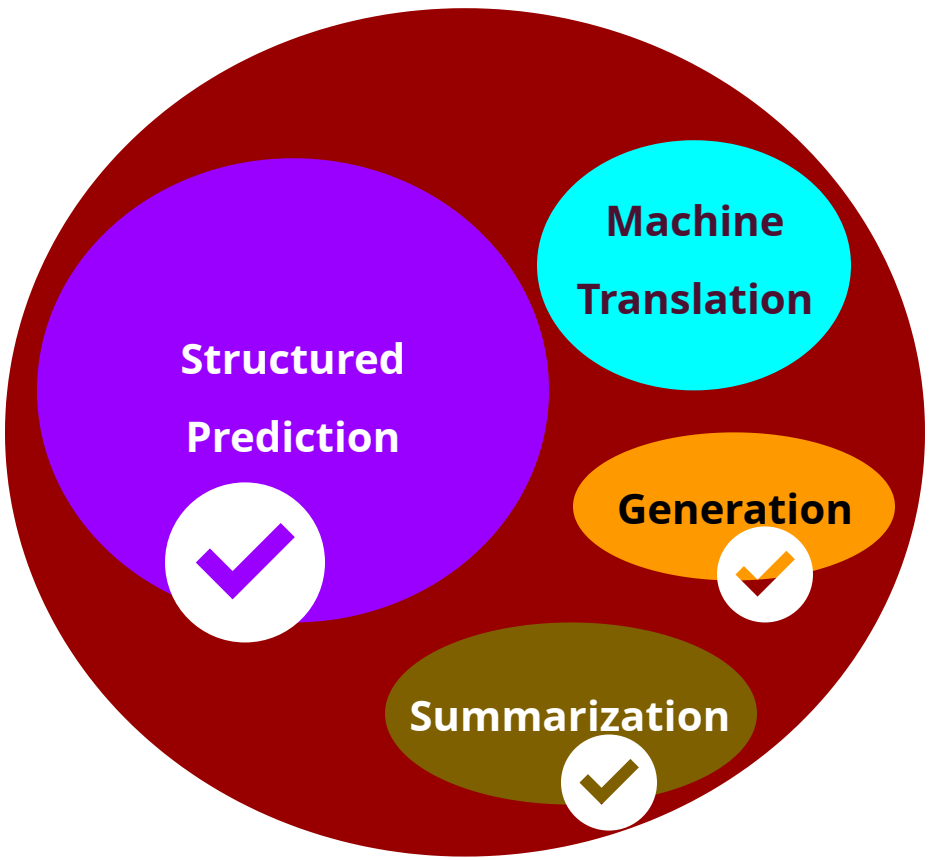
50.040
Natural Language Processing

Lu, Wei

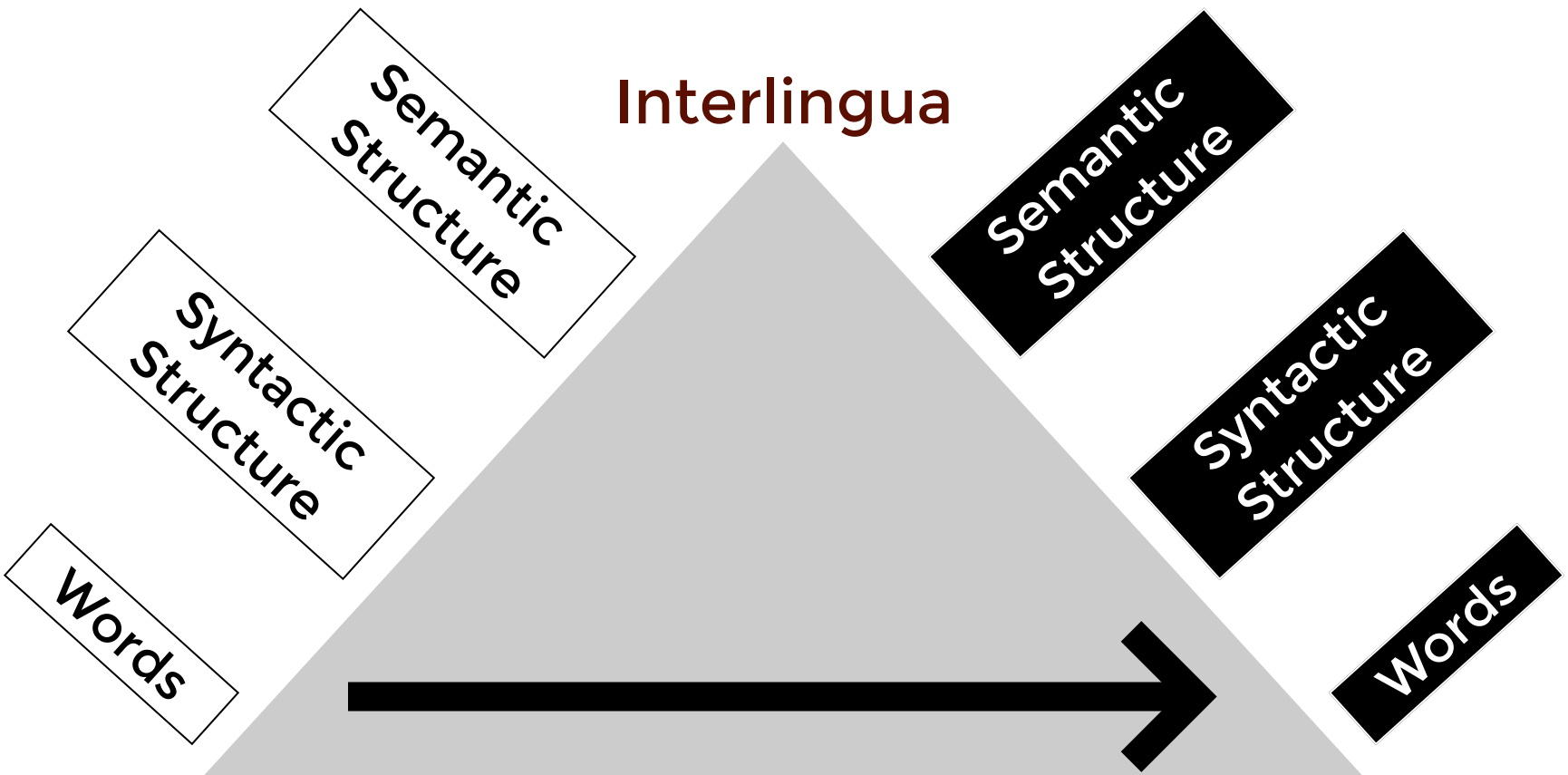
SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Tasks in NLP



