: 3.5 | 5.6 | 6.0

4-best: 2.4 | 2.5

$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue 5.6 6.0

4-best 2.4 2.5 3.5

$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue: 5.6 6.0

4-best: 2.4 2.5 3.5

$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue: 5.6 | 6.0 | 4.8

4-best: 2.4 | 2.5 | 3.5

$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue	5.6	6.0	4.8	7.0
4-best	2.4	2.5	3.5	

$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue: 4.8 | 5.6 | 6.0 | 7.0

4-best: 2.4 | 2.5 | 3.5

$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue	5.6	6.0	7.0	
4-best	2.4	2.5	3.5	4.8

$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning

queue: 5.6 6.0 7.0

4-best: 2.4 2.5 3.5 4.8

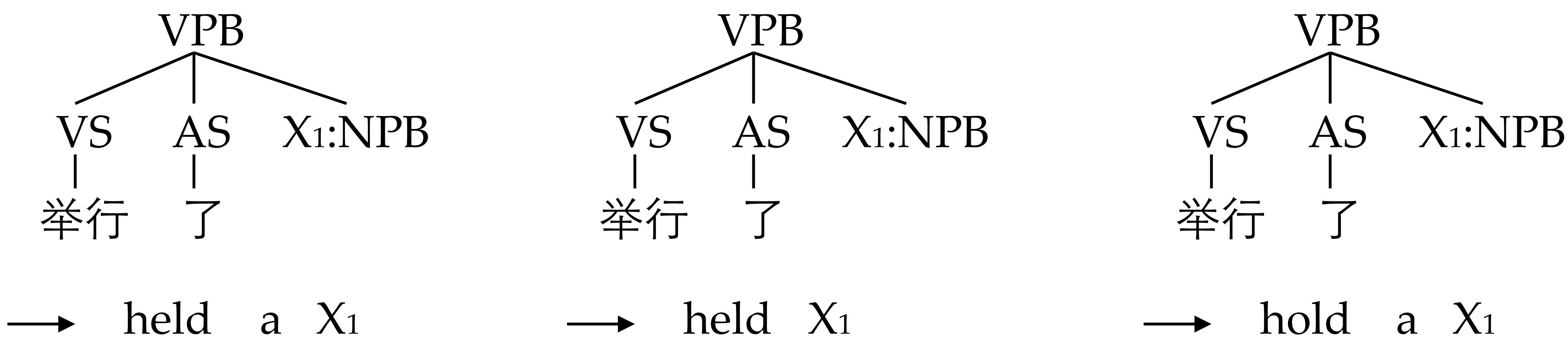
$PP_{1,3}$

$VPB_{3,6}$

		with * Sharon	and * Sharon	Sharon * with	Sharon * and
		1.0	3.0	4.0	6.5
held * talk	1.0	2.5	6.0	9.0	11.5
held * talks	1.1	2.4	5.6	8.6	10.6
hold * talk	2.0	3.5	7.0	10.0	12.5
hold * talks	3.5	4.8	8.0	11.0	13.5

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning with Rule Group



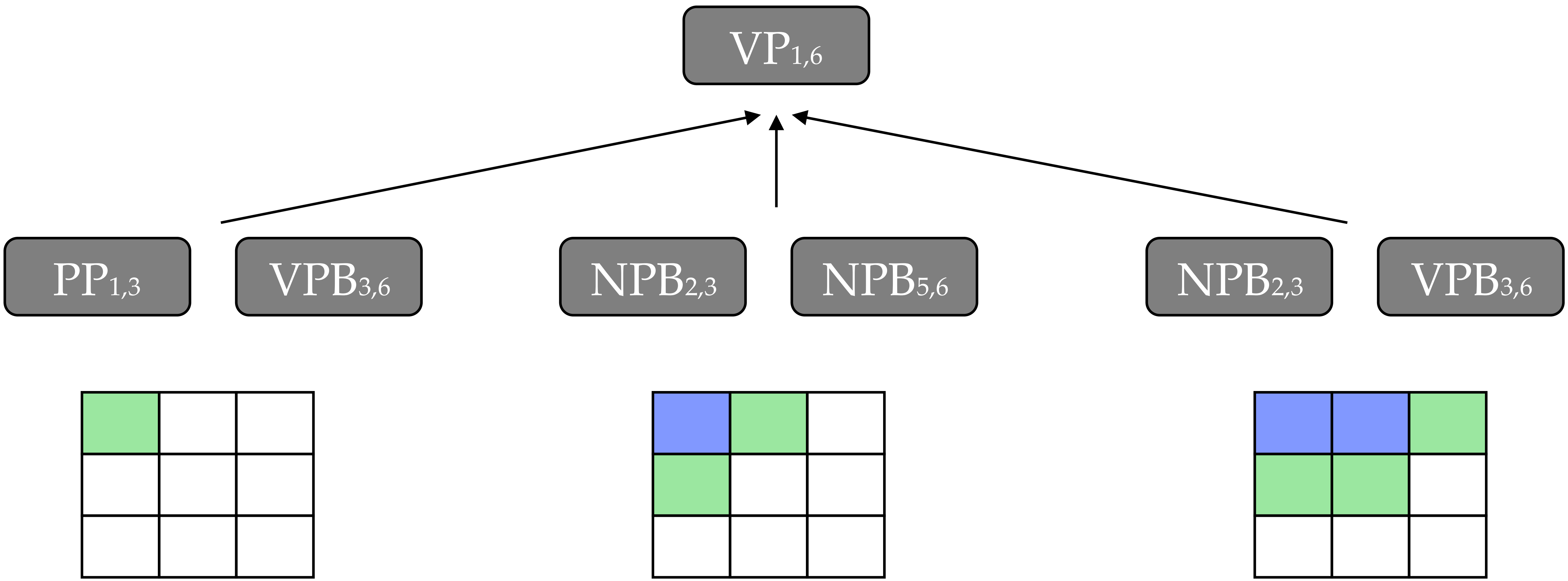
$NPB_{5,6}$

Group rules that have the same LHS

		talk	talks	meeting
		1.0	2.0	2.5
held a X_1	1.0	2.1	5.0	3.7
held X_1	1.4	3.2	4.0	5.0
hold a X_1	2.0	3.1	6.0	4.7

(Chiang 2007; Huang and Chiang, 2007)

Cube Pruning within Node



(Chiang 2007; Huang and Chiang, 2007)