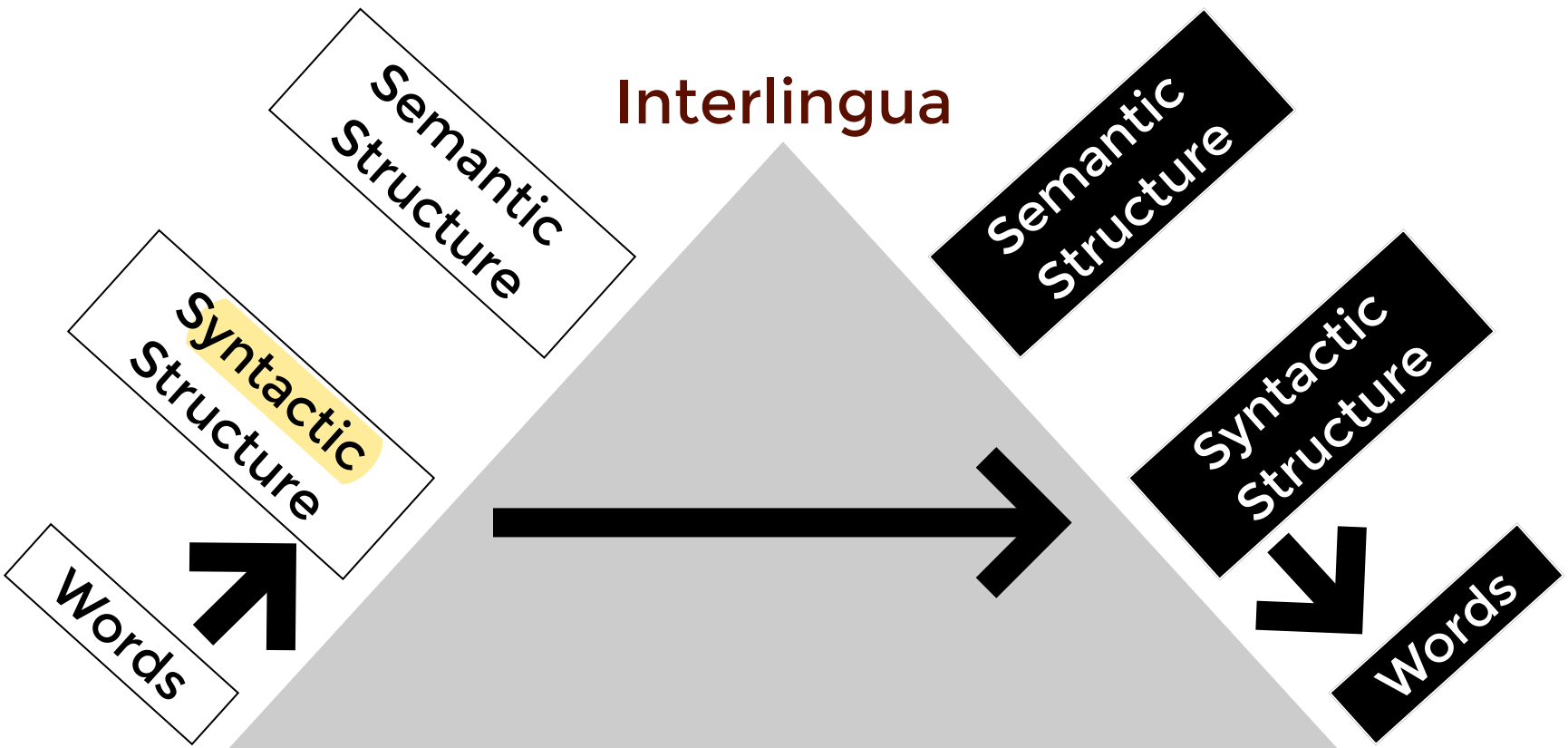
Supervised

Machine Translation

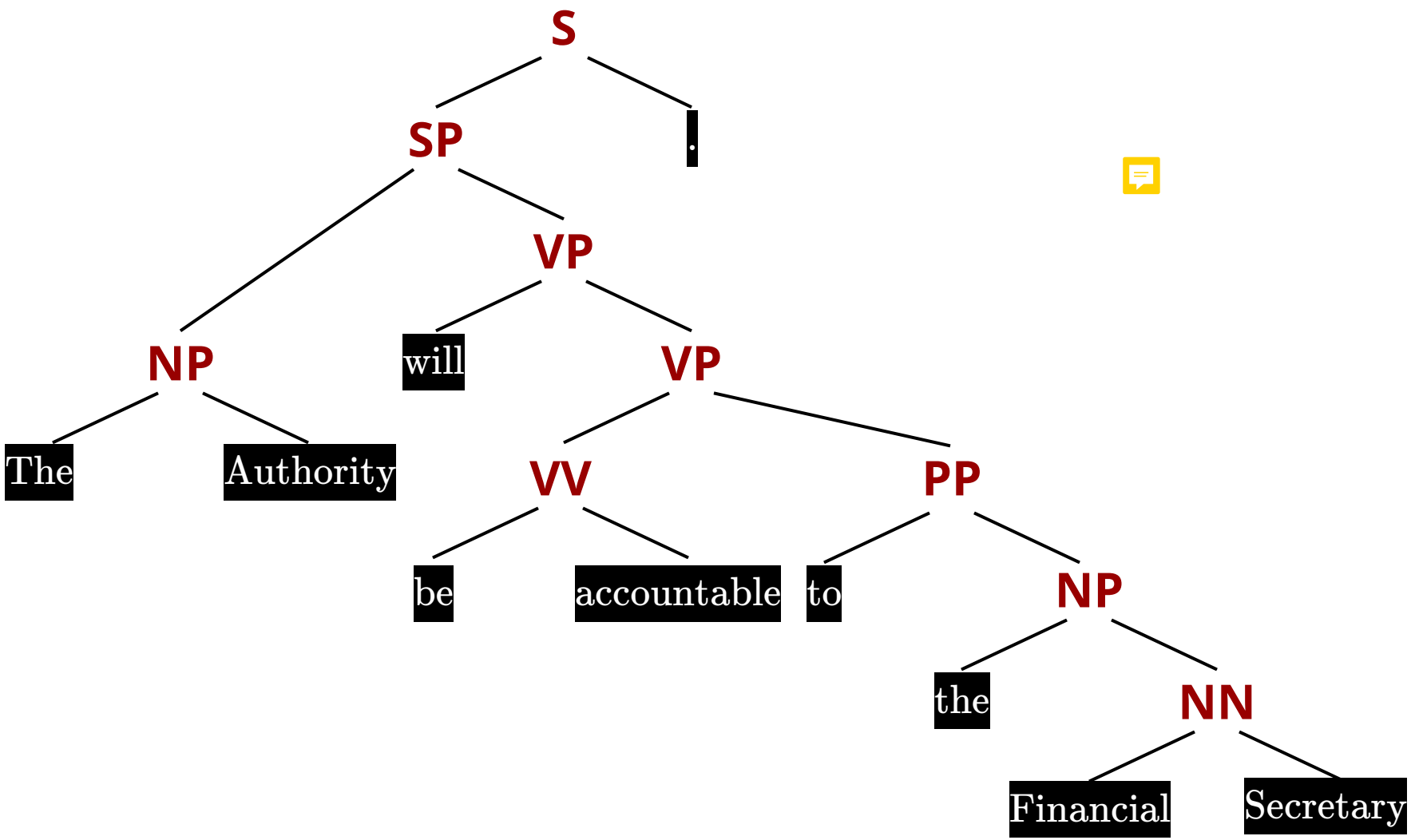


Text-to-text Problem

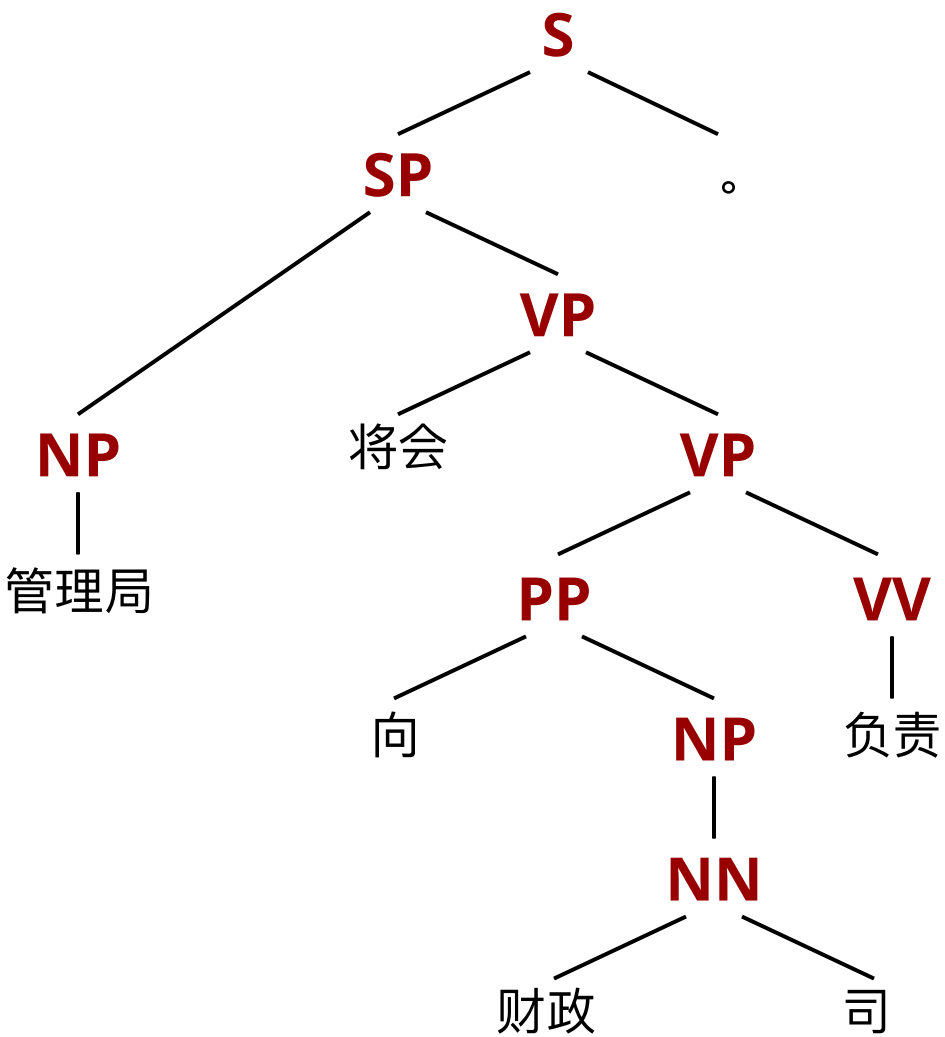
Machine Translation



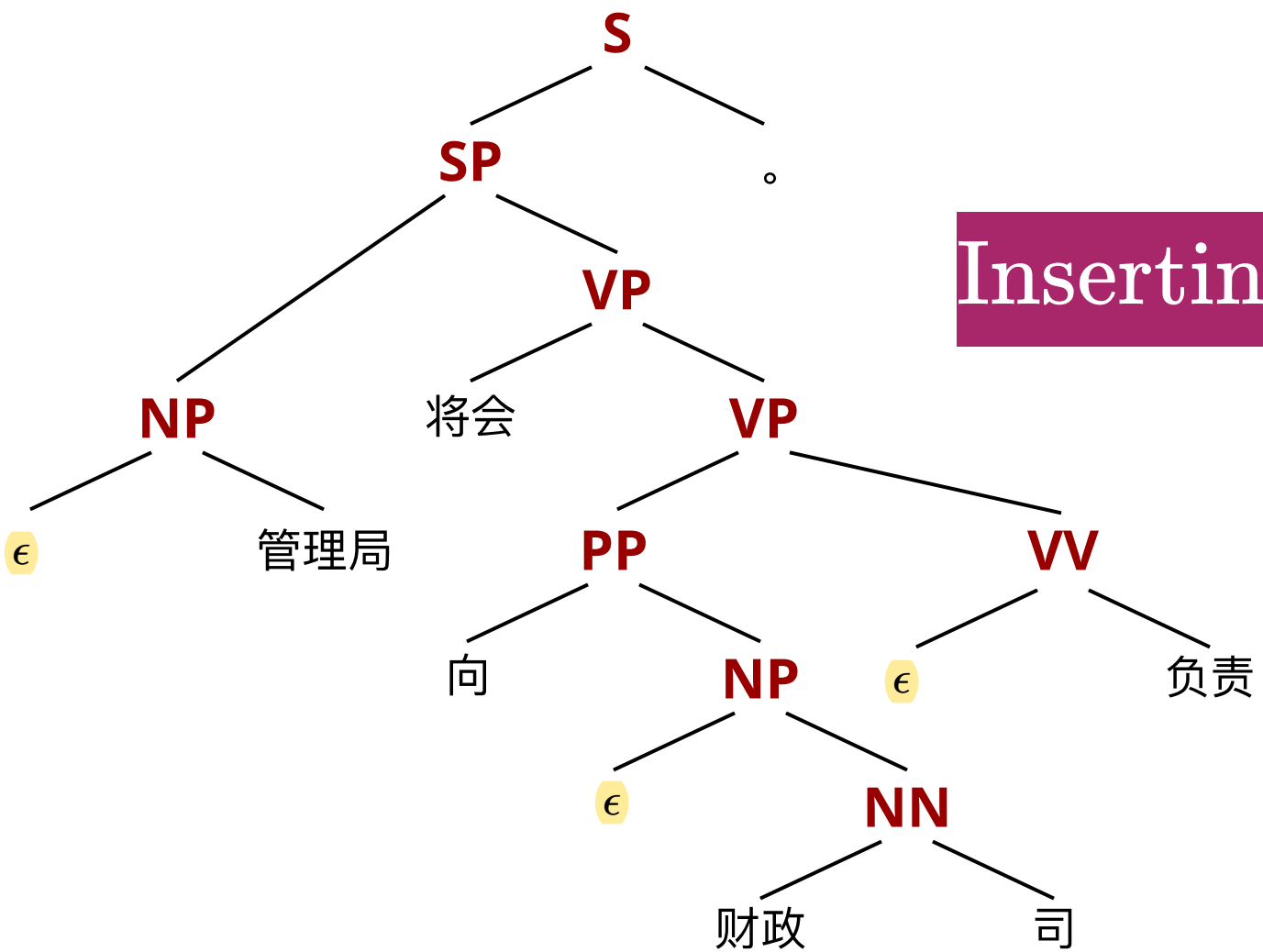
Target Parse Tree



Source Parse Tree

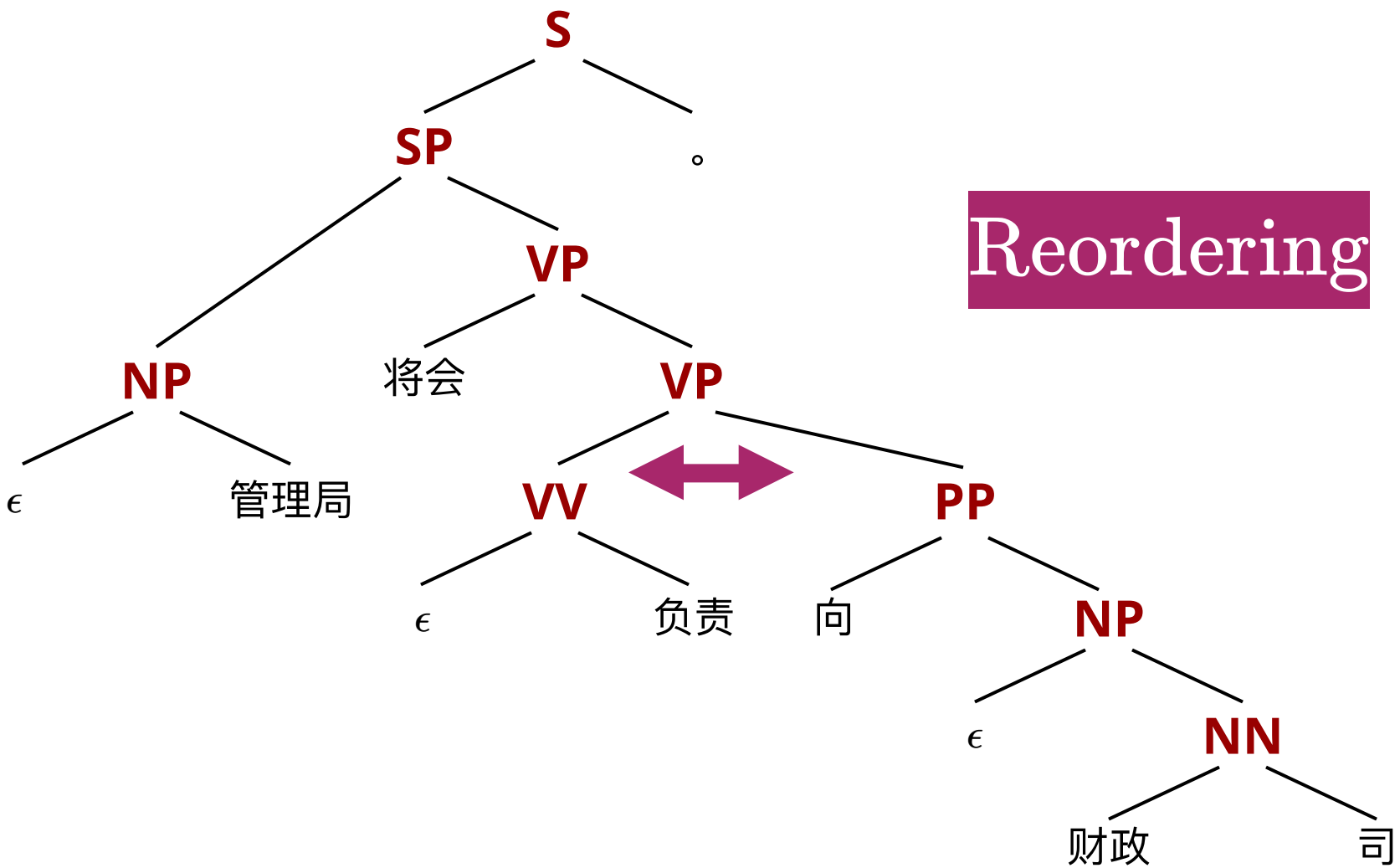


Modified Source Tree

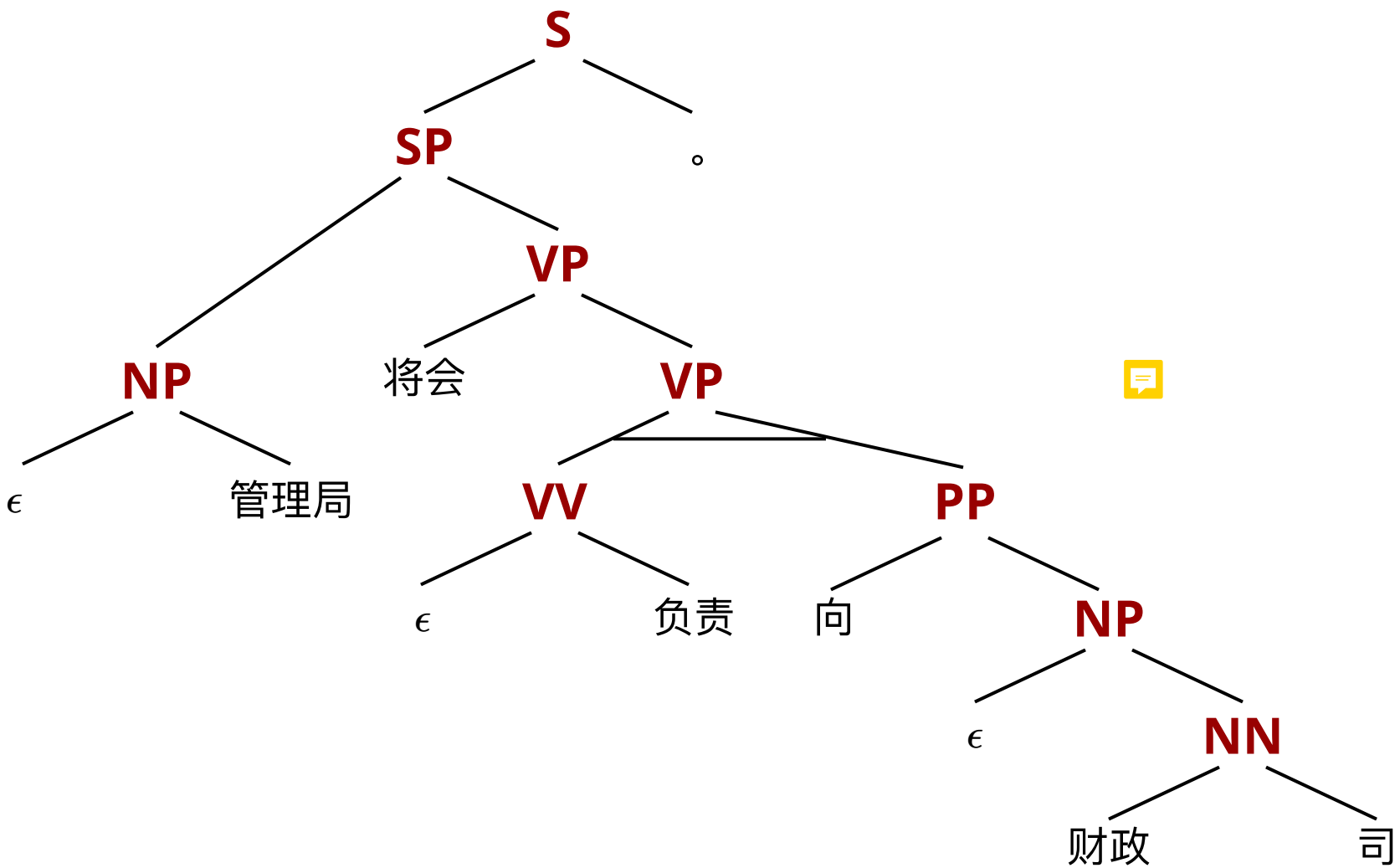


Inserting ϵ nodes

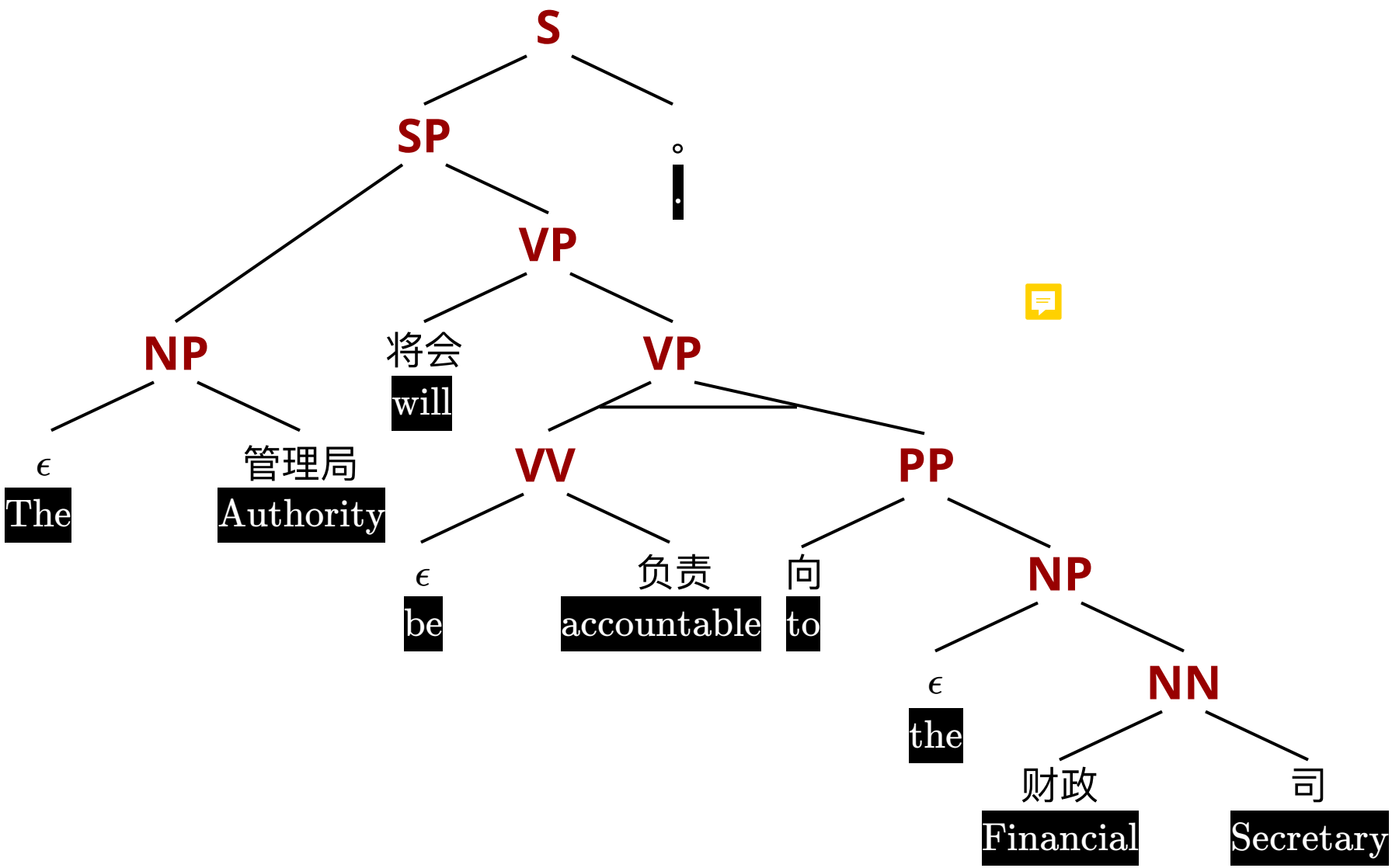
Modified Source Tree



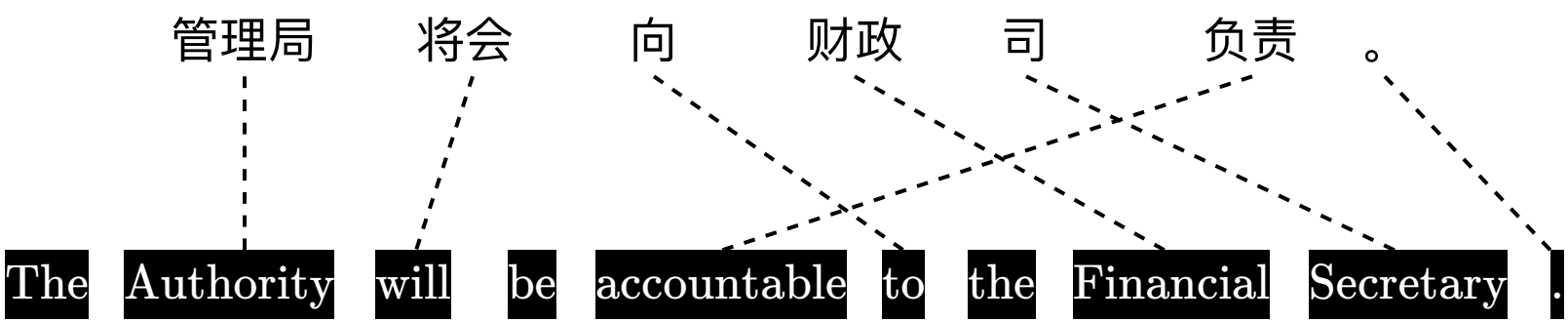
Modified Source Tree



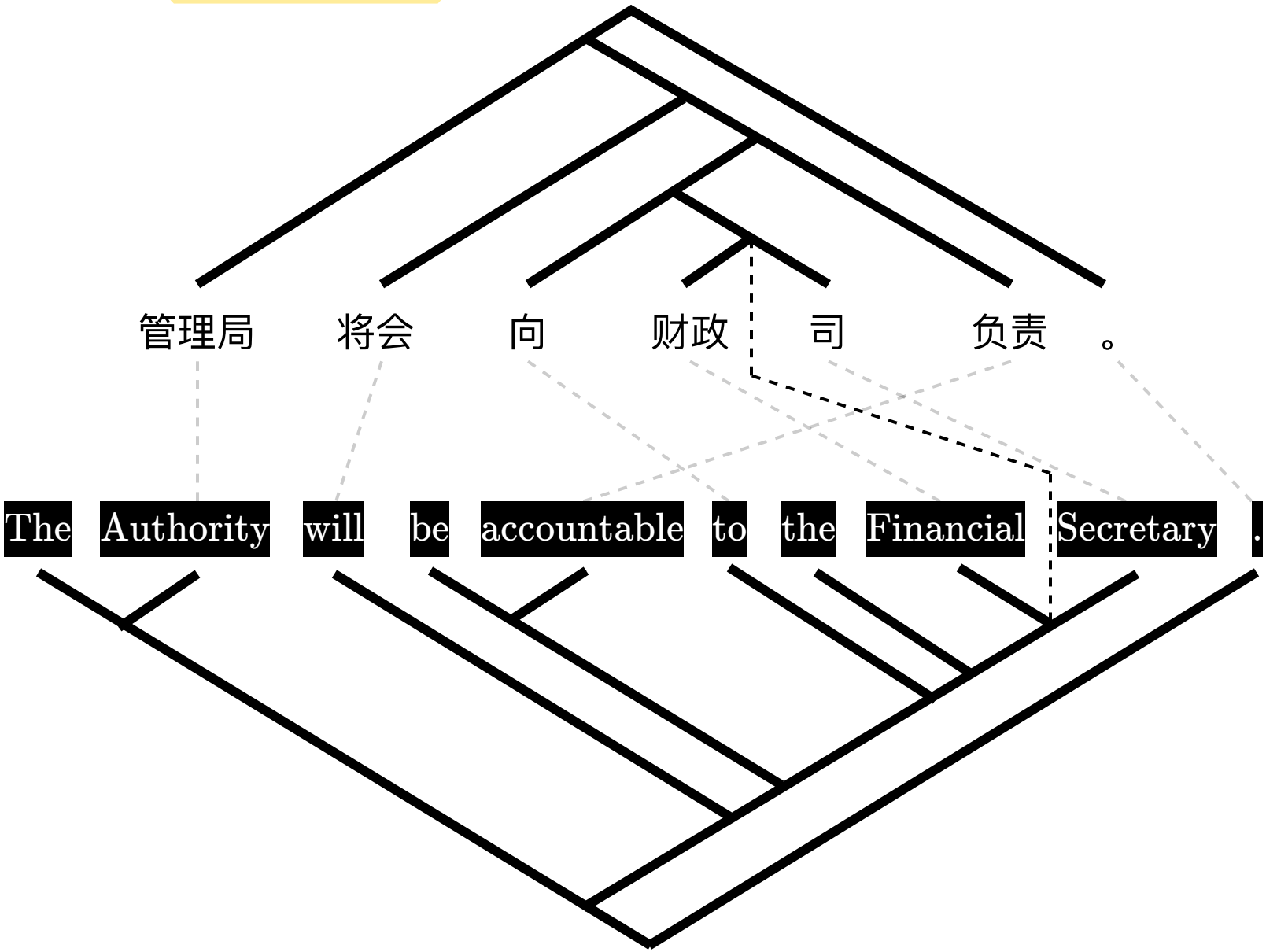
Synchronous Tree



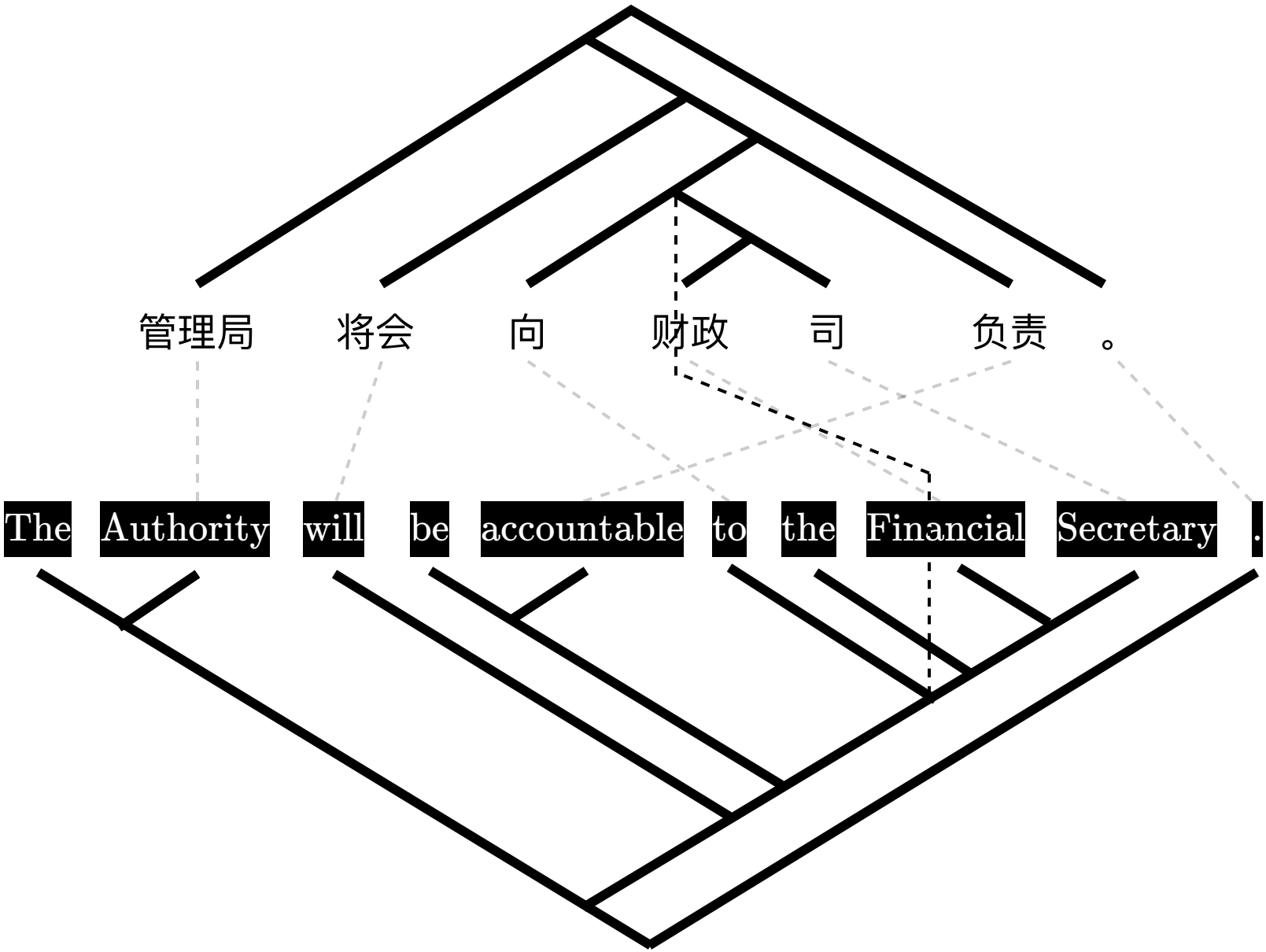
Word Alignment



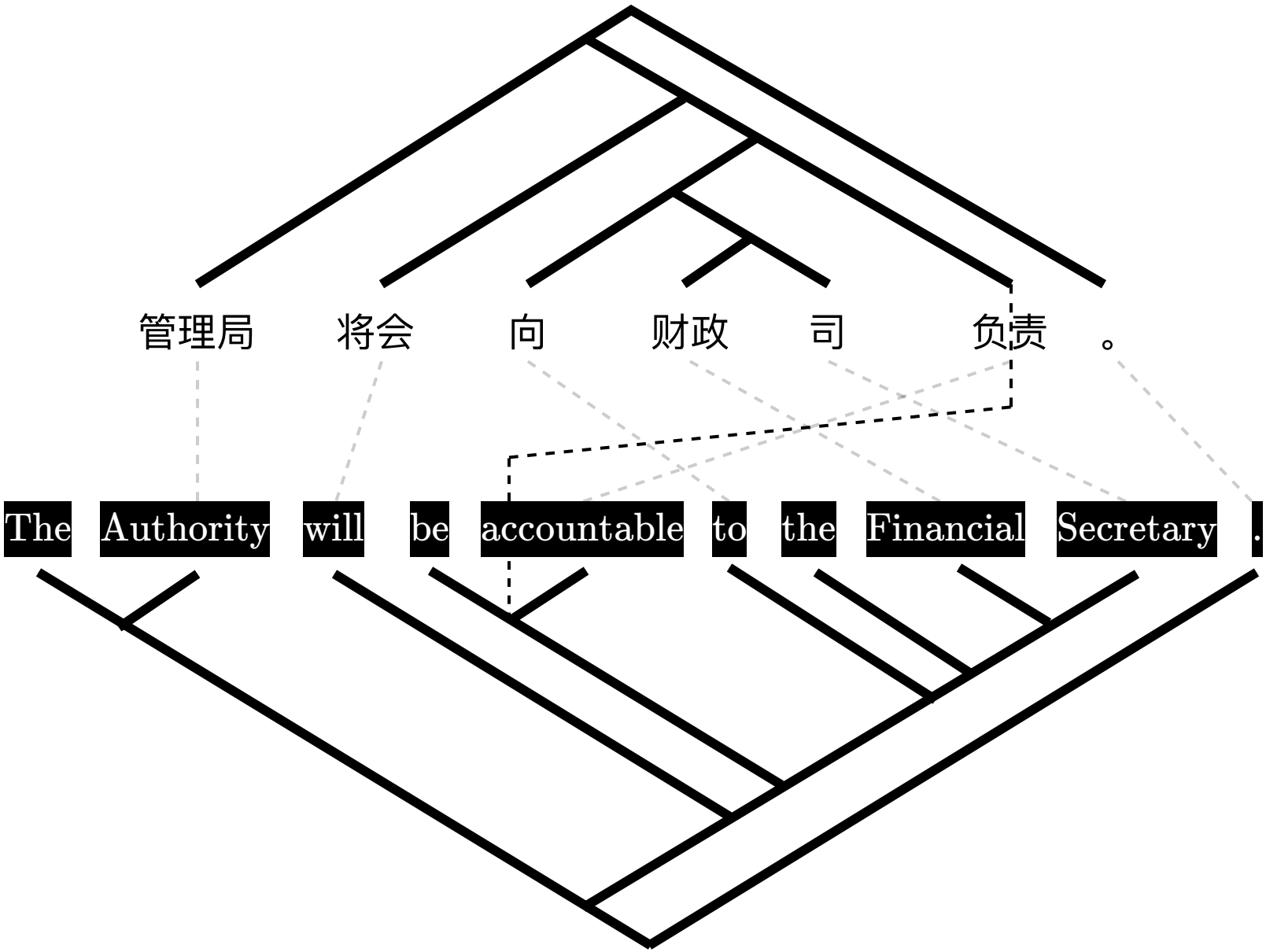
Tree Alignment



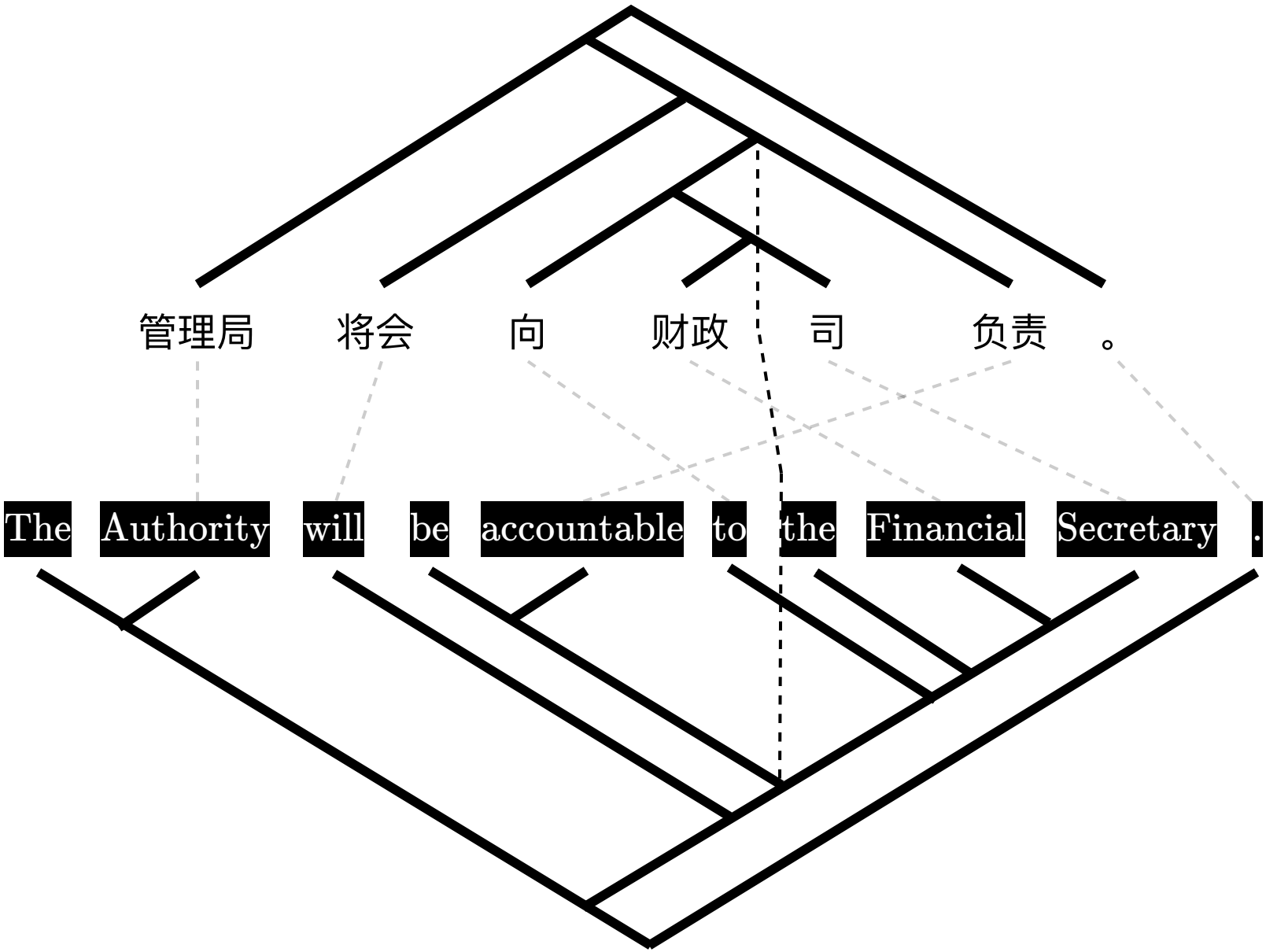
Tree Alignment



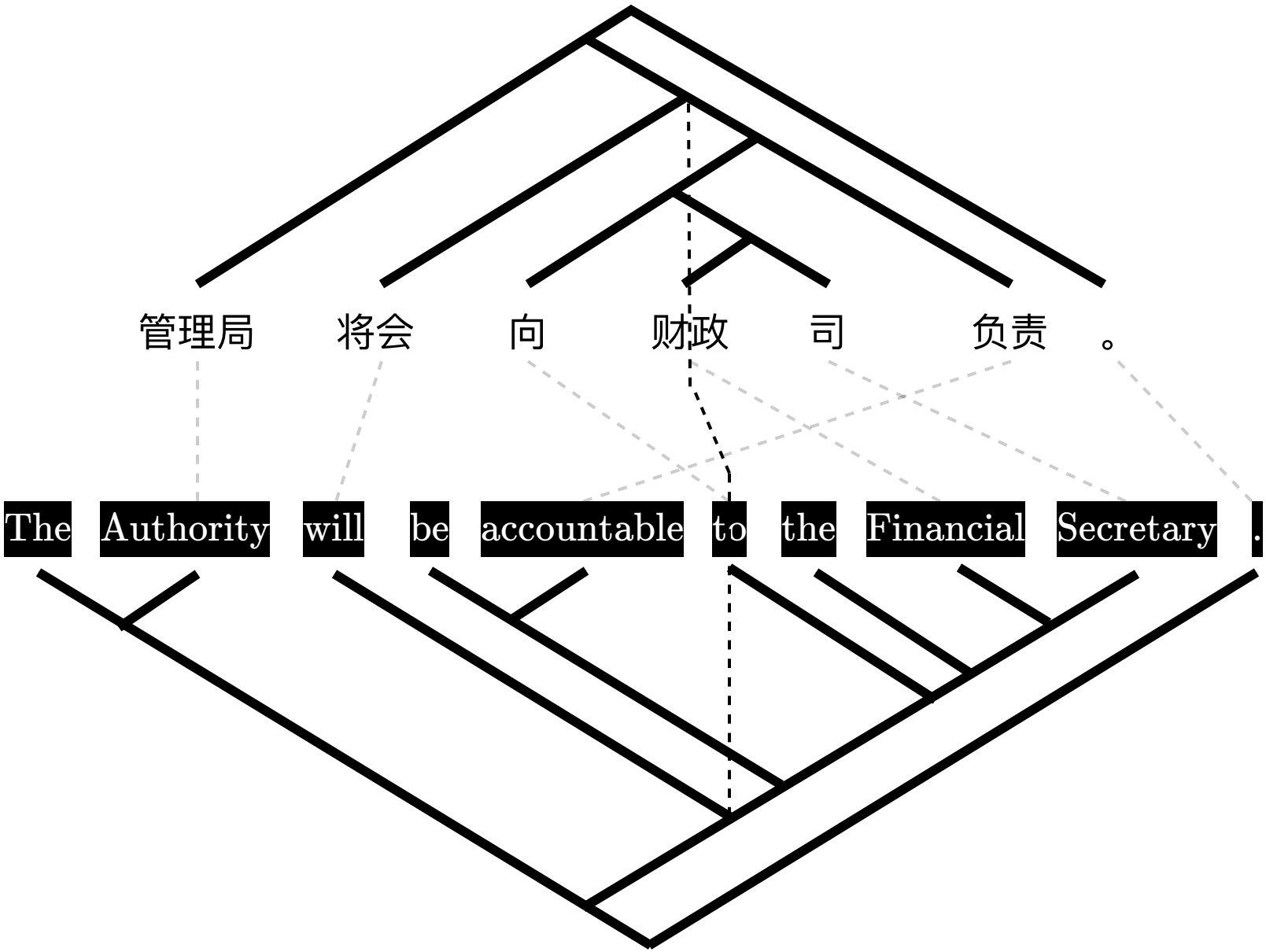
Tree Alignment



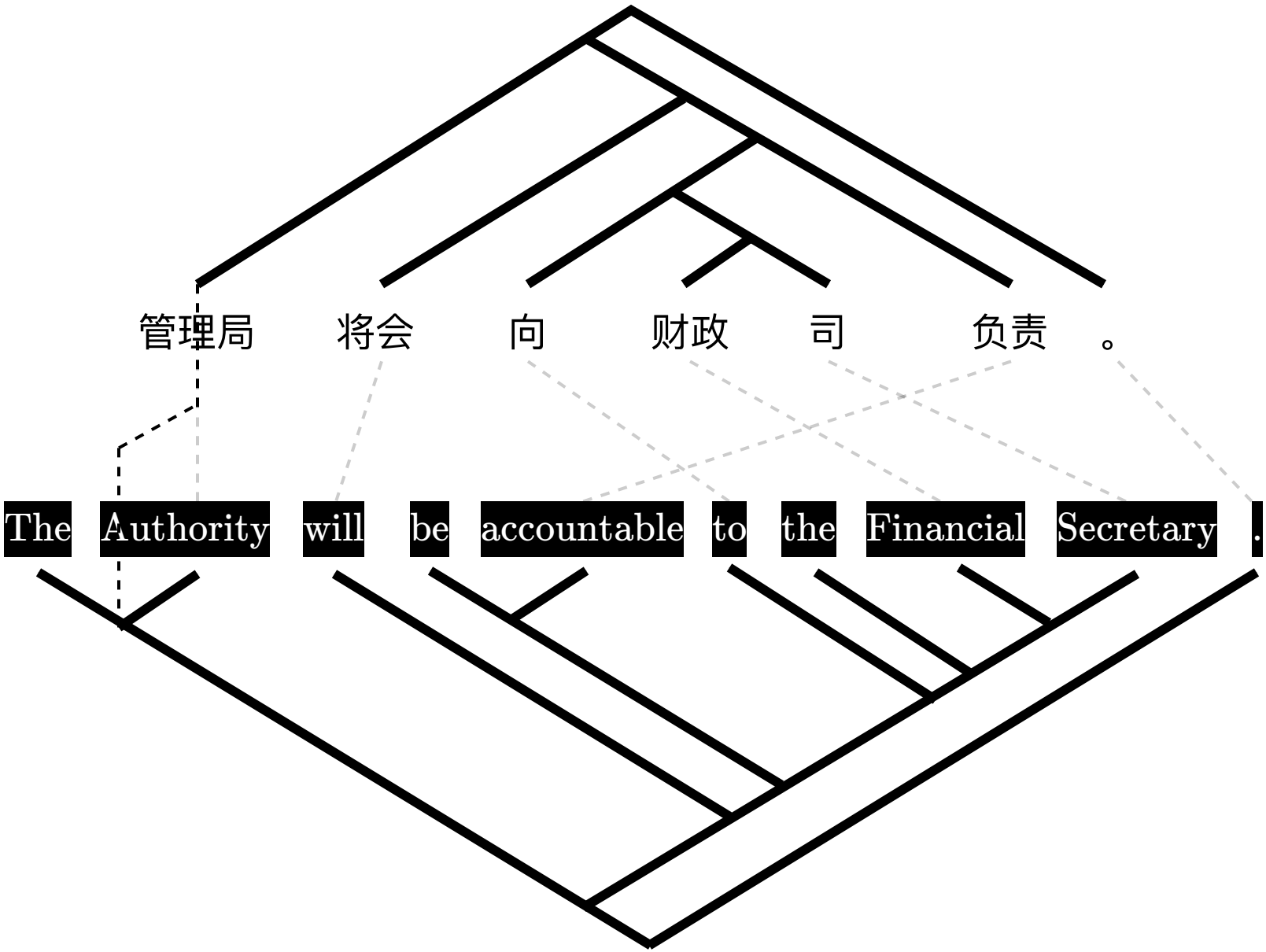
Tree Alignment



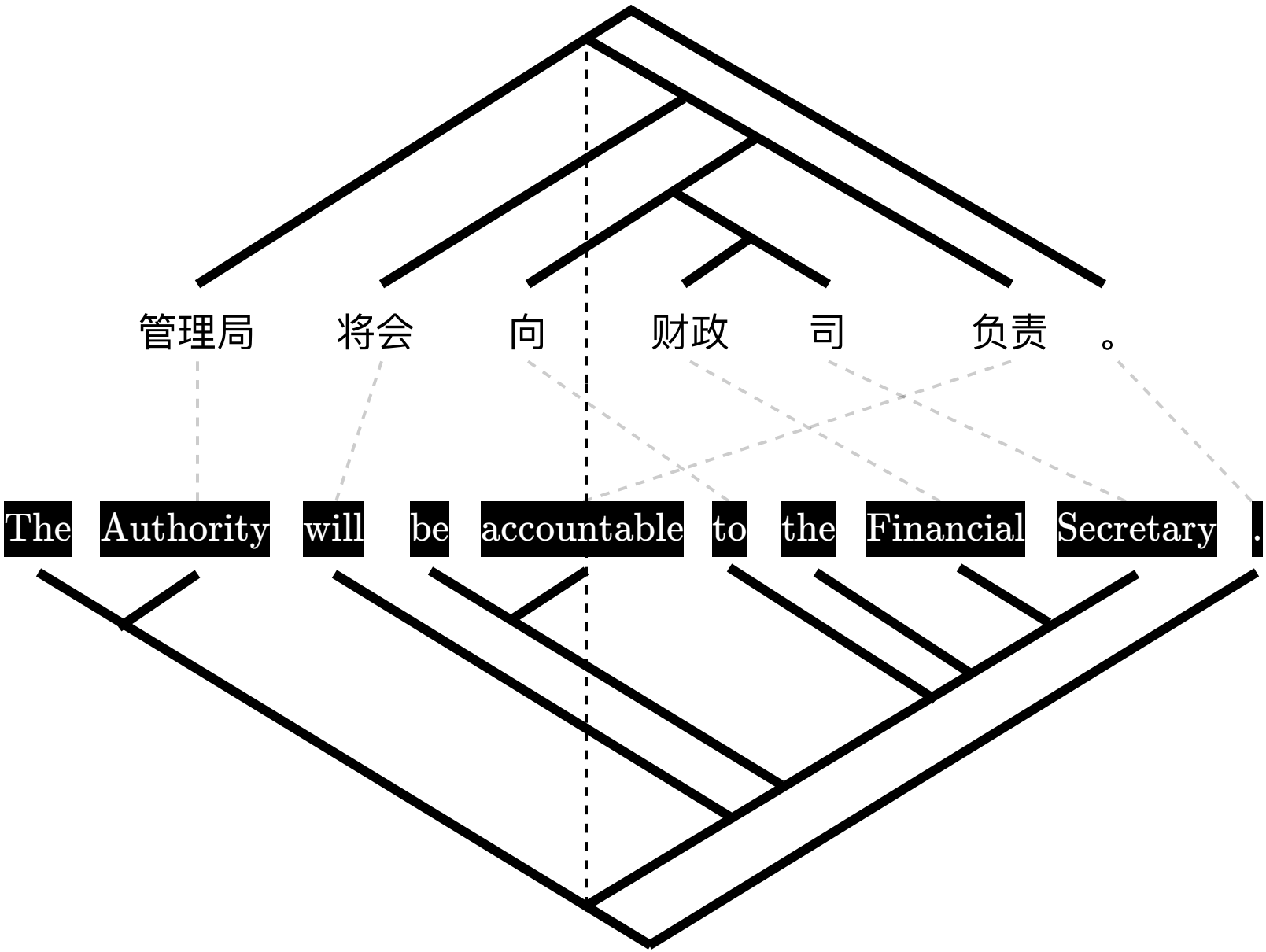
Tree Alignment



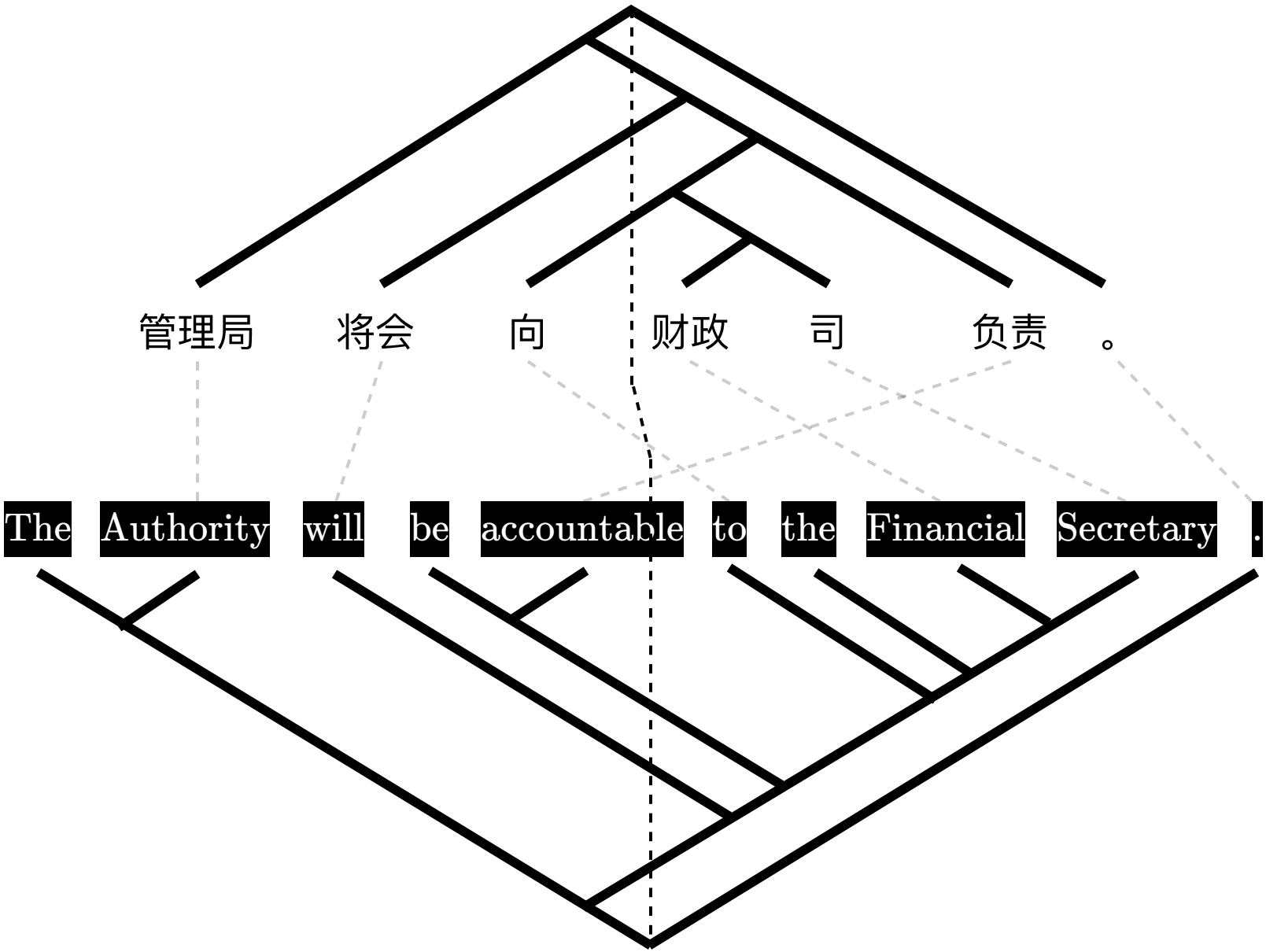
Tree Alignment



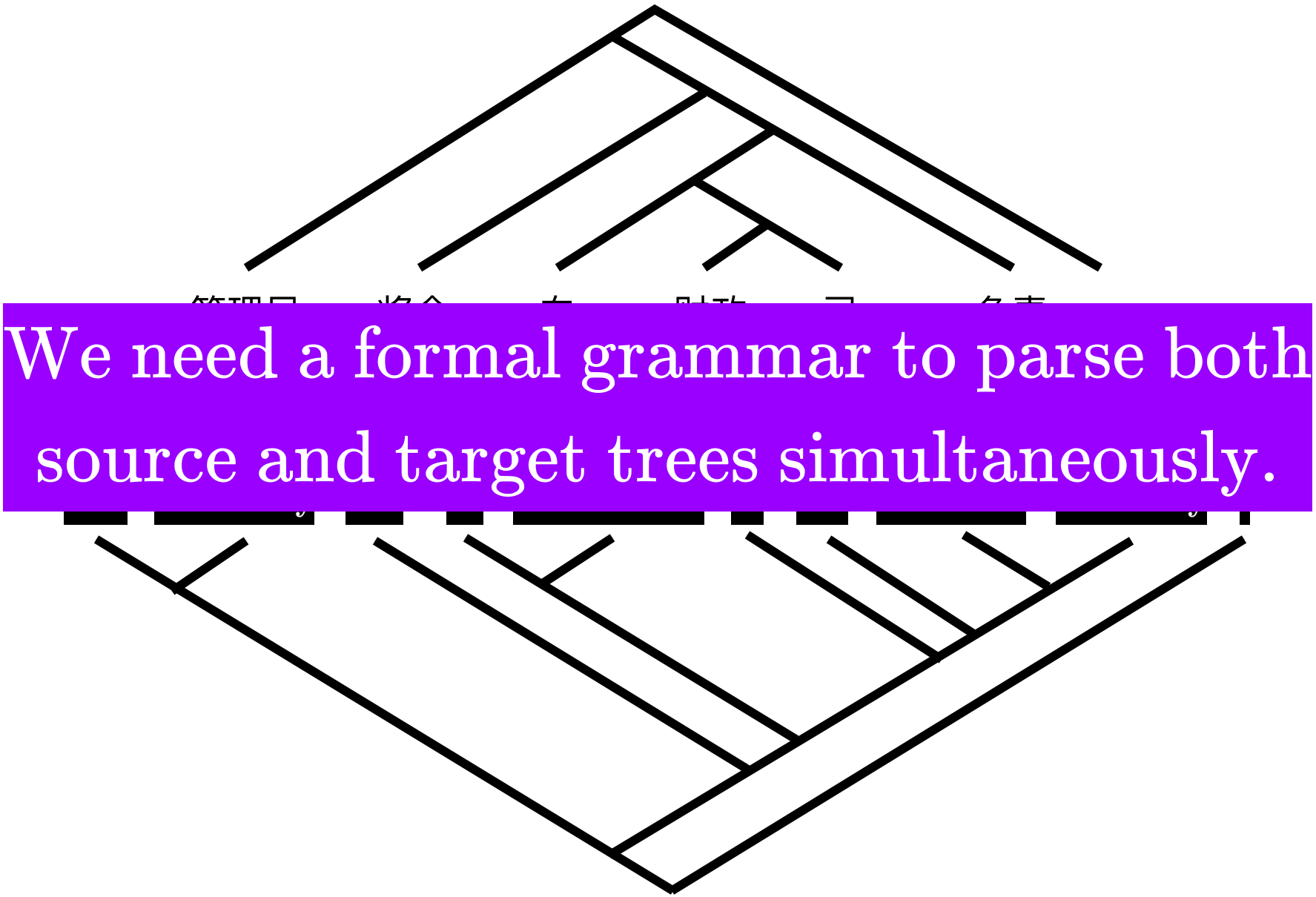
Tree Alignment



Tree Alignment

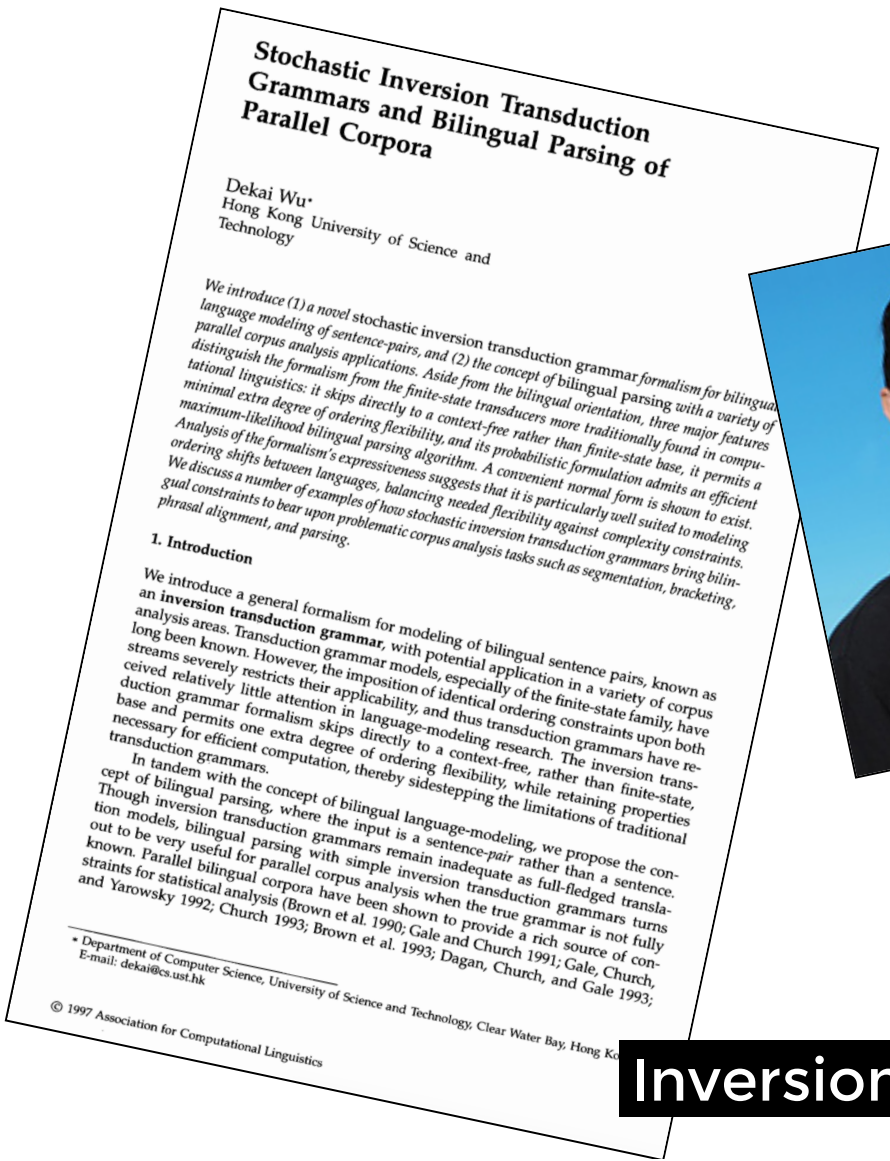


Tree Alignment



We need a formal grammar to parse both source and target trees simultaneously.

Synchronous Grammar



Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora

Dekai Wu*
Hong Kong University of Science and Technology

We introduce (1) a novel stochastic inversion transduction grammar *formalism for bilingual language modeling of sentence-pairs, and (2) the concept of* bilingual parsing *with a variety of parallel corpus analysis applications. Aside from the bilingual orientation, three major features