Syntax-based MT

SCFGs without linguistic syntax

inverted transduction grammar

hierarchical phrase-based model

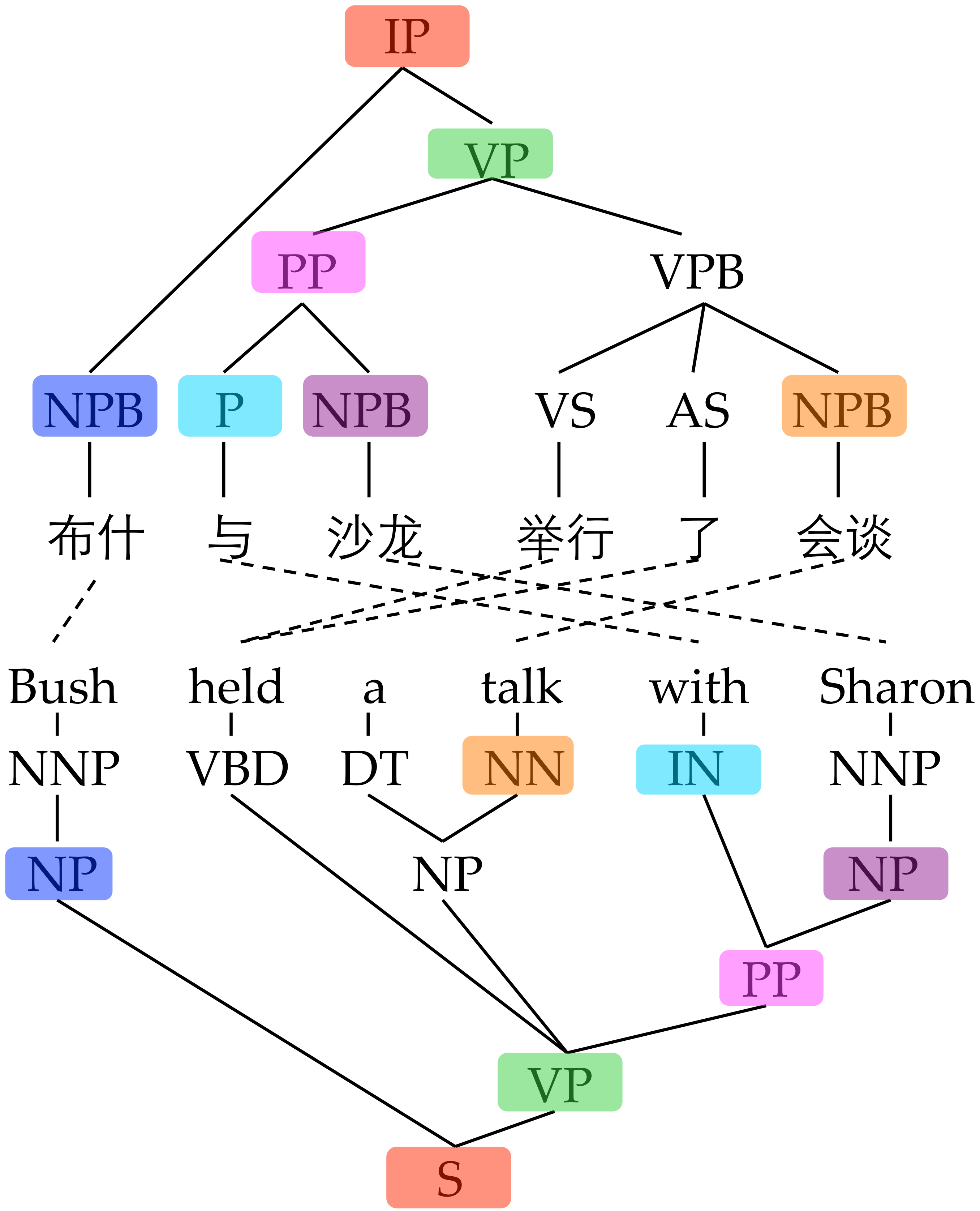
STSGs with linguistic syntax

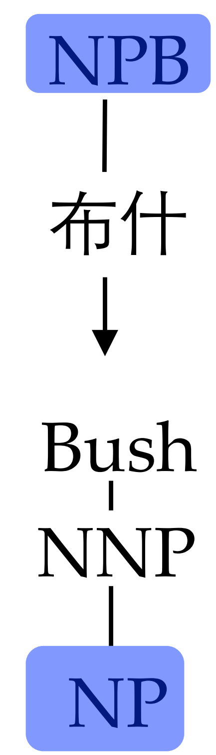
string-to-tree

tree-to-string

tree-to-tree

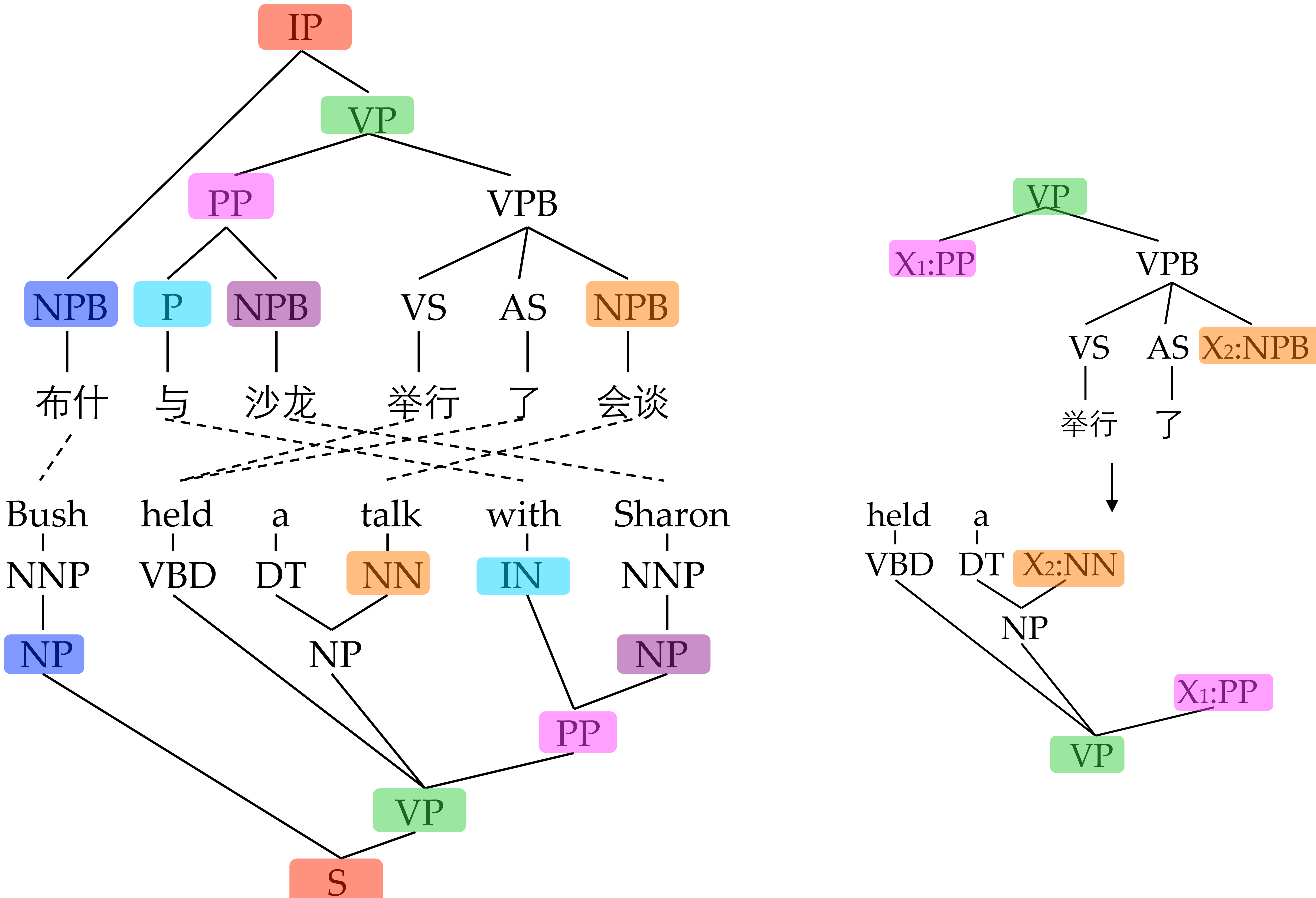
Tree-to-Tree Translation





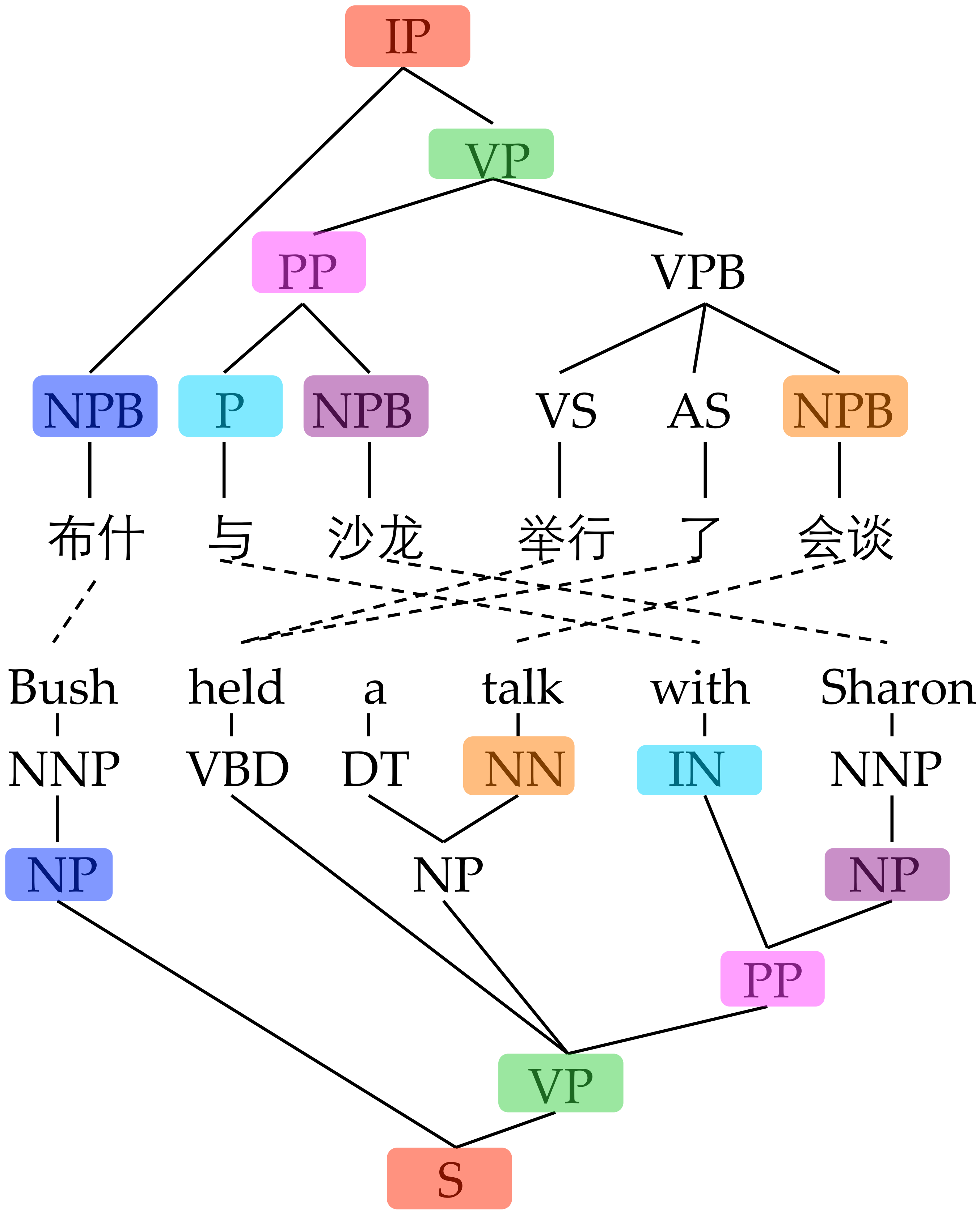
(Zhang et al. 2008; Liu et al., 2009)

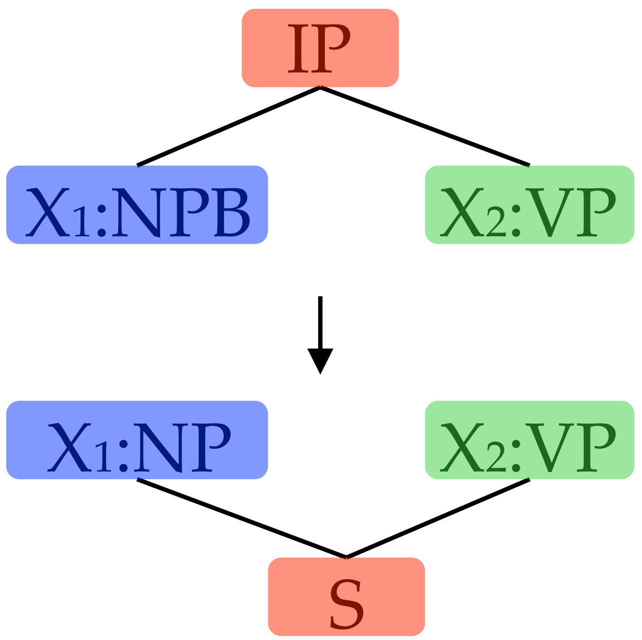
Tree-to-Tree Translation



(Zhang et al. 2008; Liu et al., 2009)

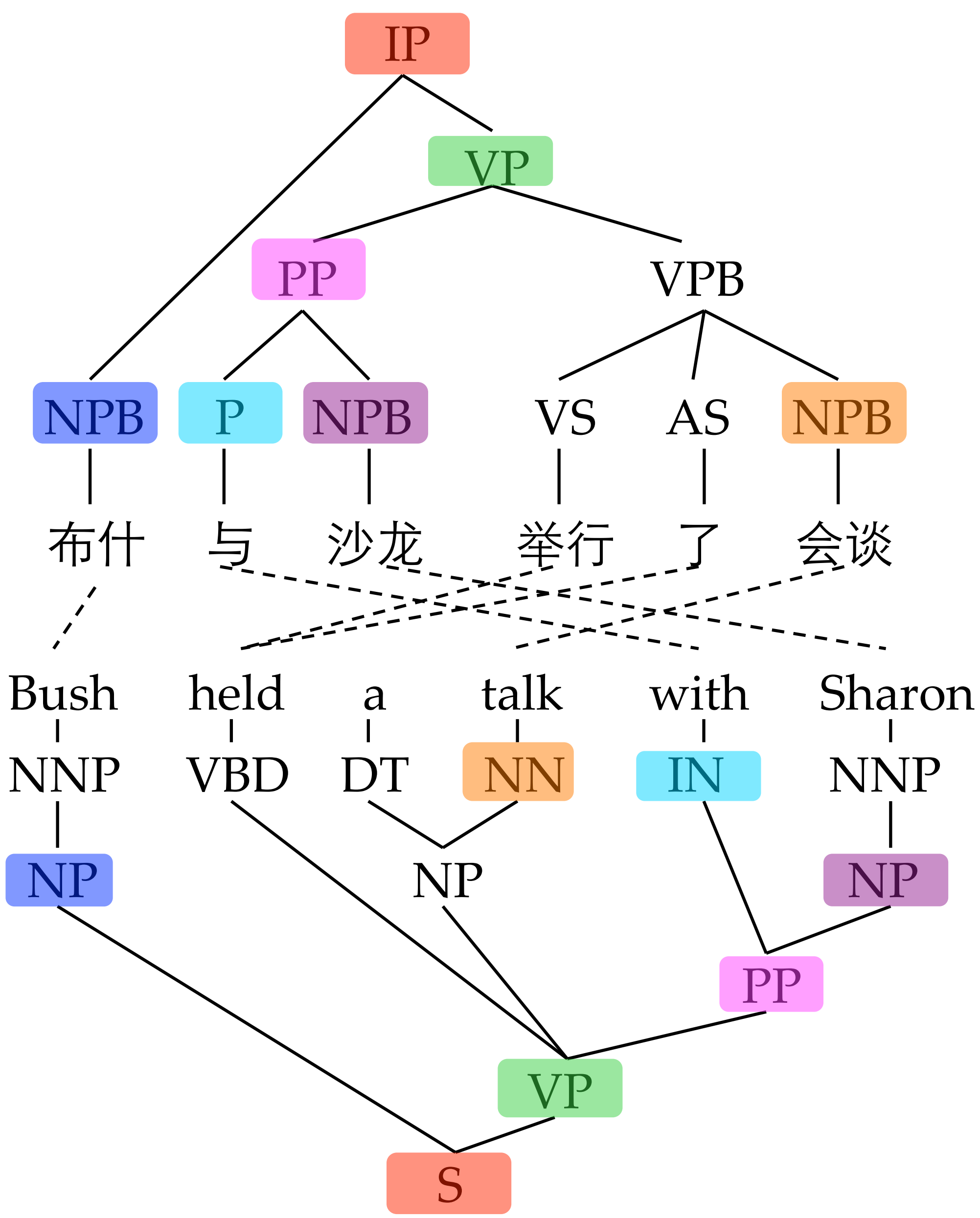
Tree-to-Tree Translation





(Zhang et al. 2008; Liu et al., 2009)

Rule Coverage



phrase pair	s2s	t2s	s2t	t2t
(布什, Bush)	√	√	√	√
(与, with)	√	√	√	√
(沙龙, Sharon)	√	√	√	√
(会谈, talk)	√	√	√	√
(与 沙龙, with Sharon)	√	√	√	√
(举行 了, held)	√	×	√	×
(举行… 会谈, held … talk)	√	√	×	×
(与 … 会谈, held … Sharon)	√	√	√	√
(布什 … 会谈, Bush … Sharon)	√	√	√	√
	100%	89%	89%	78%

Rule Coverage

model	human	automatic
string-to-string	100%	100%
tree-to-string	78%	75%
string-to-tree	76%	72%
tree-to-tree	68%	60%

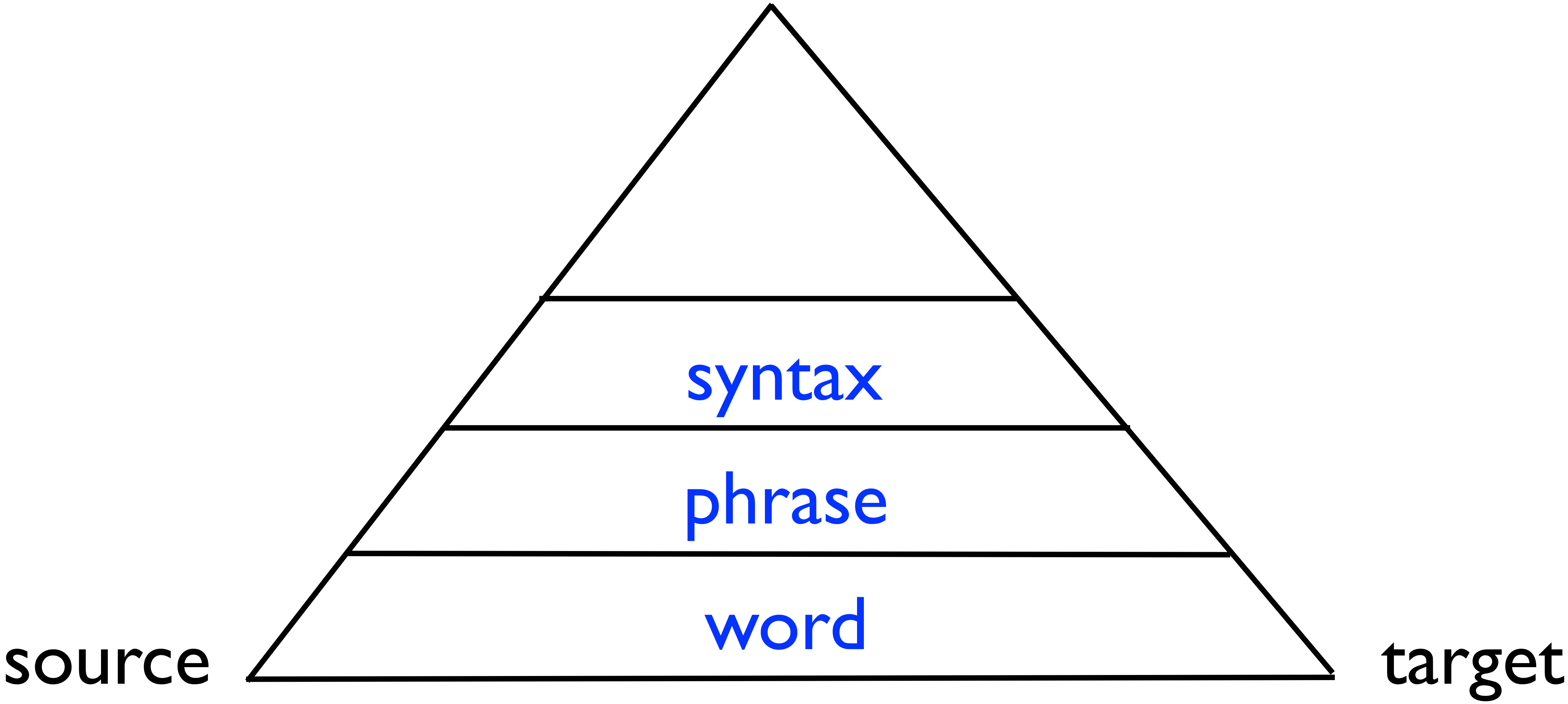
(Chiang, 2010)