distinguish the formalism from the finite-state transducers more traditionally found in computational linguistics: it skips directly to a context-free rather than finite-state base, it permits a minimal extra degree of ordering flexibility, and its probabilistic formulation admits an efficient maximum-likelihood bilingual parsing algorithm. A convenient normal form is shown to exist. Analysis of the formalism's expressiveness suggests that it is particularly well suited to modeling ordering shifts between languages, balancing needed flexibility against complexity constraints. We discuss a number of examples of how stochastic inversion transduction grammars bring bilingual constraints to bear upon problematic corpus analysis tasks such as segmentation, bracketing, phrasal alignment, and parsing.*

1. Introduction

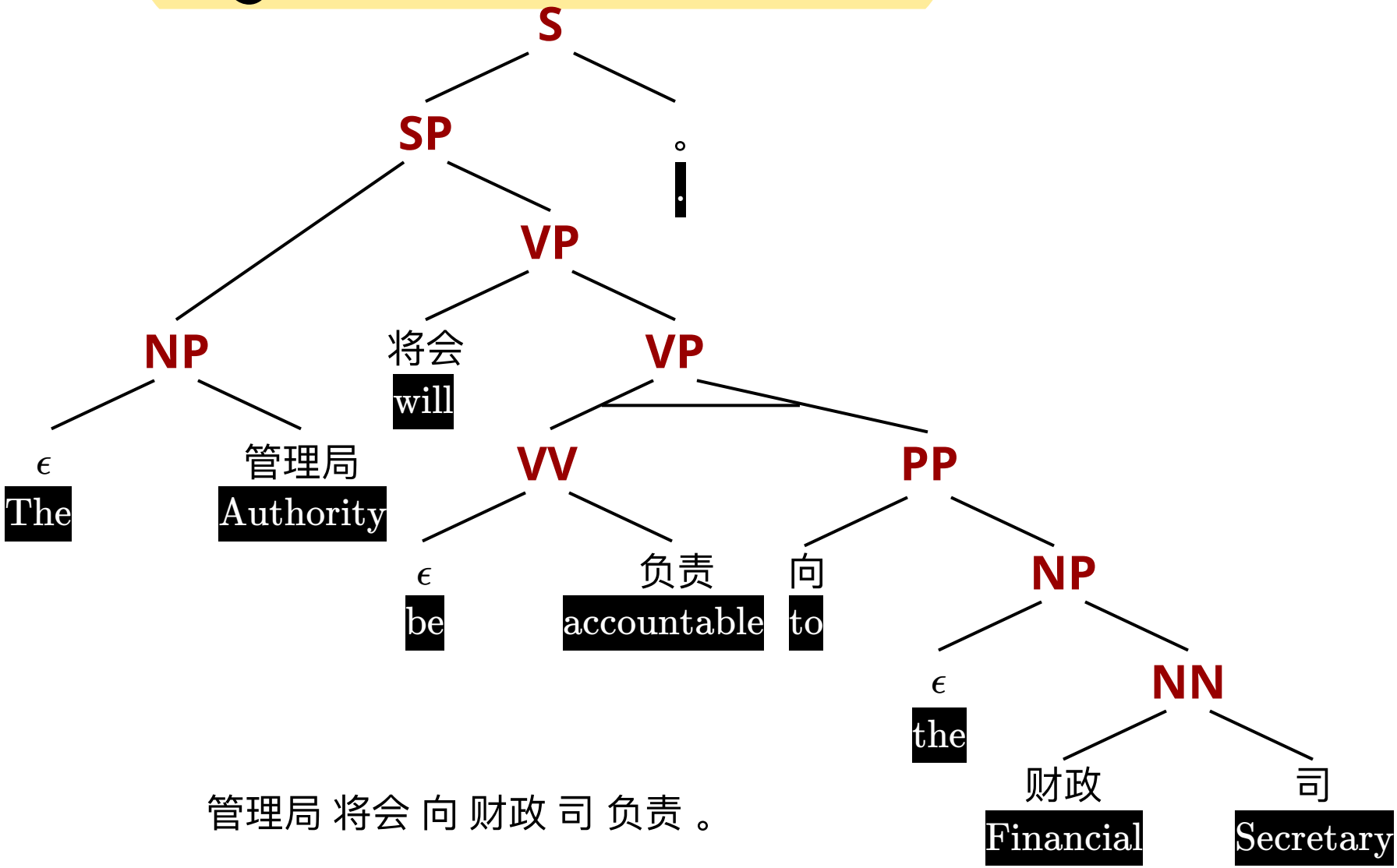
We introduce a general formalism for modeling of bilingual sentence pairs, known as an **inversion transduction grammar,** with potential application in a variety of corpus analysis areas. Transduction grammar models, especially of the finite-state family, have long been known. However, the imposition of identical ordering constraints upon both streams severely restricts their applicability, and thus transduction grammars have received relatively little attention in language-modeling research. The inversion transduction grammar formalism skips directly to a context-free, rather than finite-state, base and permits one extra degree of ordering flexibility, while retaining properties necessary for efficient computation, thereby sidestepping the limitations of traditional transduction grammars.