Summary of Syntax-based Models

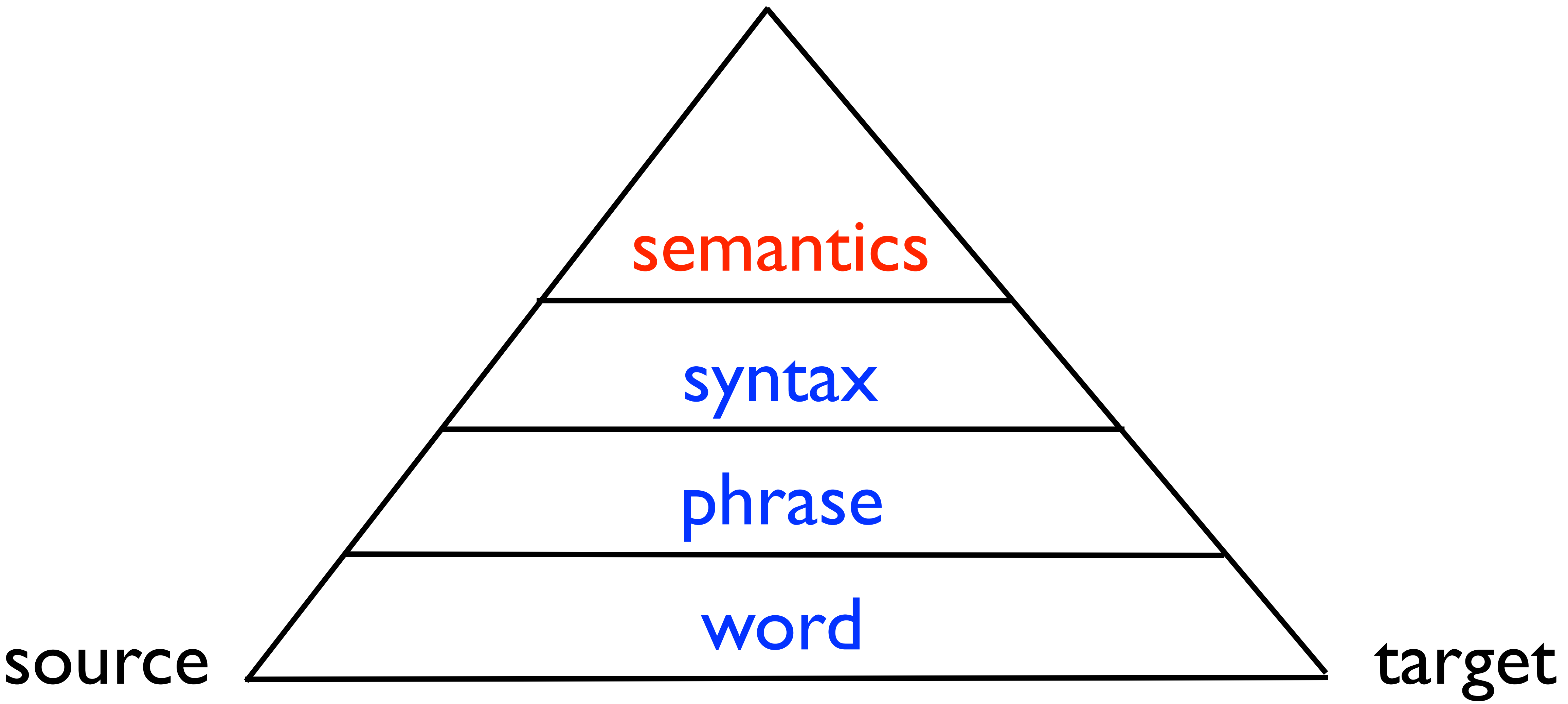
source	target	model	examples
N/A	N/A	string-to-string	Chiang (2005) Wu (1997)
N/A	syntax	string-to-tree	Galley et al. (2006) Shen et al. (2008)
syntax	N/A	tree-to-string	Liu et al. (2006) Huang et al. (2006)
syntax	syntax	tree-to-tree	Eisner (2003) Zhang et al. (2008)

Part 5: Future Directions

The SMT Pyramid



The SMT Pyramid



Semantics-based Translation

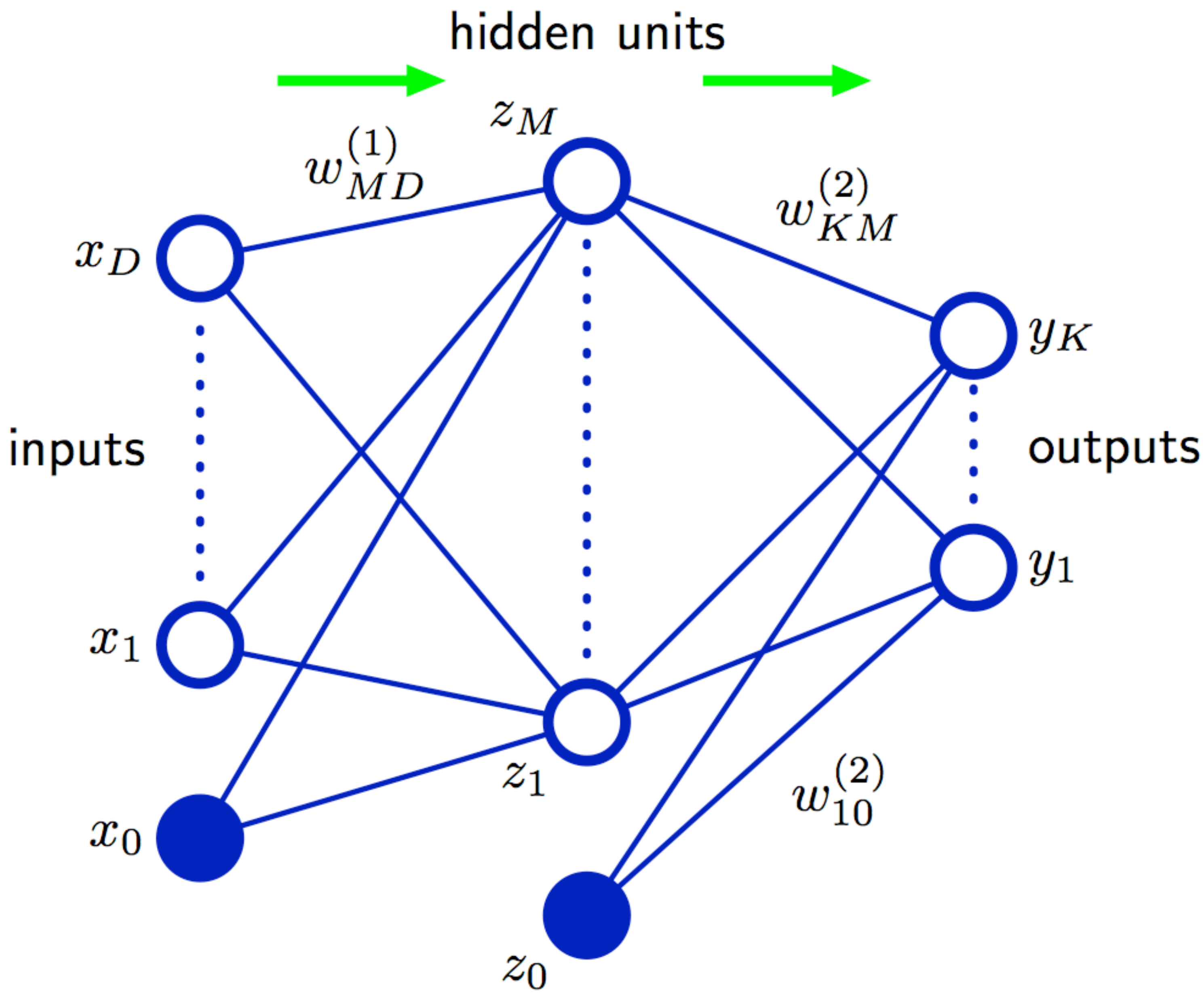
- word sense disambiguation
  - WSD does not help (Carpuat and Wu, 2005)
  - WSD does help (Carpuat and Wu, 2007; Chan et al., 2007)
- semantic role labeling
  - semantic role features (Liu and Gildea, 2010)
  - predicate-argument structure (Xiong et al., 2012)
- grammars
  - hyperedge replacement grammars (Jones et al., 2012)

Deep Learning for MT

- deep neural network for word alignment (Yang et al., 2013)
- additive neural networks for translation (Liu et al., 2013)
- recurive autoencoders for ITG-based translation (Li et al., 2013)
- recurrent continuous translation models (Kalchbrenner and Blunsom, 2013)

Neural Networks for MT

Neural Network

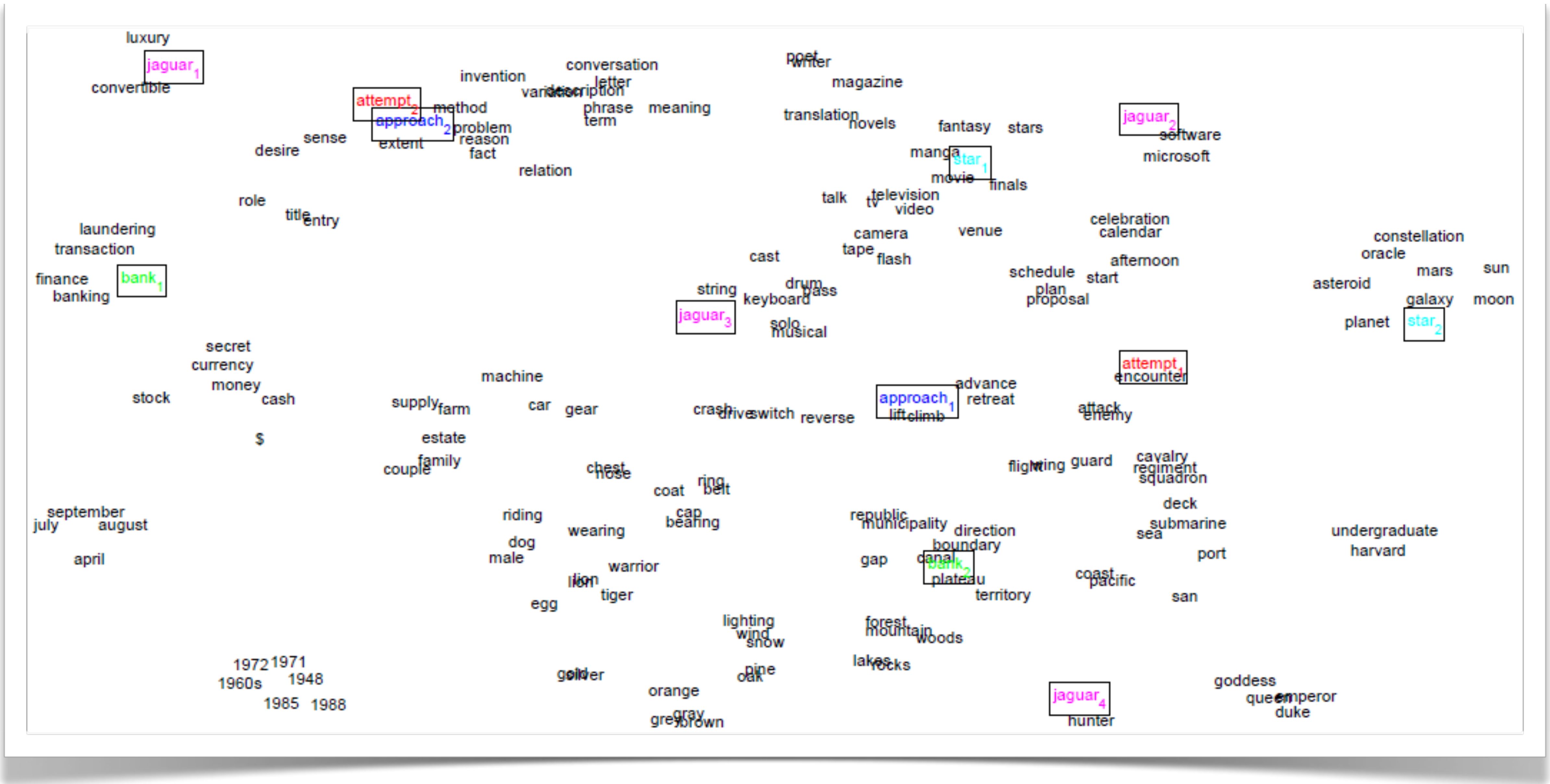


$$y_k(\mathbf{x}, \mathbf{w}) = \sigma\left(\sum_{j=1}^{M} w^{(2)}_{kj} h\left(\sum_{i=1}^{D} w^{(1)}_{ji} x_i + w^{(1)}_{j0}\right) + w^{(2)}_{k0}\right)$$

adapted from PRML by Bishop

Word Embedding

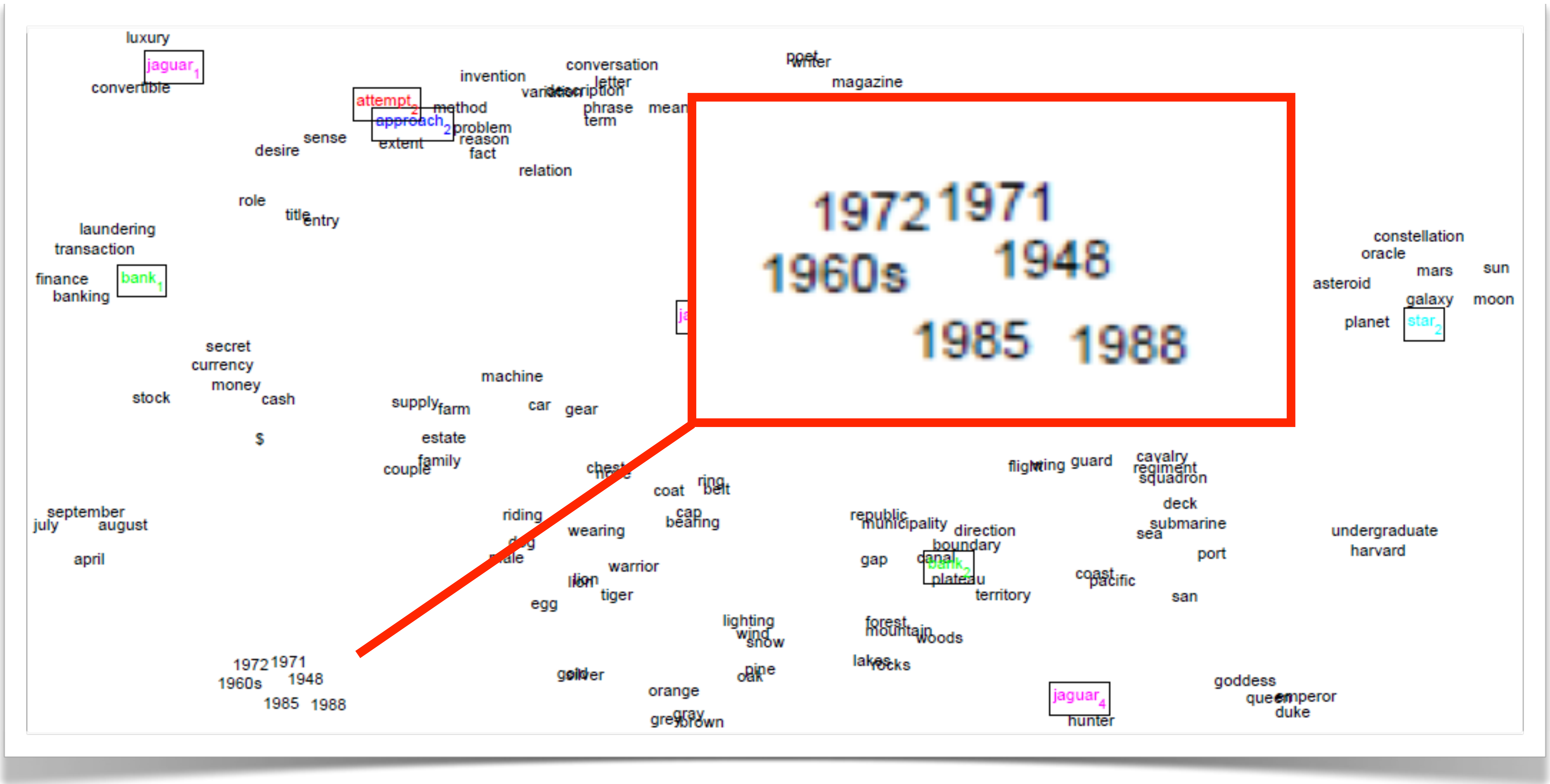
- a word is represented as a real-valued vector



from Socher's tutorial

Word Embedding

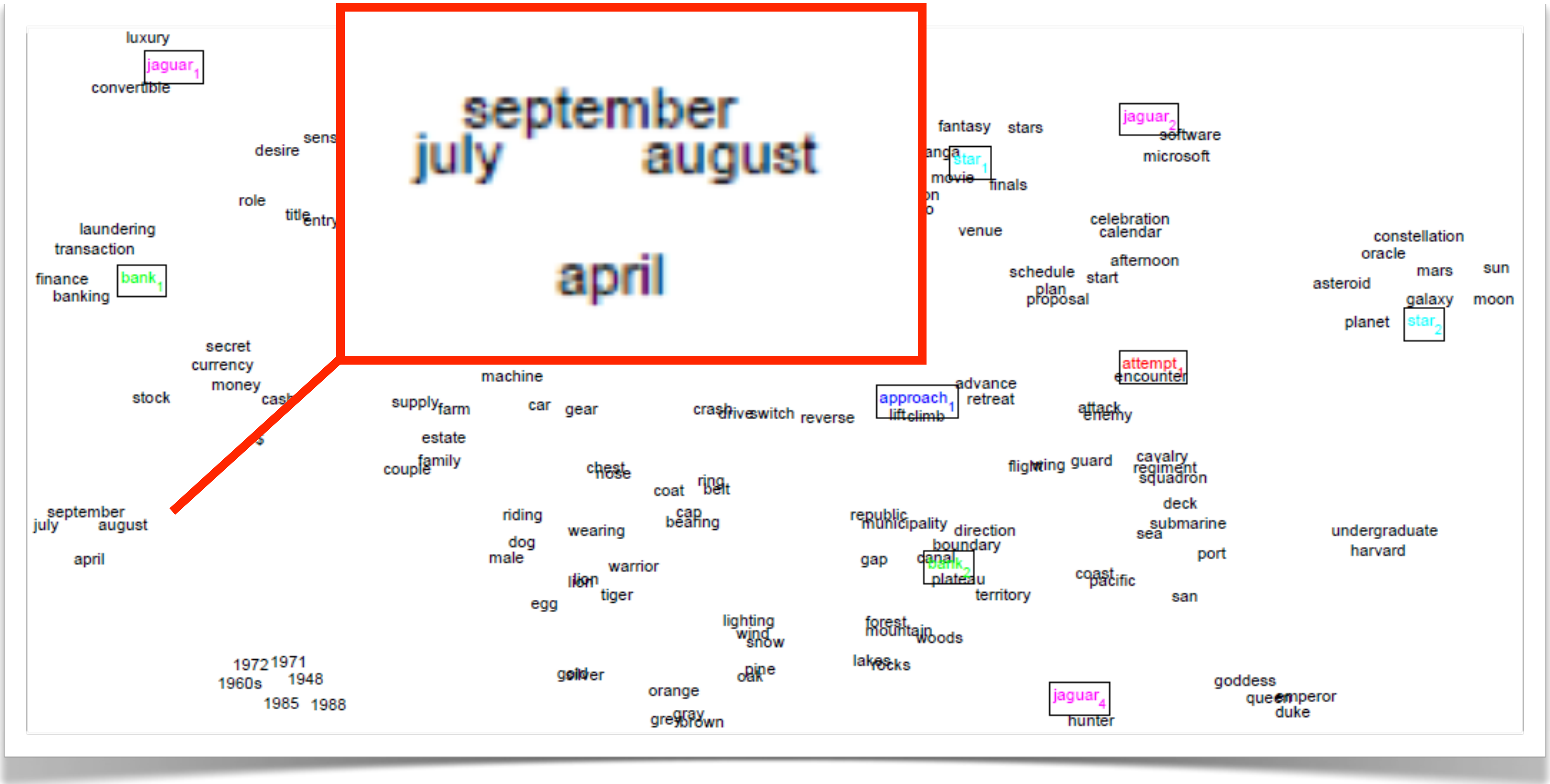
- a word is represented as a real-valued vector



from Socher's tutorial

Word Embedding

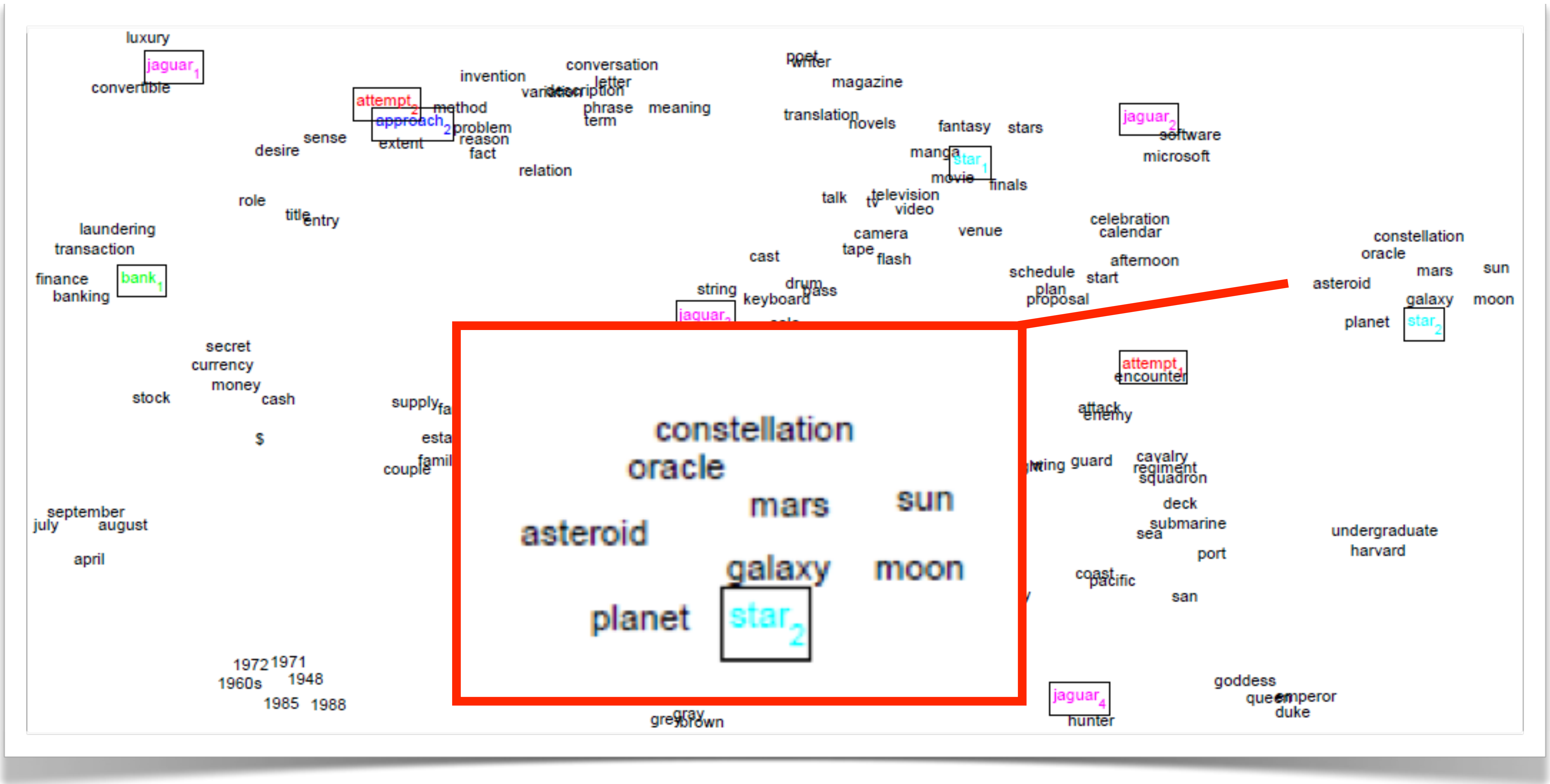
- a word is represented as a real-valued vector



from Socher's tutorial

Word Embedding

- a word is represented as a real-valued vector



from Socher’s tutorial

Word Embedding

ThisPlusThat.me

Amazing language relationships

yao ming - China + USA | Search | How it Works

The Matrix - Thoughtful + Dumb | Harry Truman - American + Russian | MIT - smart + pretentious | Mitt Romney - Experience + Celebrity | Darth Vader - Cape | Justin Bieber - man + woman

Your query was disambiguated into *+1 yao_ming -1 china +1 usa* in 3.8 seconds from ip-10-184-53-69

PERSON EXTRA, FILM ACTOR, PERSON,

Tracy McGrady

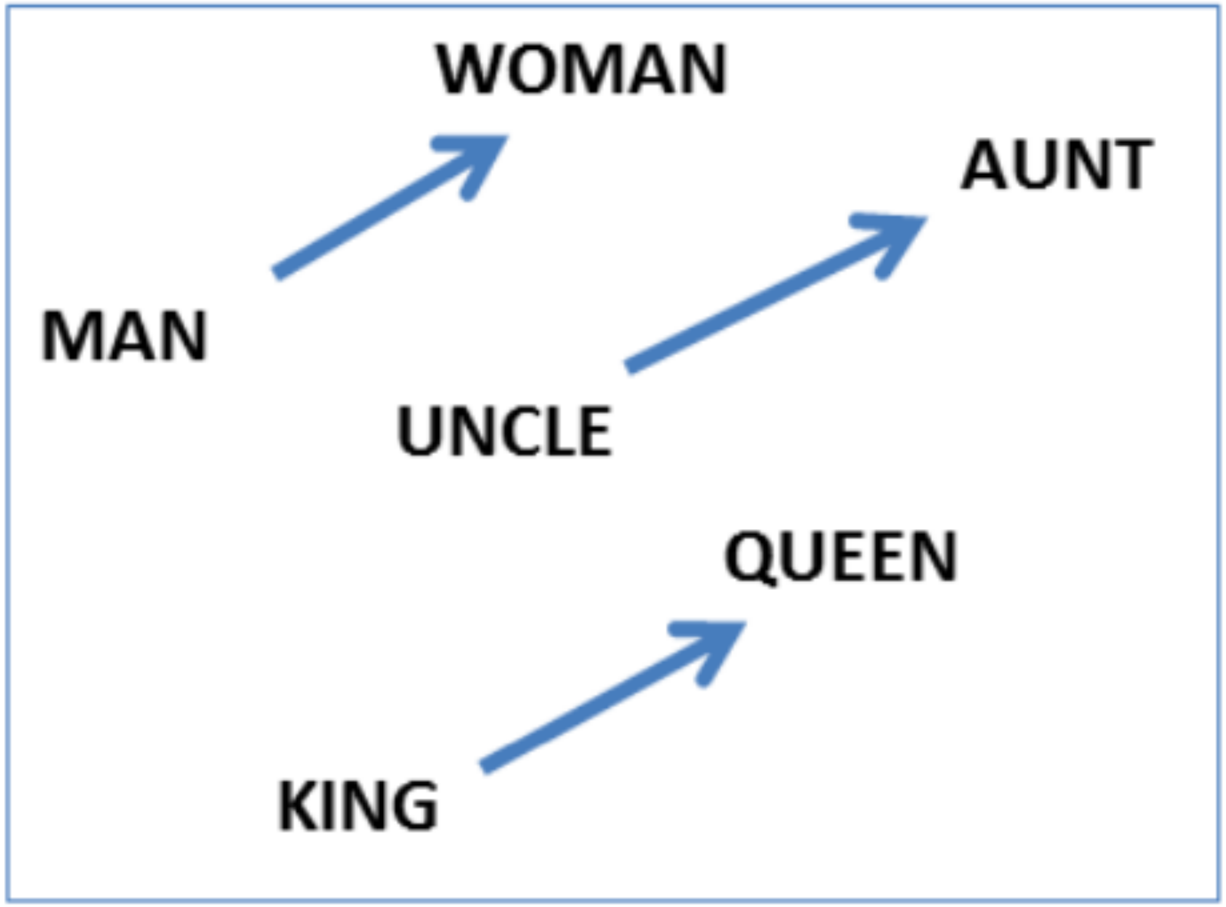
Tracy Lamar McGrady, Jr. is an American former professional basketball player who last played for the San Antonio Spurs of the National Basketball Association. He is a seven-time NBA All-Star, seven-time All-NBA selection, and a two-time NBA scoring champion.