In tandem with the concept of bilingual language-modeling, we propose the concept of bilingual parsing, where the input is a sentence-*pair* rather than a sentence. Though inversion transduction grammars remain inadequate as full-fledged translation models, bilingual parsing with simple inversion transduction grammars turns out to be very useful for parallel corpus analysis when the true grammar is not fully known. Parallel bilingual corpora have been shown to provide a rich source of constraints for statistical analysis (Brown et al. 1990; Gale and Church 1991; Gale, Church, and Yarowsky 1992; Church 1993; Brown et al. 1993; Dagan, Church, and Gale 1993;

* Department of Computer Science, University of Science and Technology, Clear Water Bay, Hong Kong
E-mail: dekai@cs.ust.hk

Inversion Transduction Grammar

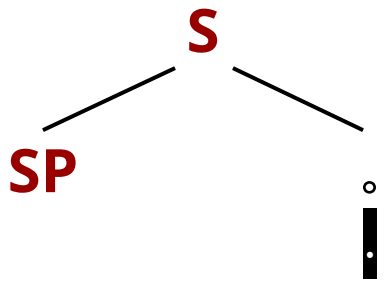
Synchronous Tree



管理局 将会 向 财政 司 负责 。

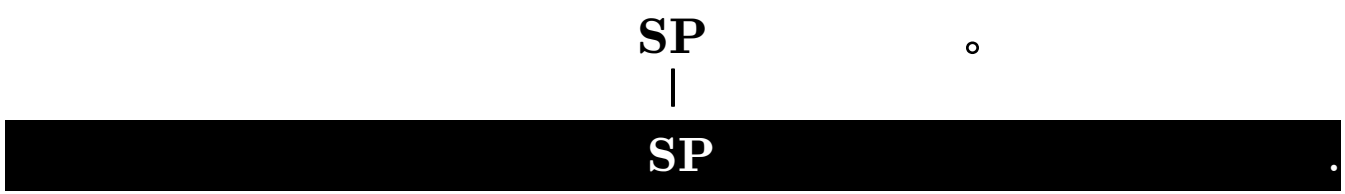
The Authority will be accountable to the Financial Secretary .

Synchronous Derivation

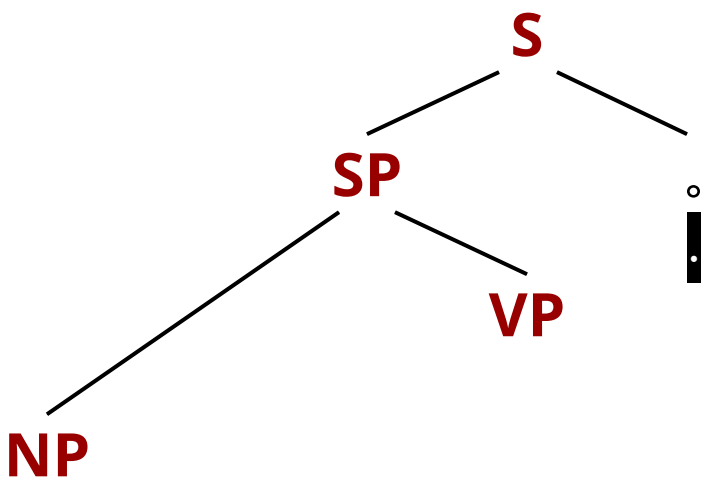


$S \rightarrow (SP\ Stop, SP\ Stop)$

$Stop \rightarrow (。\ ,.)$

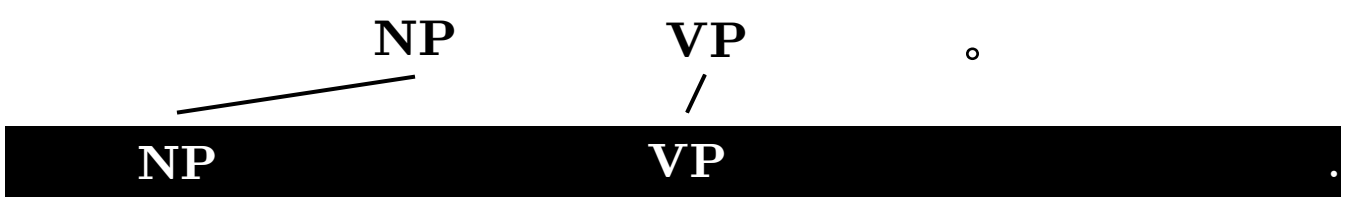


Synchronous Derivation

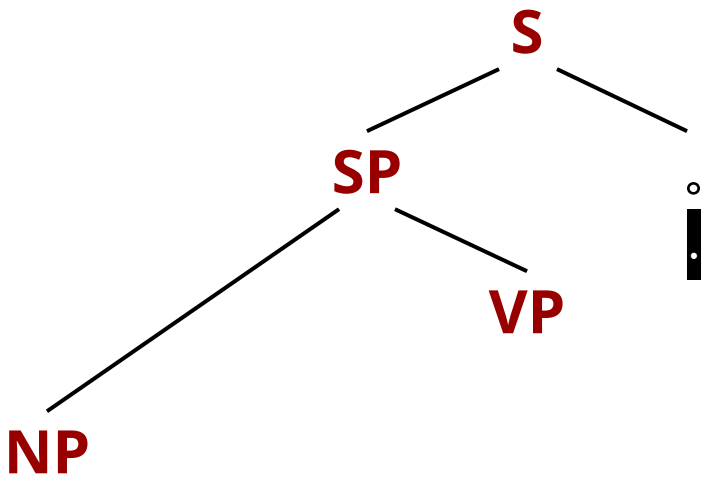


$\mathbf{SP \rightarrow (NP\ VP,\ NP\ VP)}$

We will simplify this rule!

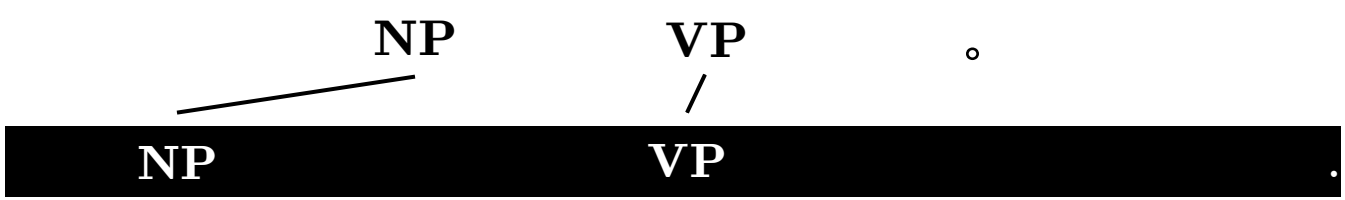


Synchronous Derivation

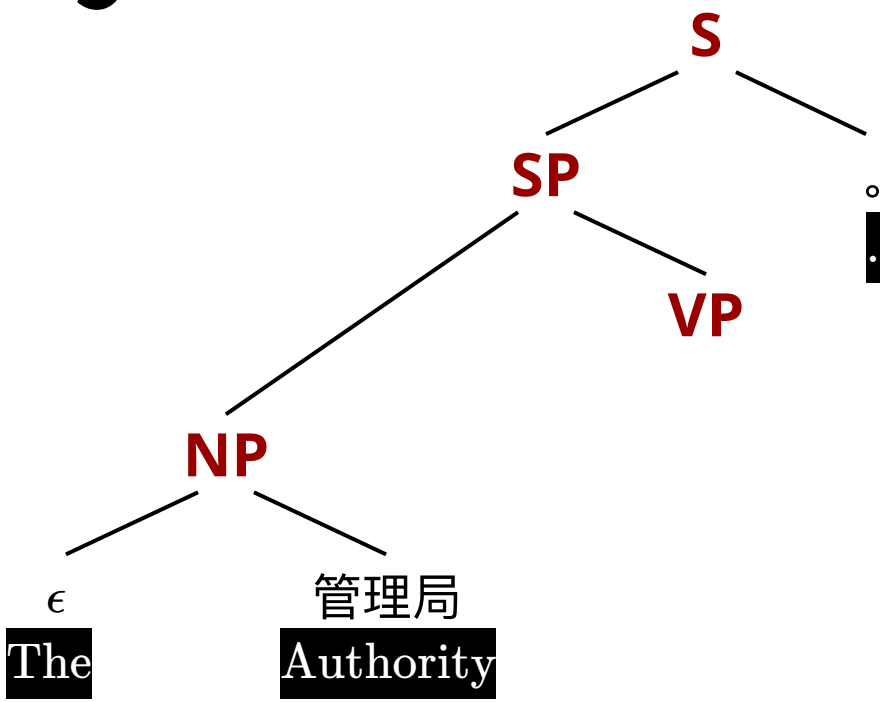


$\mathbf{SP \rightarrow (NP\ VP)}$

This means both source and target follow the same pattern.



Synchronous Derivation



$\mathbf{NP} \rightarrow (\mathbf{D}\ \mathbf{N})$

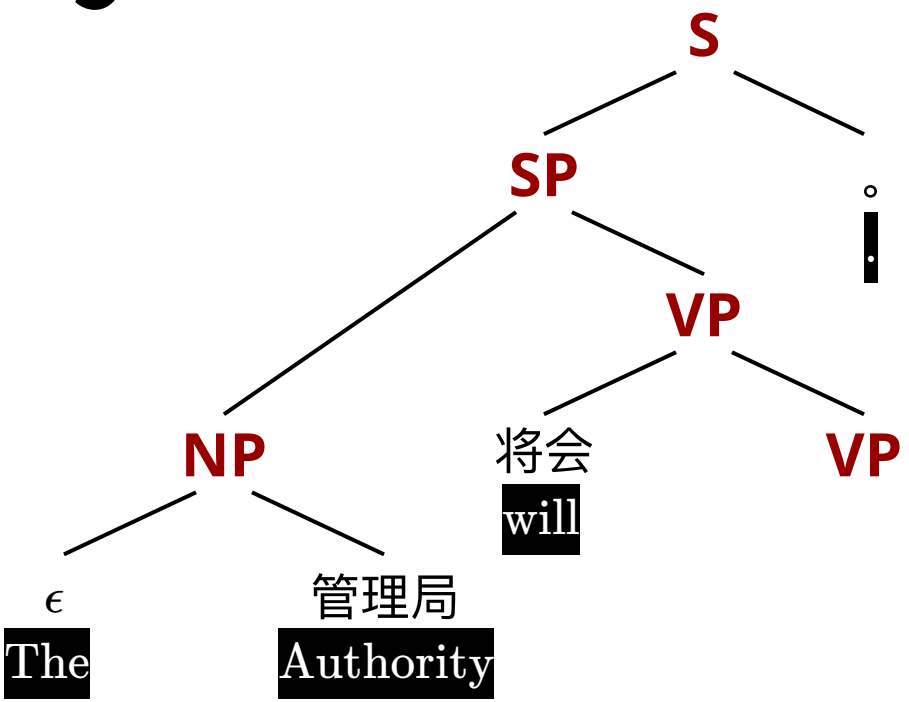
$\mathbf{D} \rightarrow (\epsilon,\ \mathrm{The})$

$\mathbf{N} \rightarrow (\text{管理局},\ \mathrm{Authority})$

管理局 **VP** 。

The Authority **VP** .

Synchronous Derivation



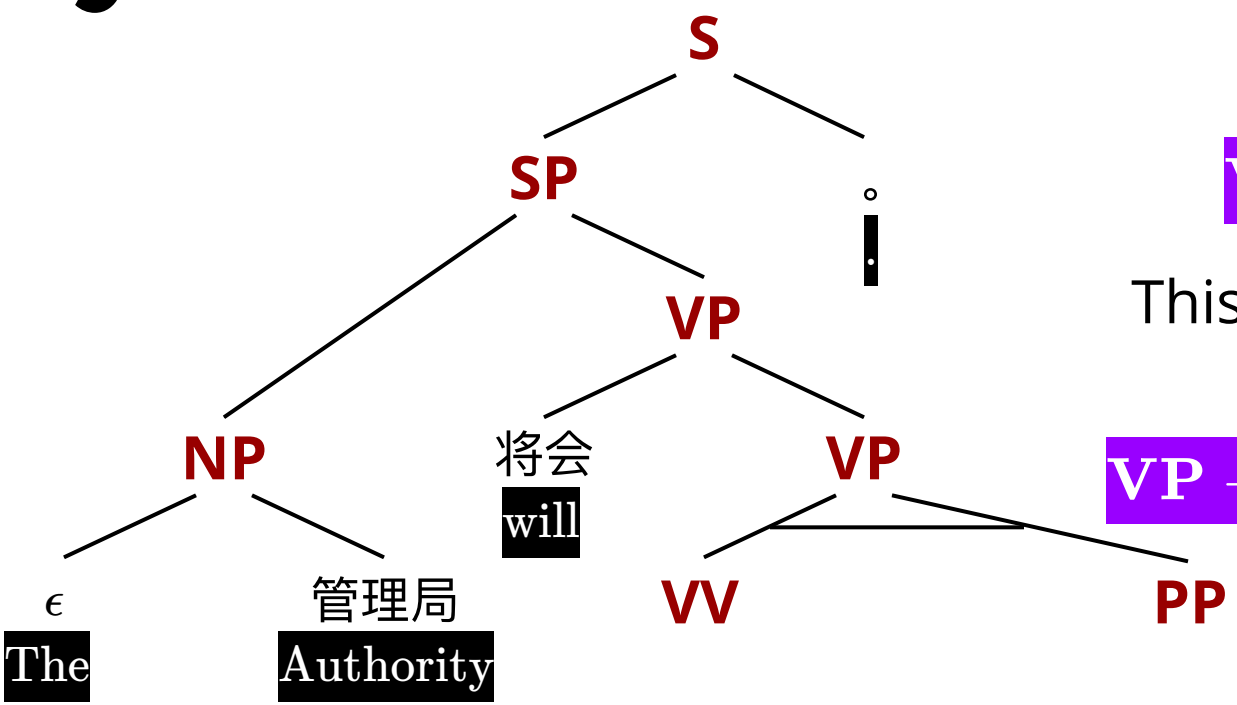
$\mathbf{VP} \rightarrow (\mathbf{V}\ \mathbf{VP})$

$\mathbf{V} \rightarrow (将会, \text{will})$

管理局 将会 **VP** 。

The Authority will **VP** .