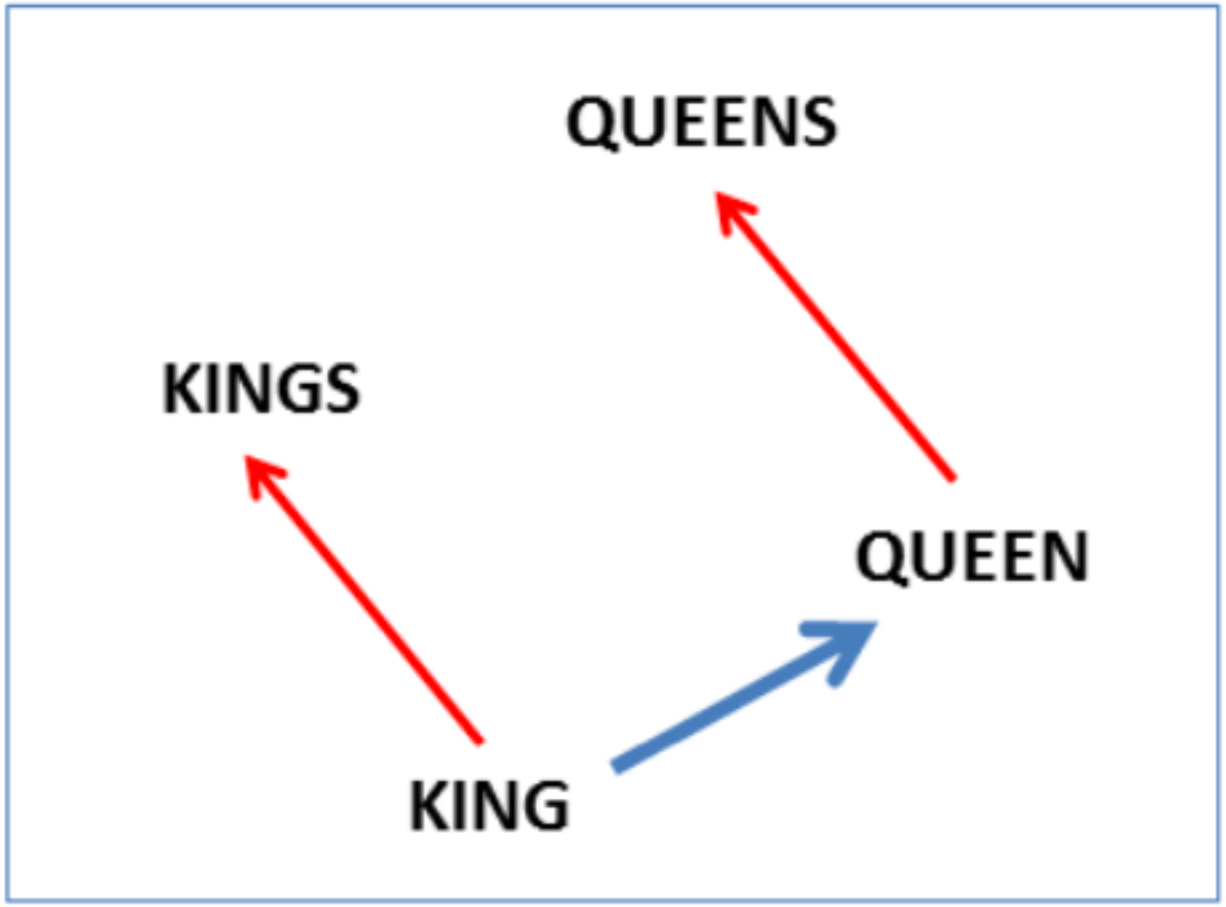
Basketball Shooting guard

Score: 0.28

http://www.thisplusthat.me

Word Embedding





(Mikolov et al., 2013)

Word Embedding

Category	Relation	Example
Adjectives	Base/Comparative	good:better rough:___
Adjectives	Base/Superlative	good:best rough:___
Adjectives	Comparative/ Superlative	better:best rougher:___
Nouns	Singular/Plural	year:years law:___
Nouns	Non-possessive/ Possessive	city:city's bank:___
Verbs	Base/Past	see:saw return:___
Verbs	Base/3rd Person Singular Present	see:sees return:___
Verbs	Past/3rd Person Singular Present	saw:sees returned:___

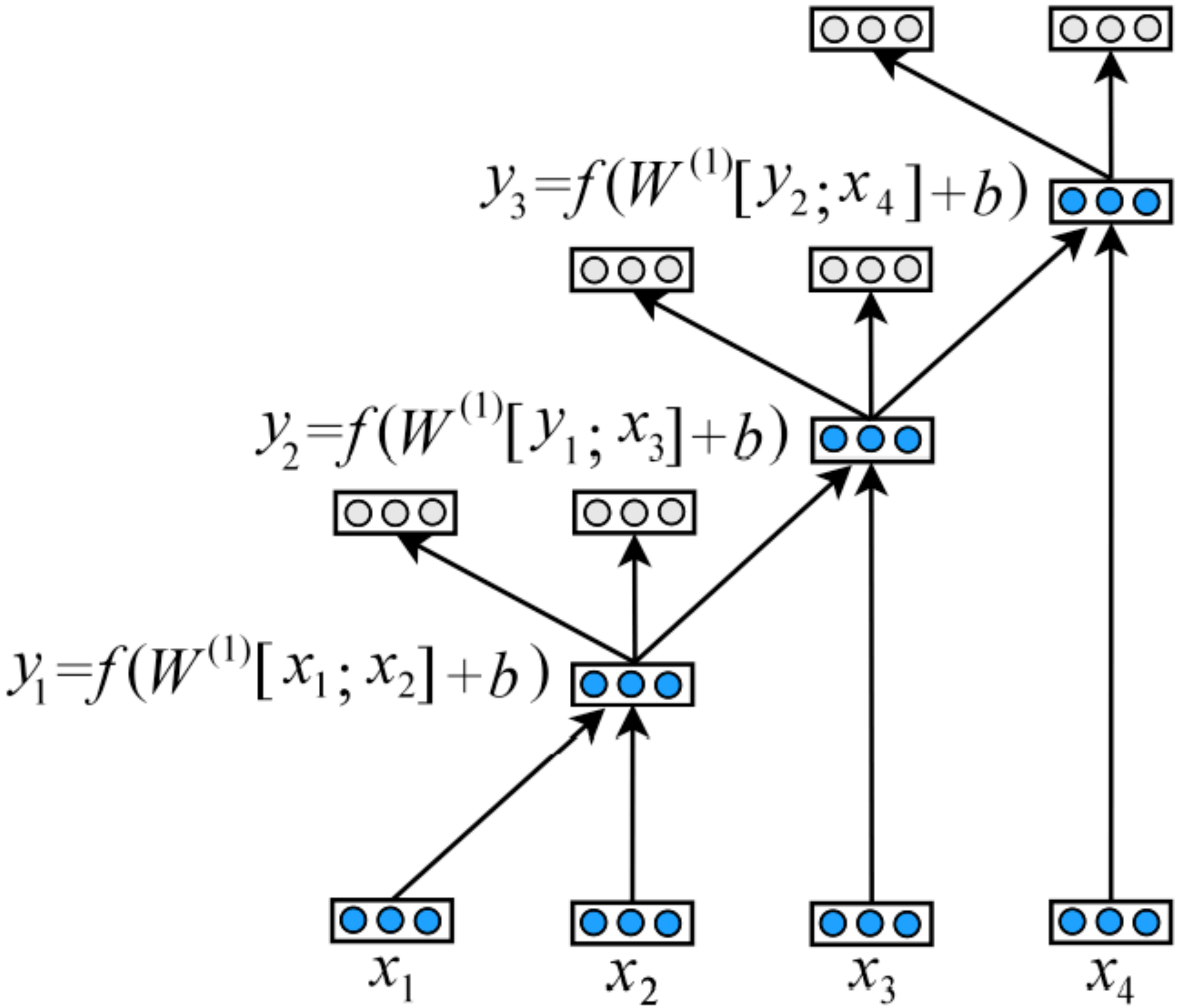
(Mikolov et al., 2013)

Word Embedding

Category	Relation	Example
Adjectives	Base/Comparative	good:better rough:___
Adjectives	Base/Superlative	good:best rough:___
Adjectives	Comparative/ Superlative	better:best rougher:___
Nouns	Singular/Plural	year:years law:___
Nouns	Non-possessive/ Possessive	city:city's bank:___
Verbs	Base/Past	see:saw return:___
Verbs	Base/3rd Person Singular Present	see:sees return:___
Verbs	Past/3rd Person Singular Present	saw:sees returned:___

Q: vectors for variable-sized phrases? (Mikolov et al., 2013)

Recursive Autoencoders



(Socher et al., 2011)

Reordering as Classification

与 沙龙	举行 了 会谈
with Sharon	held a talk

(Xiong et al., 2006)

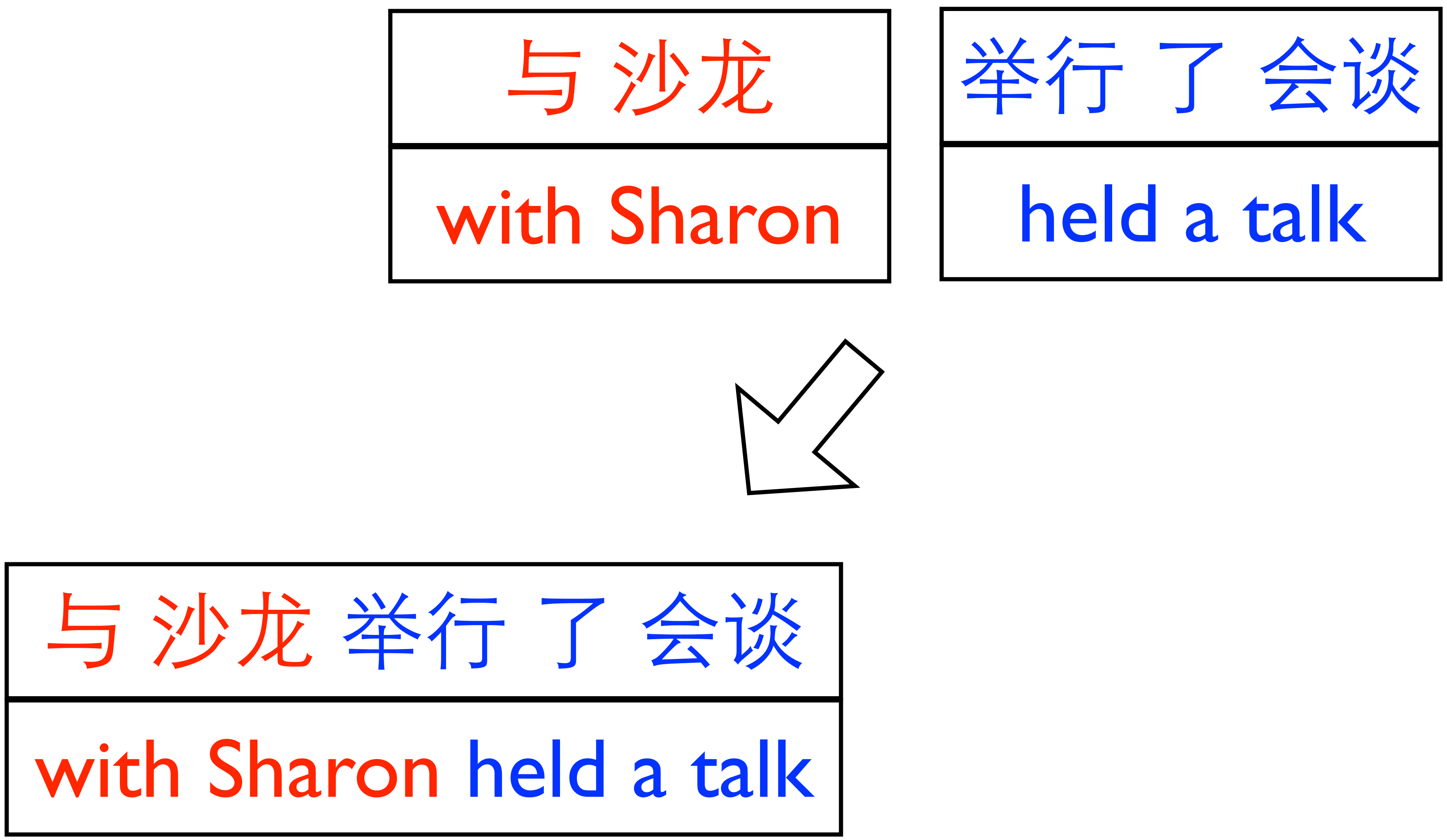
Reordering as Classification

与 沙龙	举行 了 会谈
with Sharon	held a talk

与 沙龙 举行 了 会谈
with Sharon held a talk

(Xiong et al., 2006)

Reordering as Classification



straight

(Xiong et al., 2006)

Reordering as Classification

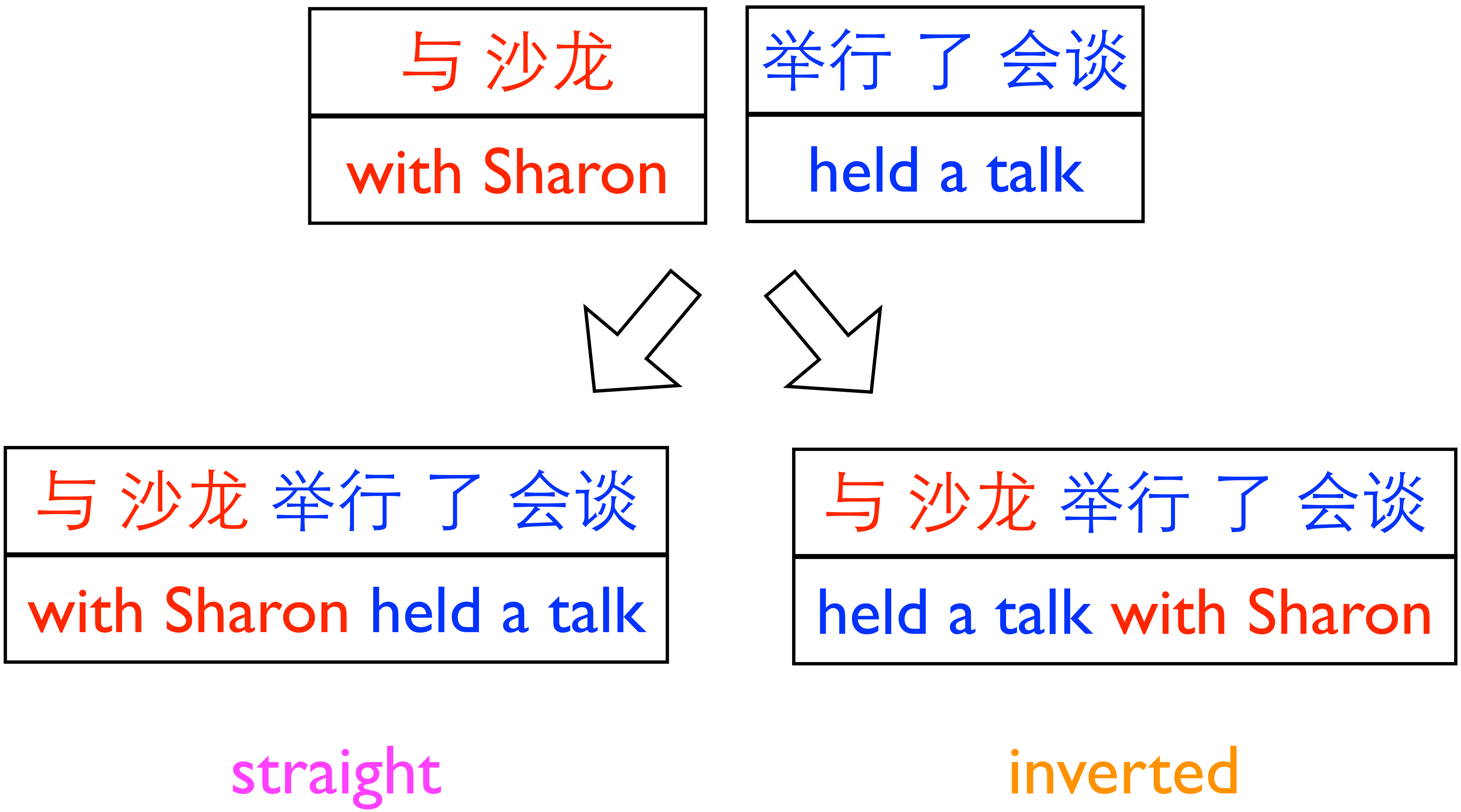
与 沙龙	举行 了 会谈
with Sharon	held a talk

与 沙龙 举行 了 会谈	与 沙龙 举行 了 会谈
with Sharon held a talk	held a talk with Sharon

straight

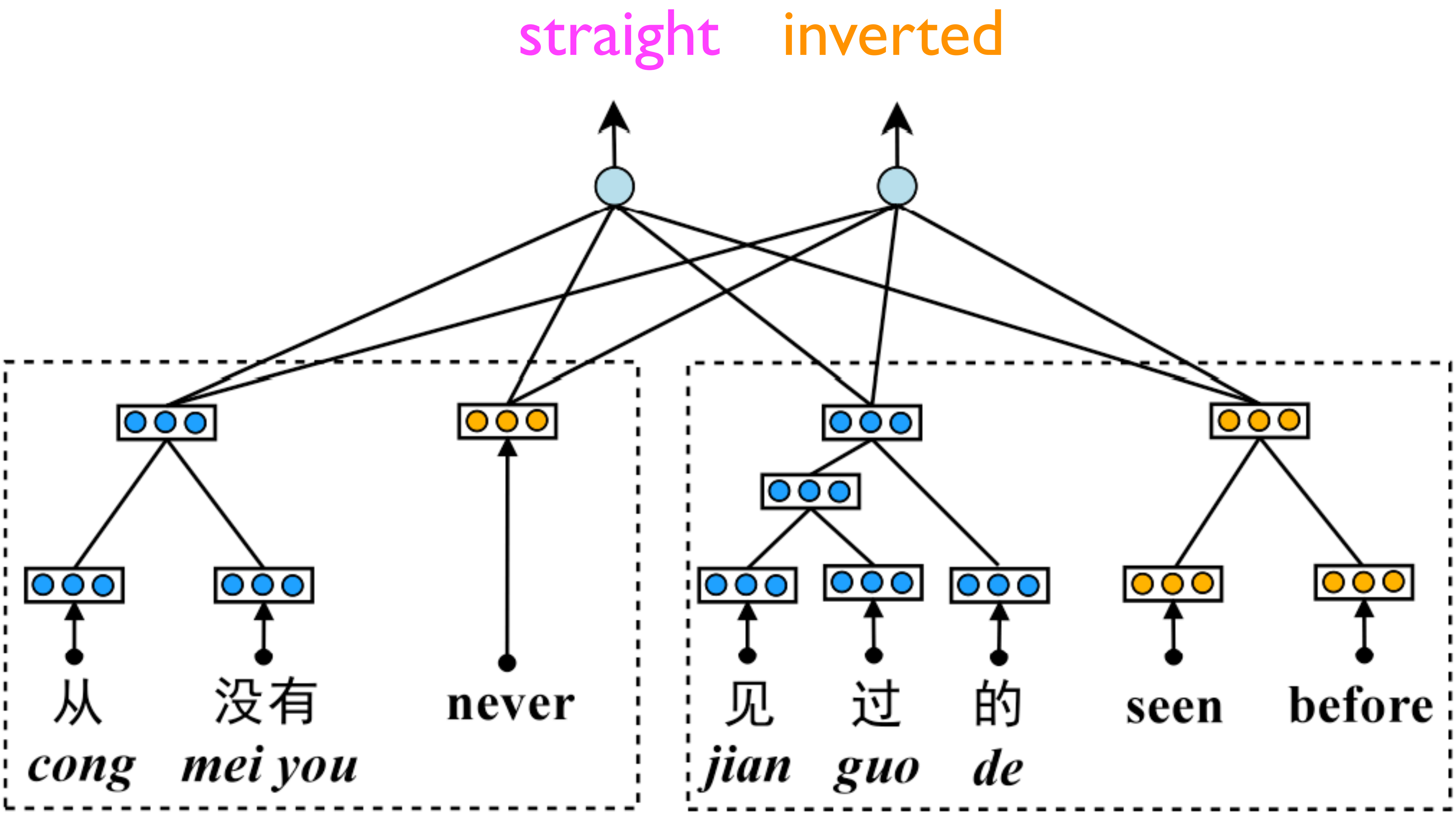
(Xiong et al., 2006)

Reordering as Classification



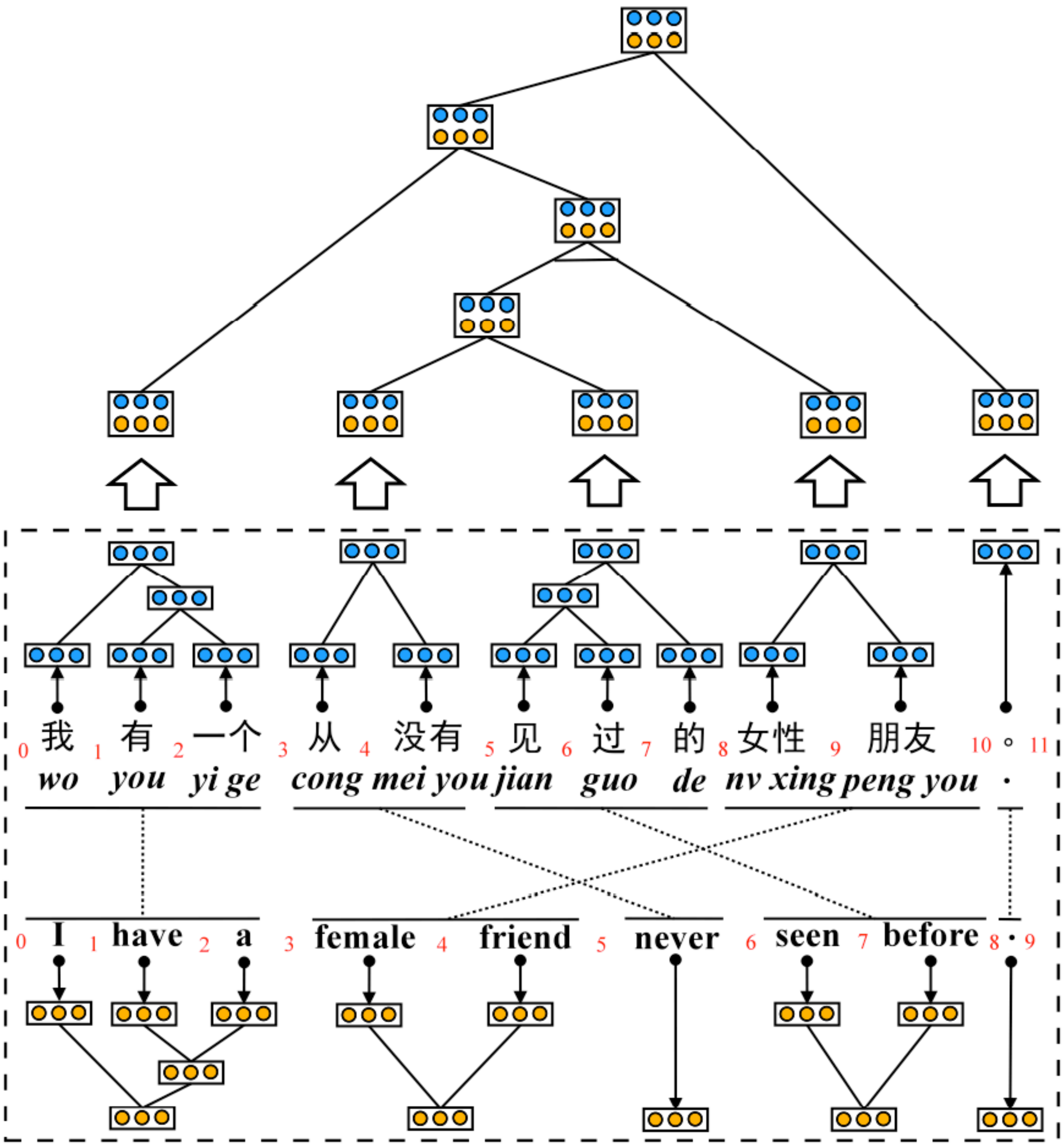
(Xiong et al., 2006)

Neural Classifier for ITG



(Li et al., 2013)

Neural ITG-based Translation



(Li et al., 2013)

Neural ITG-based Translation

cluster 1	cluster 2	cluster 3
1.18	fairly	stand alone
accessibility	harmful	one-day conference
wheelchair	shown	two-way links
candies	pretty	individual unit
cough	adopting	early july

cluster 4	cluster 5
these people who	in the same manner
their feelings about	of last century
the system which	by the year 2010
the economic sanctions against	in next week
its attutude toward	within waters

(Li et al., 2013)

Future Directions

	state-of-the-art	future
“pyramid”		
unit		
modal		
intelligence		

Future Directions

	state-of-the-art	future
“pyramid”	syntax	
unit		
modal		
intelligence		

Future Directions

	state-of-the-art	future
"pyramid"	syntax	semantics
unit		
modal		
intelligence		

Future Directions

	state-of-the-art	future
"pyramid"	syntax	semantics
unit	sentence	
modal		
intelligence		

Future Directions

	state-of-the-art	future
“pyramid”	syntax	semantics
unit	sentence	document
modal		
intelligence		

Future Directions

	state-of-the-art	future
“pyramid”	syntax	semantics
unit	sentence	document
modal	text	
intelligence		

Future Directions

	state-of-the-art	future
"pyramid"	syntax	semantics
unit	sentence	document
modal	text	text, vision, speech
intelligence		

Future Directions

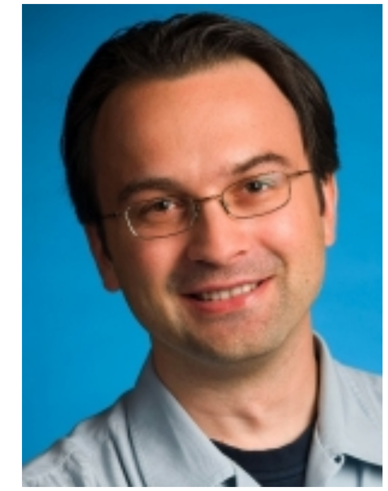
	state-of-the-art	future
“pyramid”	syntax	semantics
unit	sentence	document
modal	text	text, vision, speech
intelligence	“learning”	

Future Directions

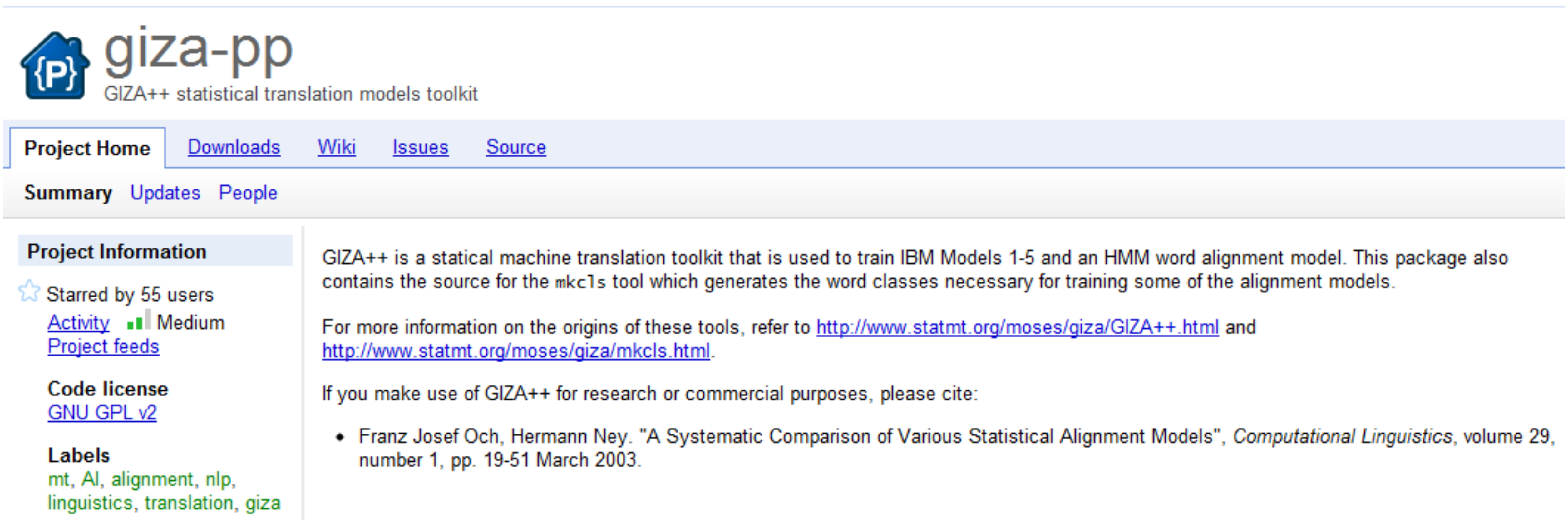
	state-of-the-art	future
"pyramid"	syntax	semantics
unit	sentence	document
modal	text	text, vision, speech
intelligence	"learning"	self-learning

Open Source Toolkits

- GIZA++
  - train IBM Models 1-5
  - developed by Franz Och
  - http://code.google.com/p/giza-pp/



Franz Och



giza-pp

GIZA++ statistical translation models toolkit

Project Home | Downloads | Wiki | Issues | Source

Summary | Updates | People

Project Information

Starred by 55 users

Activity Medium

Project feeds

Code license

GNU GPL v2

Labels

mt, AI, alignment, nlp, linguistics, translation, giza

GIZA++ is a statical machine translation toolkit that is used to train IBM Models 1-5 and an HMM word alignment model. This package also contains the source for the `mkcls` tool which generates the word classes necessary for training some of the alignment models.