Synchronous Derivation

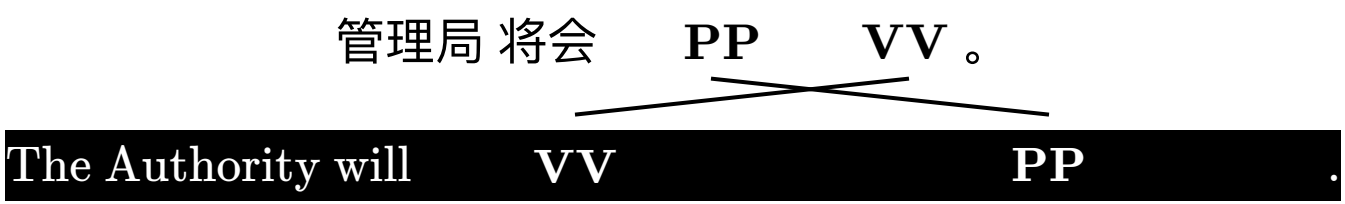


$$\mathbf{VP} \rightarrow \langle \mathbf{VV\ PP} \rangle$$

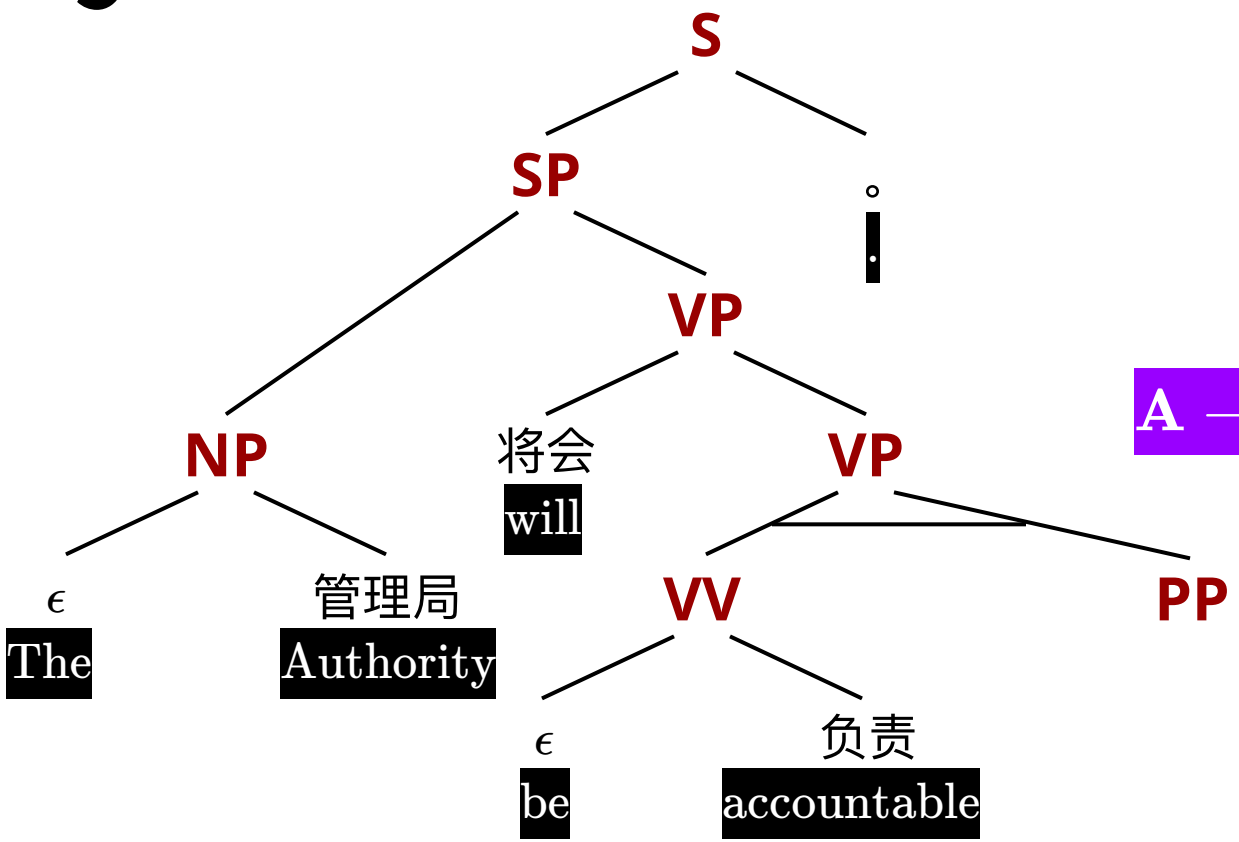
This is equivalent to the following:

$$\mathbf{VP} \rightarrow (\mathbf{PP\ VV},\ \mathbf{VV\ PP})$$

Note that this definition may be slightly different from what's described in the original paper, but is also valid and is more convenient for our illustration.



Synchronous Derivation



$\mathbf{VV} \rightarrow (\mathbf{V}\ \mathbf{A})$

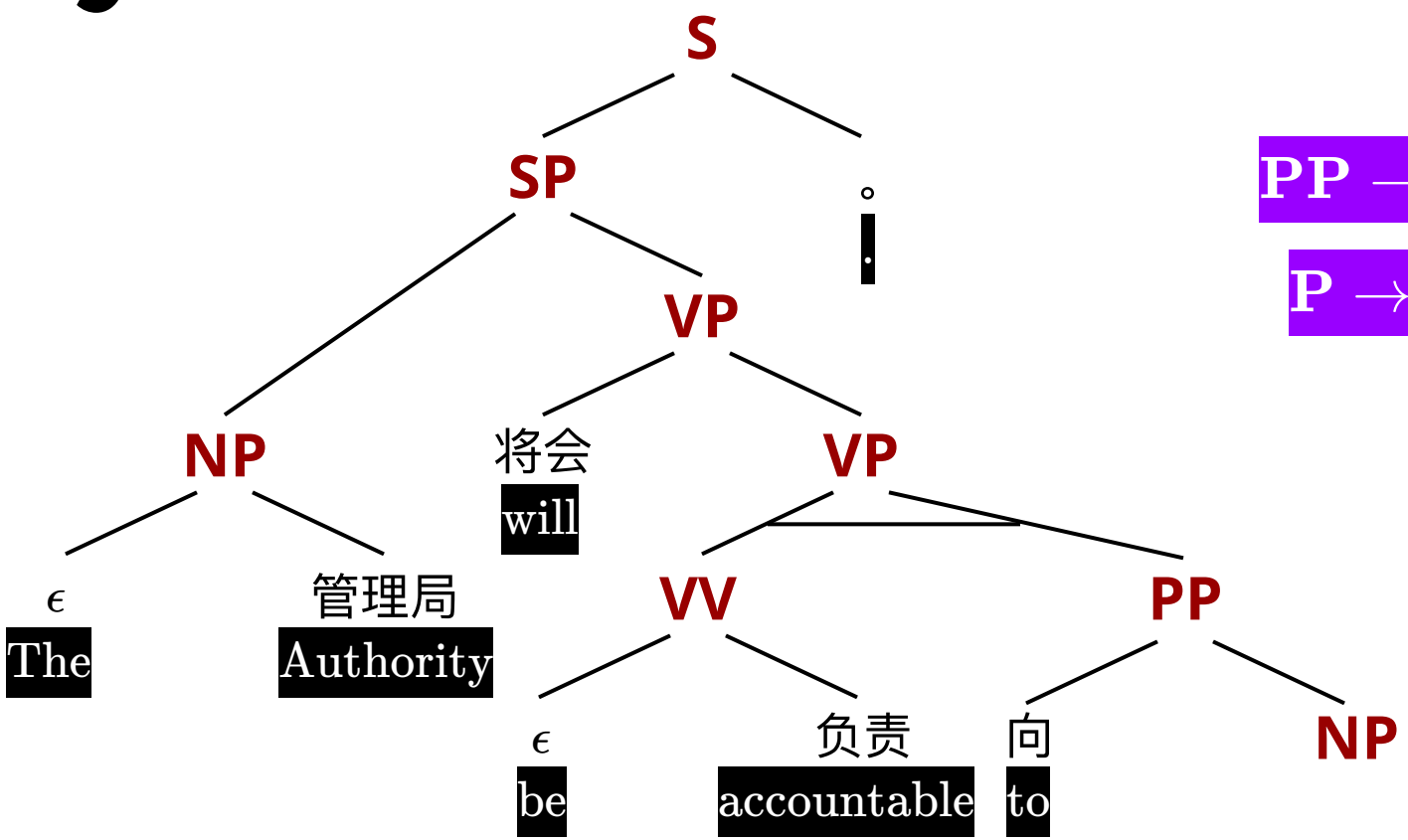
$\mathbf{V} \rightarrow (\epsilon,\ \text{be})$

$\mathbf{A} \rightarrow (\text{负责},\ \text{accountable})$

管理局 将会 **PP** 负责 。

The Authority will be accountable **PP** .

Synchronous Derivation



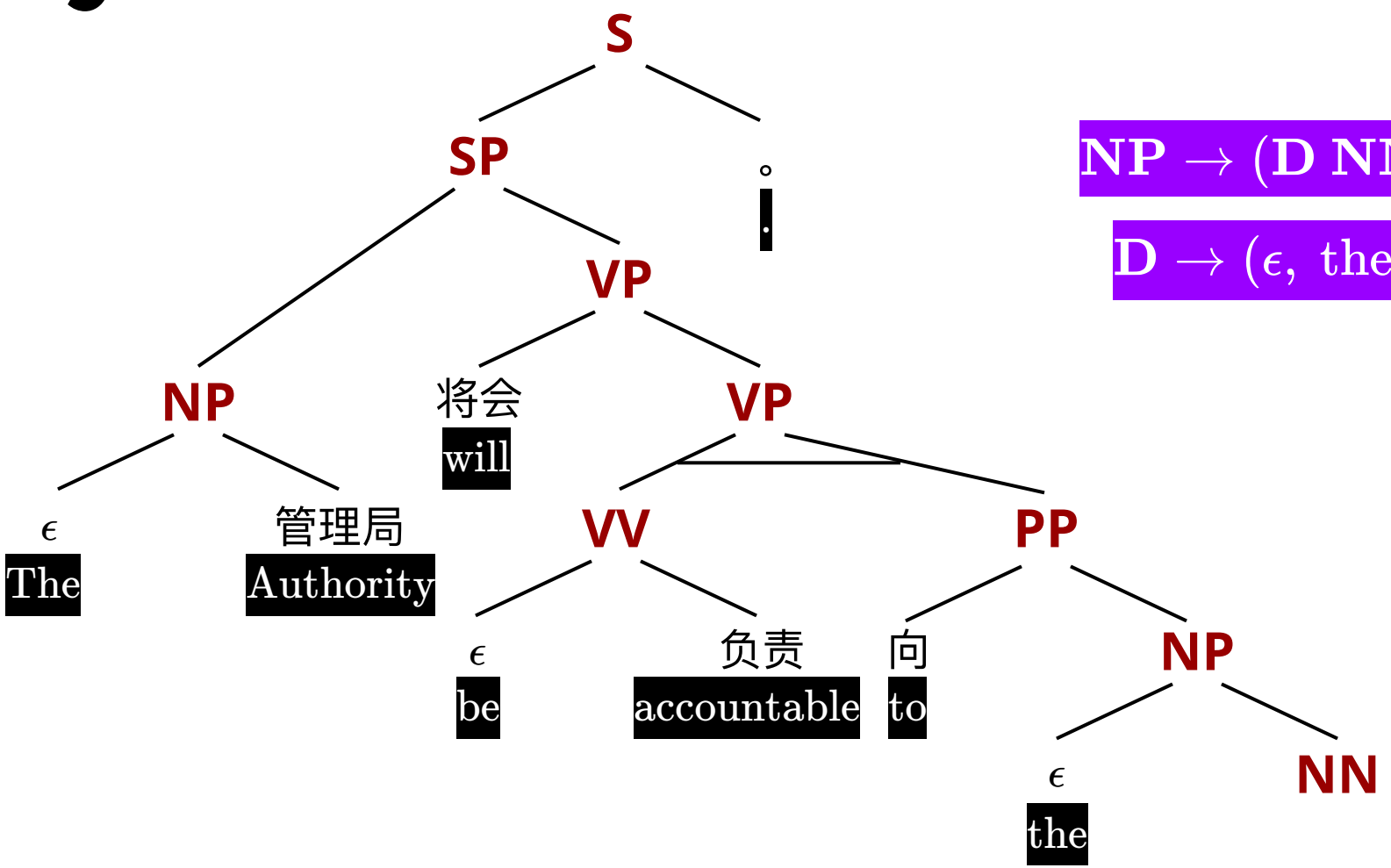
$\mathbf{PP} \rightarrow (\mathbf{P}\ \mathbf{NP})$

$\mathbf{P} \rightarrow (向,\ \text{to})$

管理局 将会 向 **NP** 负责 。

The Authority will be accountable to **NP** .

Synchronous Derivation



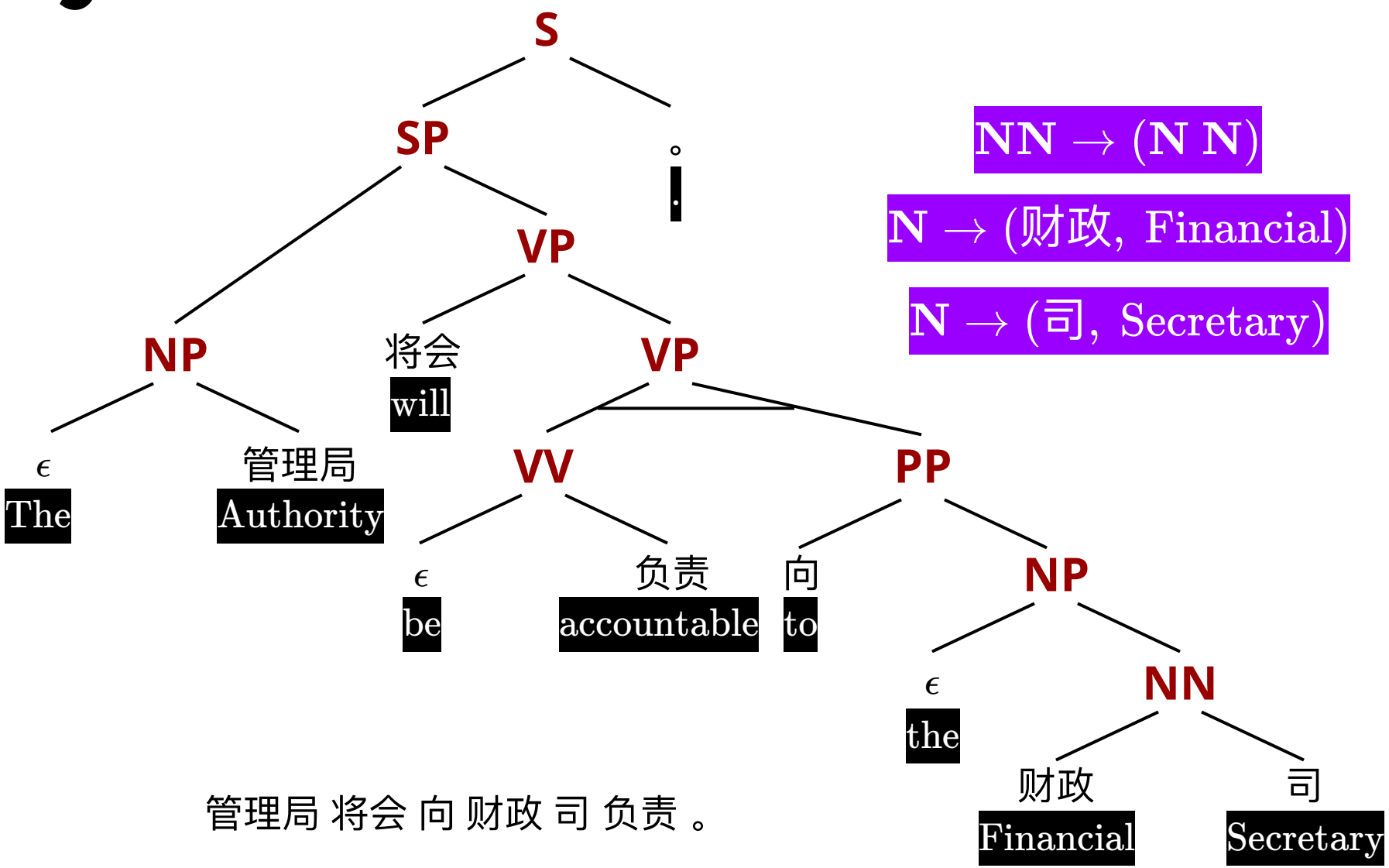
$\mathbf{NP} \rightarrow (\mathbf{D}\ \mathbf{NN})$

$\mathbf{D} \rightarrow (\epsilon,\ \text{the})$

管理局 将会 向 NN 负责 。

The Authority will be accountable to the NN .