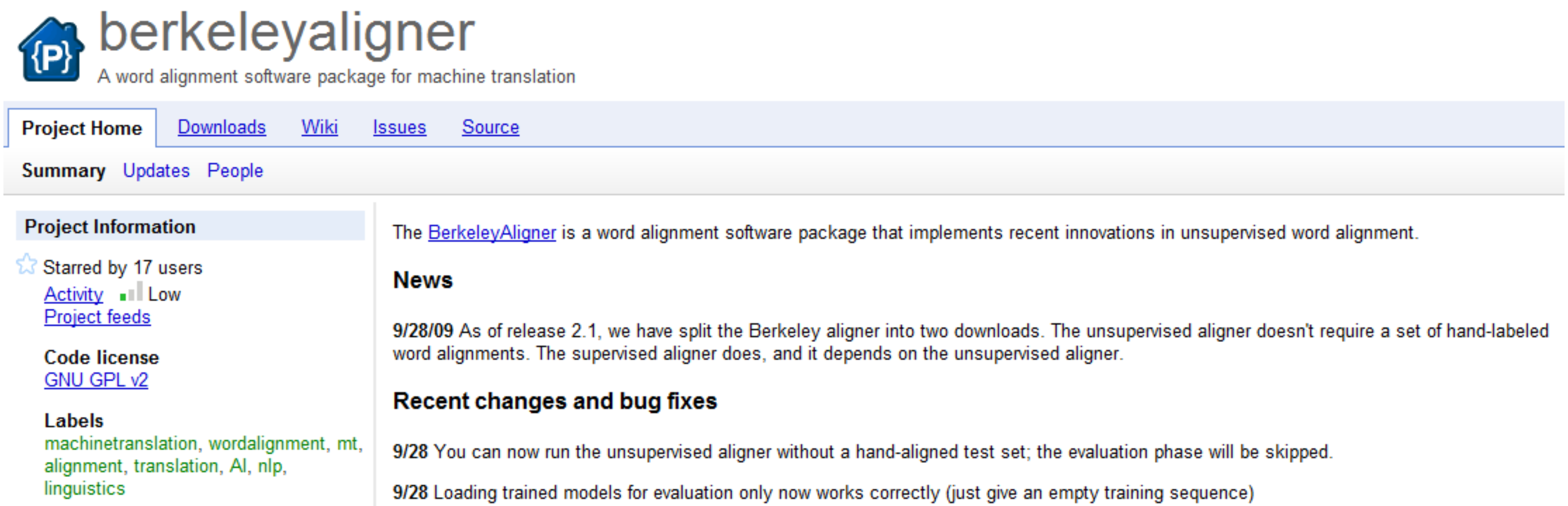
For more information on the origins of these tools, refer to http://www.statmt.org/moses/giza/GIZA++.html and http://www.statmt.org/moses/giza/mkcls.html.

If you make use of GIZA++ for research or commercial purposes, please cite:

- Franz Josef Och, Hermann Ney. "A Systematic Comparison of Various Statistical Alignment Models", *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003.

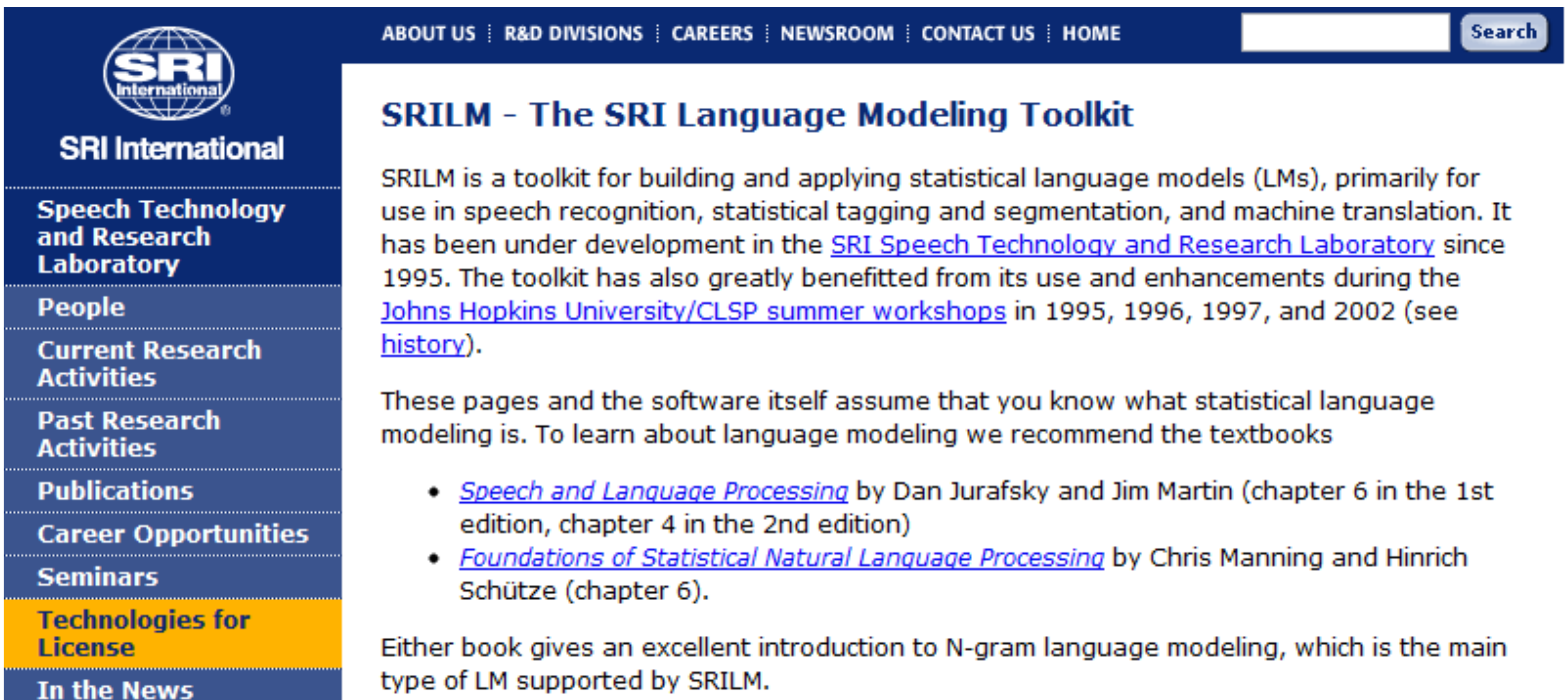
Open Source Toolkits

- Berkeley Aligner
  - train joint HMM model
  - developed by Percy Liang
  - http://code.google.com/p/berkeleyaligner/



Open Source Toolkits

- SRI Language Modeling Toolkit
  - train *n*-gram language models
  - developed by Andreas Stolcke
  - http://www.speech.sri.com/projects/srilm/


ABOUT US | R&D DIVISIONS | CAREERS | NEWSROOM | CONTACT US | HOME
Search

SRI International

Speech Technology and Research Laboratory
People
Current Research Activities
Past Research Activities
Publications
Career Opportunities
Seminars
Technologies for License
In the News

SRILM - The SRI Language Modeling Toolkit

SRILM is a toolkit for building and applying statistical language models (LMs), primarily for use in speech recognition, statistical tagging and segmentation, and machine translation. It has been under development in the SRI Speech Technology and Research Laboratory since 1995. The toolkit has also greatly benefitted from its use and enhancements during the Johns Hopkins University/CLSP summer workshops in 1995, 1996, 1997, and 2002 (see history).

These pages and the software itself assume that you know what statistical language modeling is. To learn about language modeling we recommend the textbooks

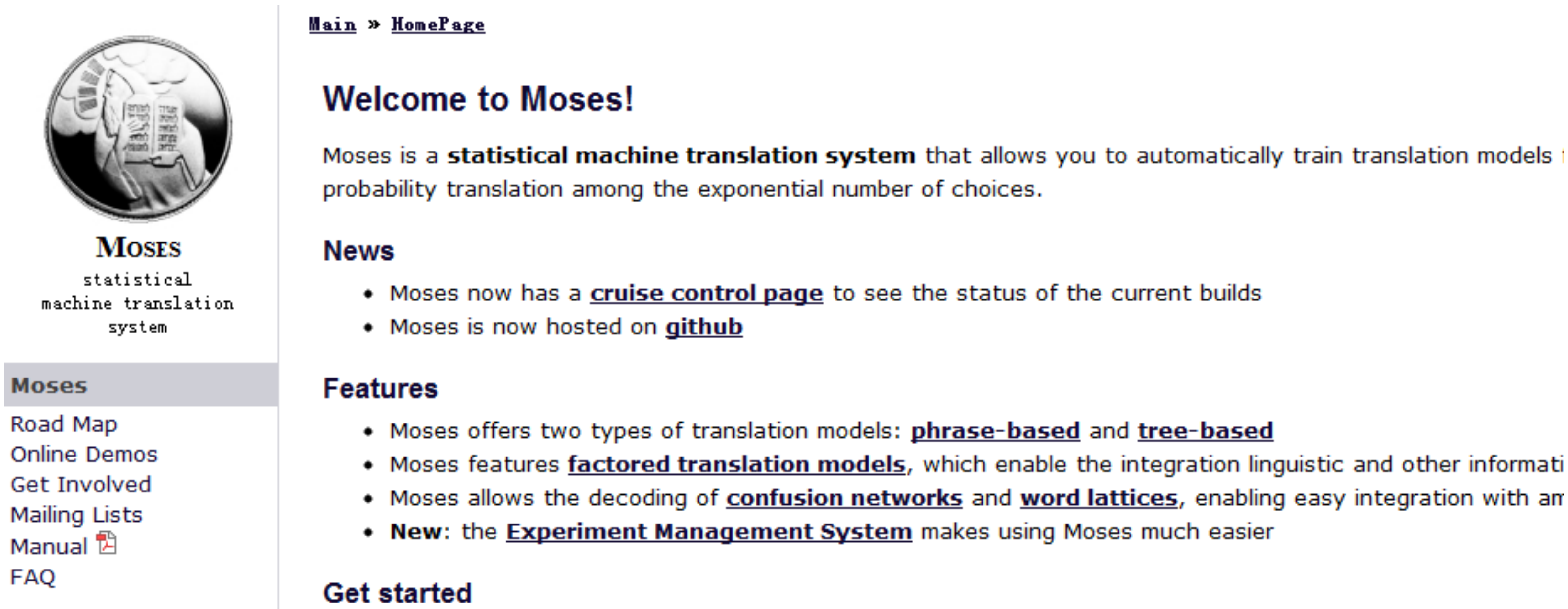
- *Speech and Language Processing* by Dan Jurafsky and Jim Martin (chapter 6 in the 1st edition, chapter 4 in the 2nd edition)
- *Foundations of Statistical Natural Language Processing* by Chris Manning and Hinrich Schütze (chapter 6).

Either book gives an excellent introduction to N-gram language modeling, which is the main type of LM supported by SRILM.

Open Source Toolkits

- Moses
  - phrase-based and tree-based systems
  - the main contributor: Philipp Koehn
  - http://www.statmt.org/moses/


Philipp Koehn



Open Source Toolkits

- Joshua
  - SCFG-based SMT system
  - the main contributor: Zhifei Li
  - http://joshua.sourceforge.net/Joshua/

Open Source Toolkits

- Phrasal
  - Phrase-based system
  - the main contributor: Michel Galley
  - http://nlp.stanford.edu/software/phrasal/

Stanford Phrasal: A Phrase-Based Translation System

About | Usage | Download | Contributors | Citation | Mailing lists |

The Beta3 release of the Stanford Phrasal open source machine translation package has just been released!

About

Stanford Phrasal is a state-of-the-art phrase-based machine translation system. It provides an easy to use API for implementating new decoding model features and supports unique capabilities such as translating using phrases that include gaps (Galley et al. 2010) and conditional extraction of phrase-tables and lexical reordering models.

Usage

- Quick Start Guide.

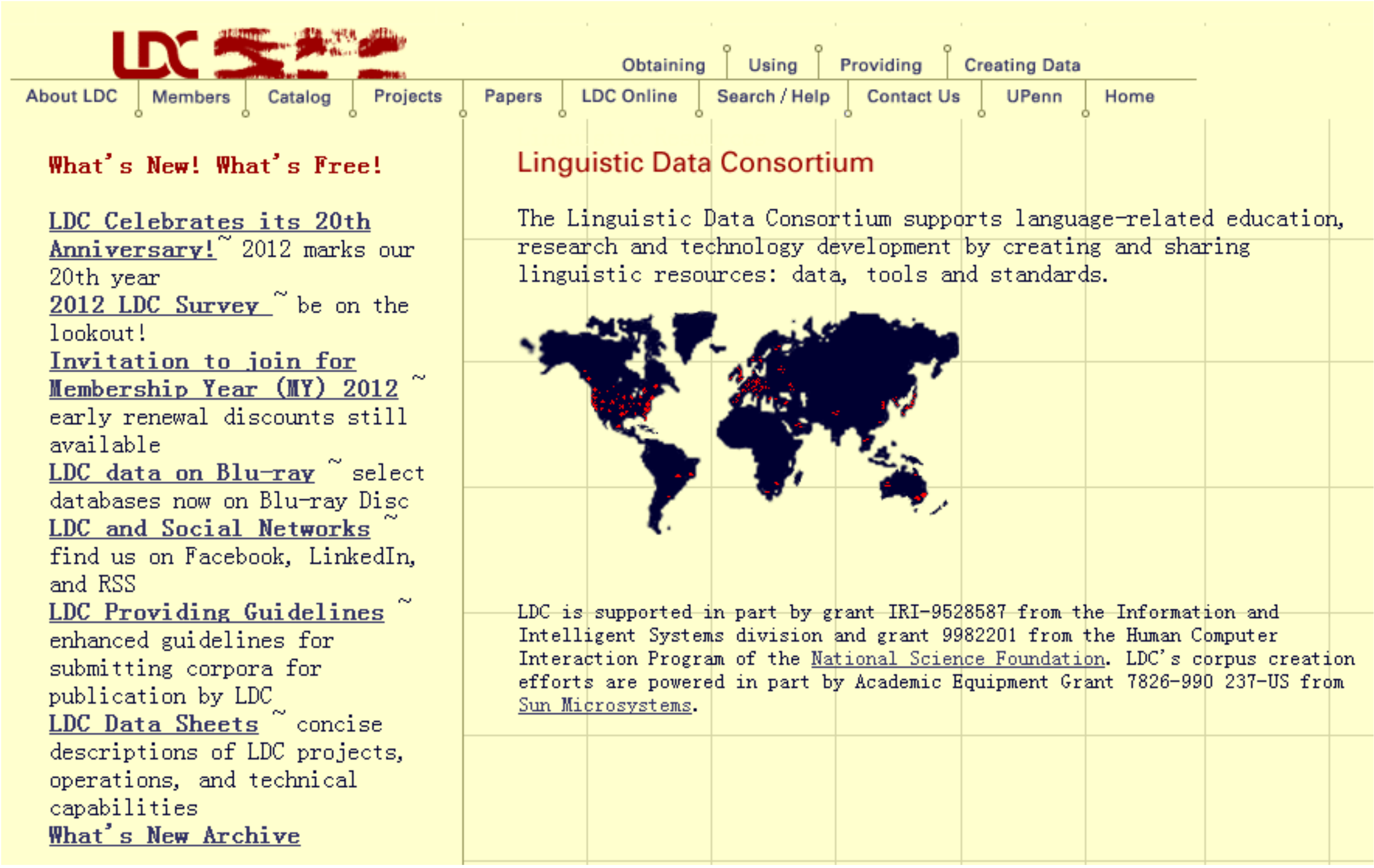
Download

Phrasal is available for download, **licensed under the GNU General Public License** (v2 or later). Source is included. The package includes components for command-line invocation, and a Java API.

Stanford Phrasal Beta3

Data Resources

- Linguistic Data Consortium
  - Major source of monolingual and bilingual corpora for SMT research
  - http://www.ldc.upenn.edu/



Data Resources

- Chinese Linguistic Data Consortium
  - Many useful monolingual and bilingual corpora for SMT research
  - http://www.chineseldc.org/

Data Resources

- Europarl
  - It is free! 10 European language pairs
  - http://www.statmt.org/europarl/

European Parliament Proceedings Parallel Corpus 1996-2009

For a detailed description of this corpus, please read:

Europarl: A Parallel Corpus for Statistical Machine Translation, *Philipp Koehn*, MT Summit 2005, pdf.

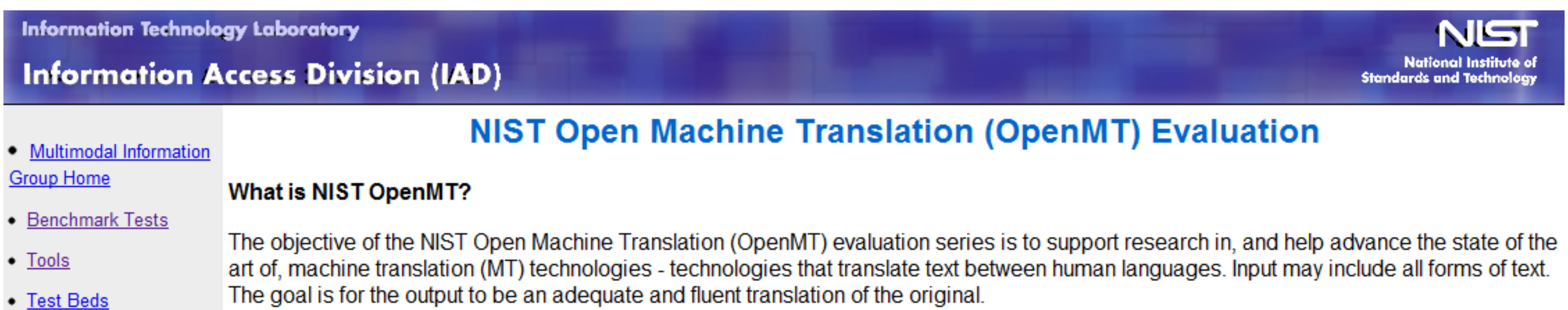
Please cite the paper, if you use this corpus in your work. See also the extended (but earlier) version of the report (ps, pdf).

The Europarl parallel corpus is extracted from the proceedings of the European Parliament. It includes versions in 11 European languages: Romanic (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish.

The goal of the extraction and processing was to generate sentence aligned text for statistical machine translation systems. For this purpose we extracted matching items and labeled them with corresponding document IDs. Using a preprocessor we identified sentence boundaries. We sentence aligned the data using a tool based on the Church and Gale algorithm.

Evaluation

- NIST
  - the most influential
  - Tasks: Arabic-English, Chinese-English
  - http://www.itl.nist.gov/iad/mig/tests/mt/



Evaluation

- IWSLT
  - spoken language translation
  - Tasks: European languages and English
  - http://iwslt2011.org/

Evaluation

- WMT
  - workshop on machine translation
  - Tasks: European languages and English
  - http://www.statmt.org/wmt11/

EMNLP 2011
SIXTH WORKSHOP ON
STATISTICAL MACHINE TRANSLATION

July 30-31, 2011
Edinburgh, UK

[HOME] | [TRANSLATION TASK] | [FEATURED TRANSLATION TASK] | [SYSTEM COMBINATION TASK] | [EVALUATION TASK]
[BASELINE SYSTEM] | [BASELINE SYSTEM 2]
[SCHEDULE] | [PAPERS] | [AUTHORS]

Evaluation

- CWMT
  - the most influential MT evaluation in China
  - Tasks: English and languages in China
  - http://nlp.ict.ac.cn/new/CWMT/index.php

全国机器翻译研讨会评测简介

2005年由中科院自动化所、计算所和厦门大学联合发起并组织了第一届统计机器翻译技术评测及学术研讨会，会议在厦门大学成功举办。随后，会议由中科院计算所、自动化所、软件所、哈尔滨工业大学和厦门大学五家单位联合组织。2006年、2007年，第二、第三届全国统计机器翻译研讨会（SSMT）分别在中科院计算所、哈尔滨工业大学成功召开。2008年，第四届会议于中科院自动化所成功举办，并由此届起会议名称更改为全国机器翻译研讨会（China Workshop on Machine Translation，简称CWMT）。2009年，第五届全国机器翻译研讨会在南京大学成功举办。2010年，由于同行们将很大精力都投入到了在北京召开的COLING 2010，没有举办大规模的全国机器翻译研讨会，而是进行了小范围的机器翻译战略研讨会，范围虽小，但也相当热烈和成功，该研讨会算是本系列会议的第六次。前六届会议的成功举办，对加强国内外同行的学术交流，促进中国机器翻译事业的发展，起到了很好的推动作用。

Journals and Conferences

- Journals
  - Computational Linguistics
  - Machine Translation
  - ACM TALIP
- Conferences
  - ACL
  - EMNLP
  - NAACL
  - COLING

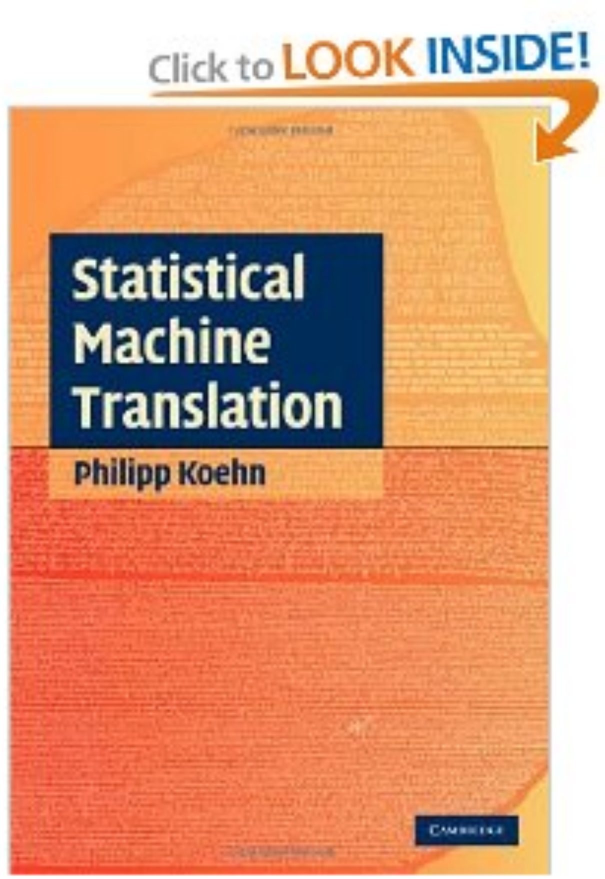
Other Useful Resources

- ACL Anthology
  - ~20,000 (free) papers in the NLP field
  - http://aclweb.org/anthology-new/
- ACL Anthology Network
  - Paper network, author network, ranking
  - http://clair.si.umich.edu/clair/anthology/index.cgi
- ACL Wiki
  - Many useful information for NLP researchers
  - http://aclweb.org/aclwiki

Other Tutorials

- Statistical Machine Translation (Adam Lopez, 2010)
- What's New in Statistical Machine Translation (Kevin Knight, 2006)
- Statistical Machine Translation: the Basic, the Novel, and the Speculative (Philipp Koehn, 2006)
- Statistical Machine Translation: Foundations and Recent Advances (Franz Och, 2005)

Book



Share your own customer images
Search inside this book

Statistical Machine Translation [Hardcover]
Philipp Koehn (Author)
(2 customer reviews) | Like (0)

List Price: ~~$67.00~~
Price: $55.57 & this item ships for FREE with **Super Saver Shipping**. Details
You Save: $11.43 (17%)

In Stock.
Ships from and sold by **Amazon.com**. Gift-wrap available.

Only 6 left in stock--order soon (more on the way).

Want it delivered Wednesday, February 8? Order it in the next 16 hours and 45 minutes,

20 new from $55.56 **13 used** from $42.99

FREE Two-Day Shipping for students on millions of items. Learn more

Book Trade-In

Sell Back Your Copy for $28.44
Whether you buy it used on Amazon for $42.99 or somewhere else, you can
price of $28.44.

Conclusions

- Statistical machine translation learns translation knowledge from data
- Big data makes more training instances available to SMT
- SMT evolves from word-based to phrase-based and syntax-based models
- We look forward to more intelligent MT systems

Thanks

http://nlp.csai.tsinghua.edu.cn/~ly/

References

- Sylvie Billot and Bernard Lang. 1989. The structure of shared forests in ambiguous parsing. In *Proceedings of ACL1989*.
- Peter F. Brown, Stephan A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*.
- M. Carpuat and D. Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *Proceedings of ACL 2005*.
- M. Carpuat and D. Wu. 2007. Improving Statistical Machine Translation using Word Sense Disambiguation. In *Proceedings of EMNLP 2007*.
- Y. Chan, H. Ng, D. Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of ACL 2007*.

References

- David Chiang, 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL 2005*.
- David Chiang, 2007. Hierarchical phrase-based translation. *Computational Linguistics*.
- David Chiang, 2010. Learning to translate with source and target syntax. In *Proceedings of ACL 2010*.
- Jason Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In *Proceedings of ACL 2003*.
- Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of HLT-NAACL 2004*.

References

- Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of COLING-ACL 2006*.
- Michel Galley and Chris Manning. 2010. Accurate Non-Hierarchical Phrase-based Translation. In *Proceedings of NAACL 2010*.
- Michel Galley and Chirs Manning. 2008. A Simple and Effective Hierarchical Reordering Model. In *Proceedings of EMNLP 2008.*
- Liang Huang, Kevin Knight, and Aravind Joshi. 2006. Statistical syntax-directed translation with extended domain of locality. In *Proceedings of AMTA 2006*.
- Liang Huang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proceedings of ACL 2007*.
- B. Jones, J. Andreas, D. Bauer, K. Hermann, and K. Knight. 2012. Semantic-based machine translation with hyperedge replacement grammars. In *Proceedings of COLING 2012*.

References

- Nal Kalchbrenner and Phil Blunsom. 2013. Two Recurrent Continuous Translation Models. In *Proceedings of EMNLP 2013*.
- Kevin Knight and Jonathan Graehl. 2005. An overview of probabilistic tree transducers for natural language processing. In *Proceedings of CICLing 2005*.
- Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation In *Proceedings of HLT-NAACL 2003*.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine transla- tion. In *Proceedings of ACL 2007*.
- Peng Li, Yang Liu, and Maosong Sun. 2013. Recursive Autoencoders for ITG-based Translation. In *Proceedings of EMNLP 2013*.

References

- Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of COLING-ACL 2006*.
- Yang Liu, Yajuan Lu, and Qun Liu. 2009a. Improving tree-to-tree translation with packed forests. In *Proceedings of ACL-IJCNLP 2009*.
- D. Liu and D. Gildea. 2010. Semantic role features for machine translation. In *Proceedings of COLING 2010*.
- Yang Liu and Qun Liu. 2010. Joint parsing and translation. In *Proceedings of COLING 2010*.
- Lemao Liu, Taro Watanabe, Eiichiro Sumita, and Tiejun Zhao. 2012. Additive Neural Networks for Statistical Machine Translation. In *Proceedings of ACL 2013*.

References

- Haitao Mi, Liang Huang, and Qun Liu. 2008. Forest-based translation. In *Proceedings of ACL-HLT 2008*.
- Tomas Mikolov, Wen-tau Yih, and Geoffey Zweig. 2013. Linguistic regularities in continuous vector space word representations. In *Proceedings of NAACL 2013*.
- Franz J. Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL 2002*.
- Franz Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*.
- Franz J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL 2003*.
- Libin Shen, Jinxi Xu, and Ralph Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of ACL-HLT 2008.*

References

- Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*.
- Deyi Xiong, Qun Liu, and Shouxun Lin. 2006. Maximum entropy based phrase reordering model for statistical machine translation. In *Proceedings of ACL 2006*.
- Ddeyi Xiong, M. Zhang, and H. Li. 2012. Modeling the translation of predicate-argument structure for SMT. In *Proceedings of ACL 2012*.
- Nan Yang, Shujie Liu, Mu Li, Ming Zhou, and Nenghai Yu. 2013. Word Alignment Modeling with Context Dependent Deep Neural Network. In *Proceedings of ACL 2013*.
- Min Zhang, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lin Tan, and Sheng Li. 2008. A tree sequence alignment-based tree-to-tree translation model. In *Proceedings of ACL-HLT 2008*.