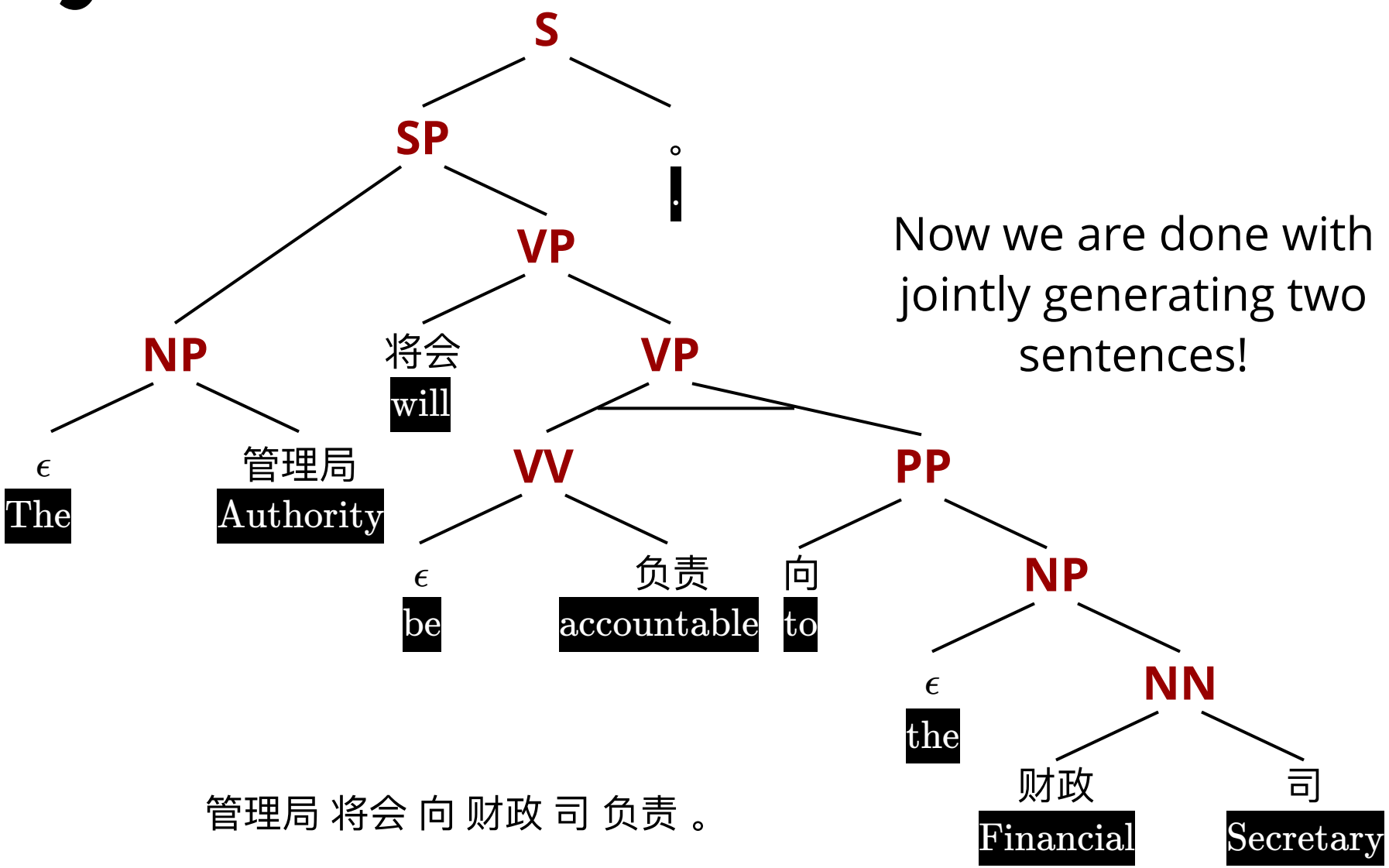
Synchronous Derivation



管理局 将会 向 财政 司 负责 。

The Authority will be accountable to the Financial Secretary .

Synchronous Derivation

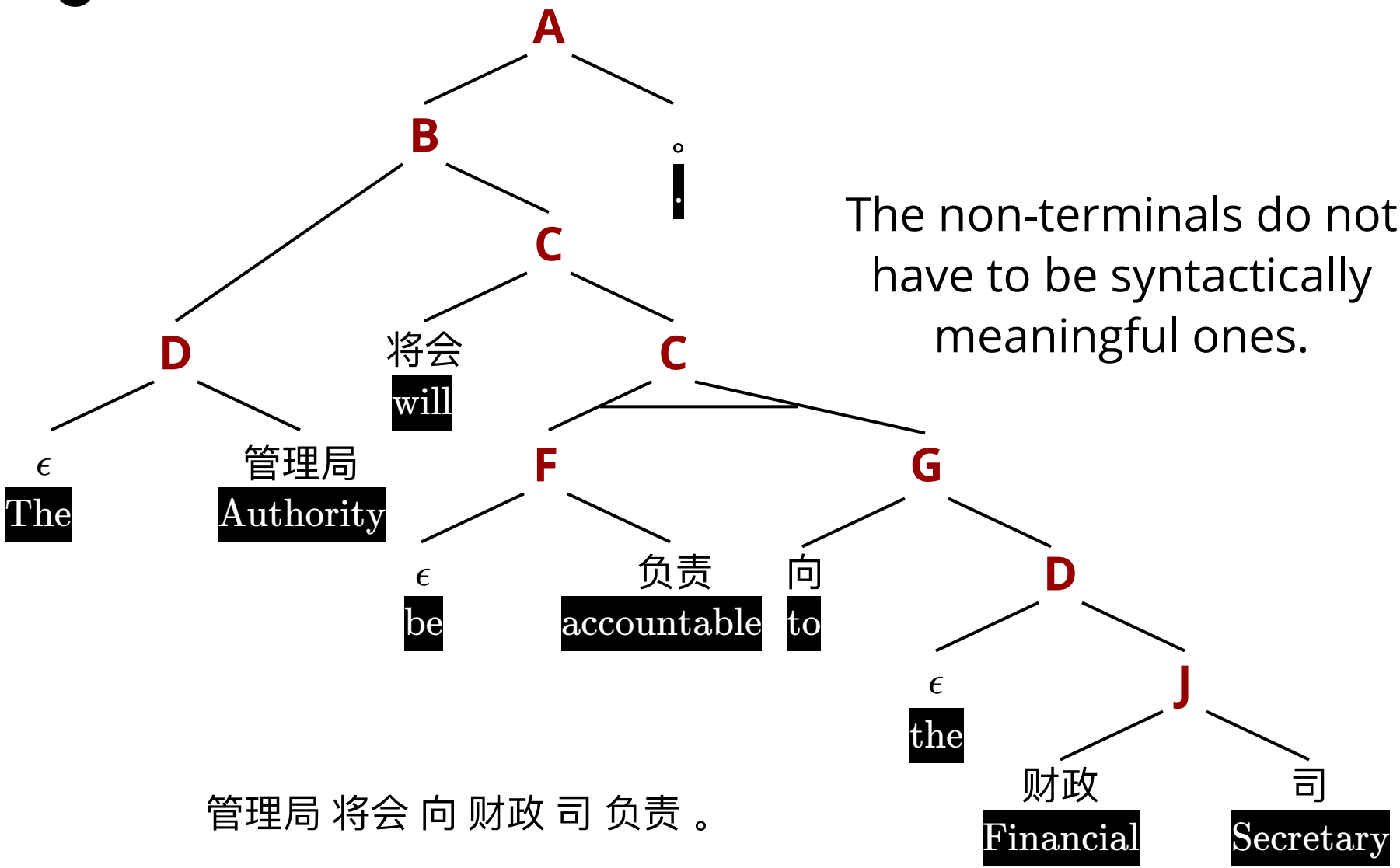


Now we are done with jointly generating two sentences!

管理局 将会 向 财政 司 负责 。

The Authority will be accountable to the Financial Secretary .

Synchronous Derivation

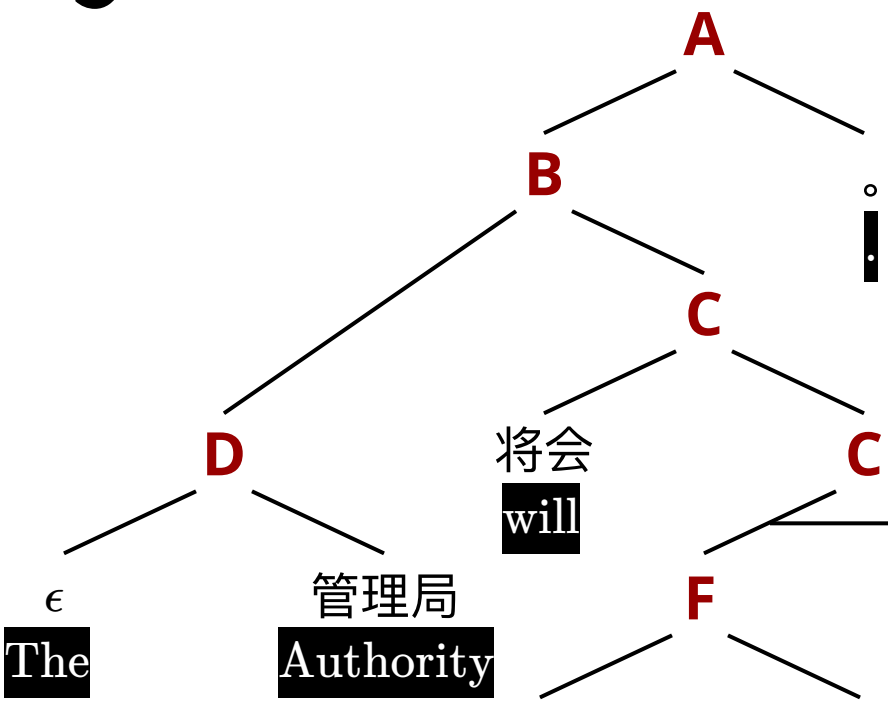


The non-terminals do not have to be syntactically meaningful ones.

管理局 将会 向 财政 司 负责 。

The Authority will be accountable to the Financial Secretary .

Synchronous Derivation

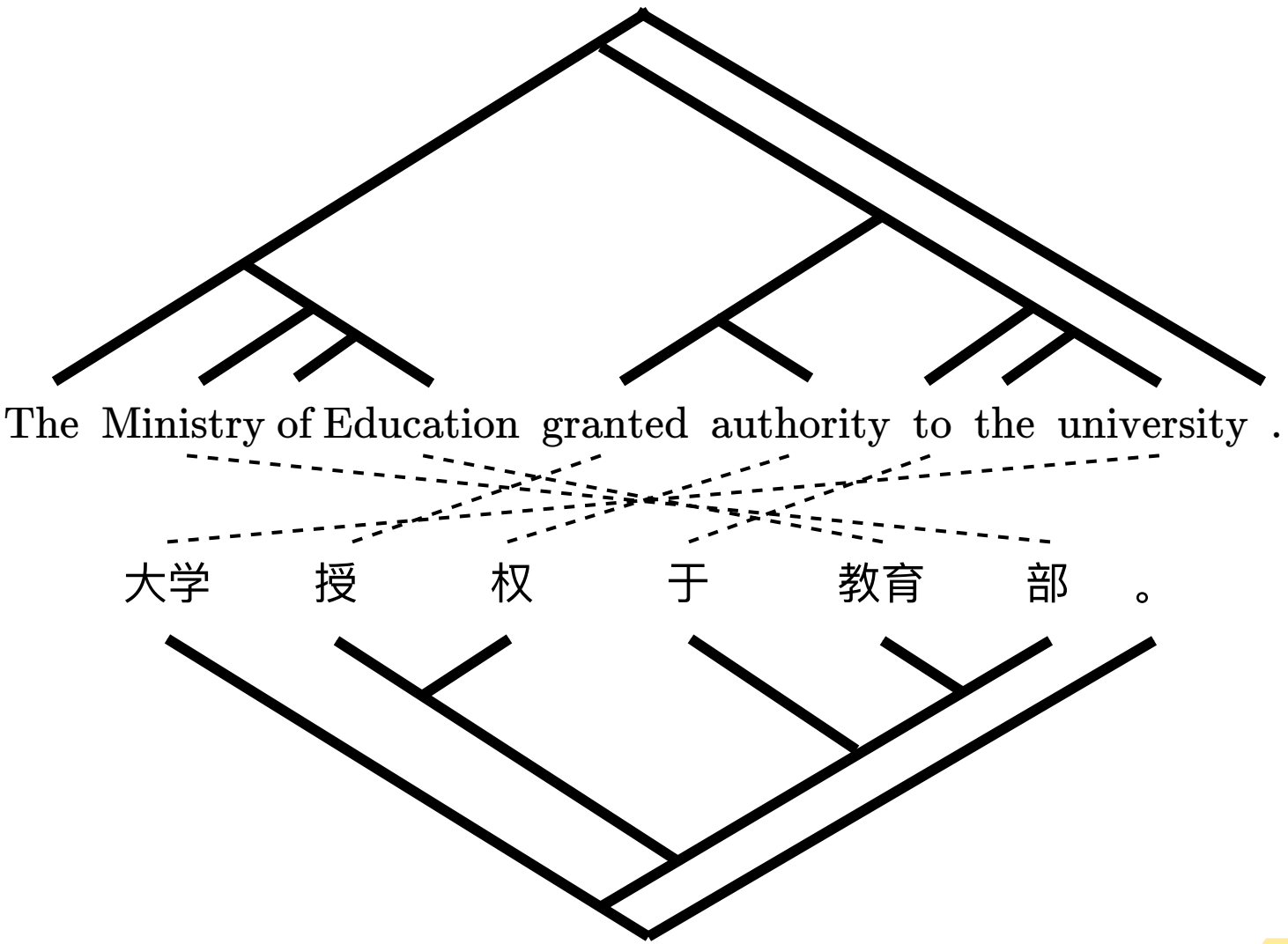


Given the generative story, how do we learn the model?

管理局 将会 向 财政 司 负责 。

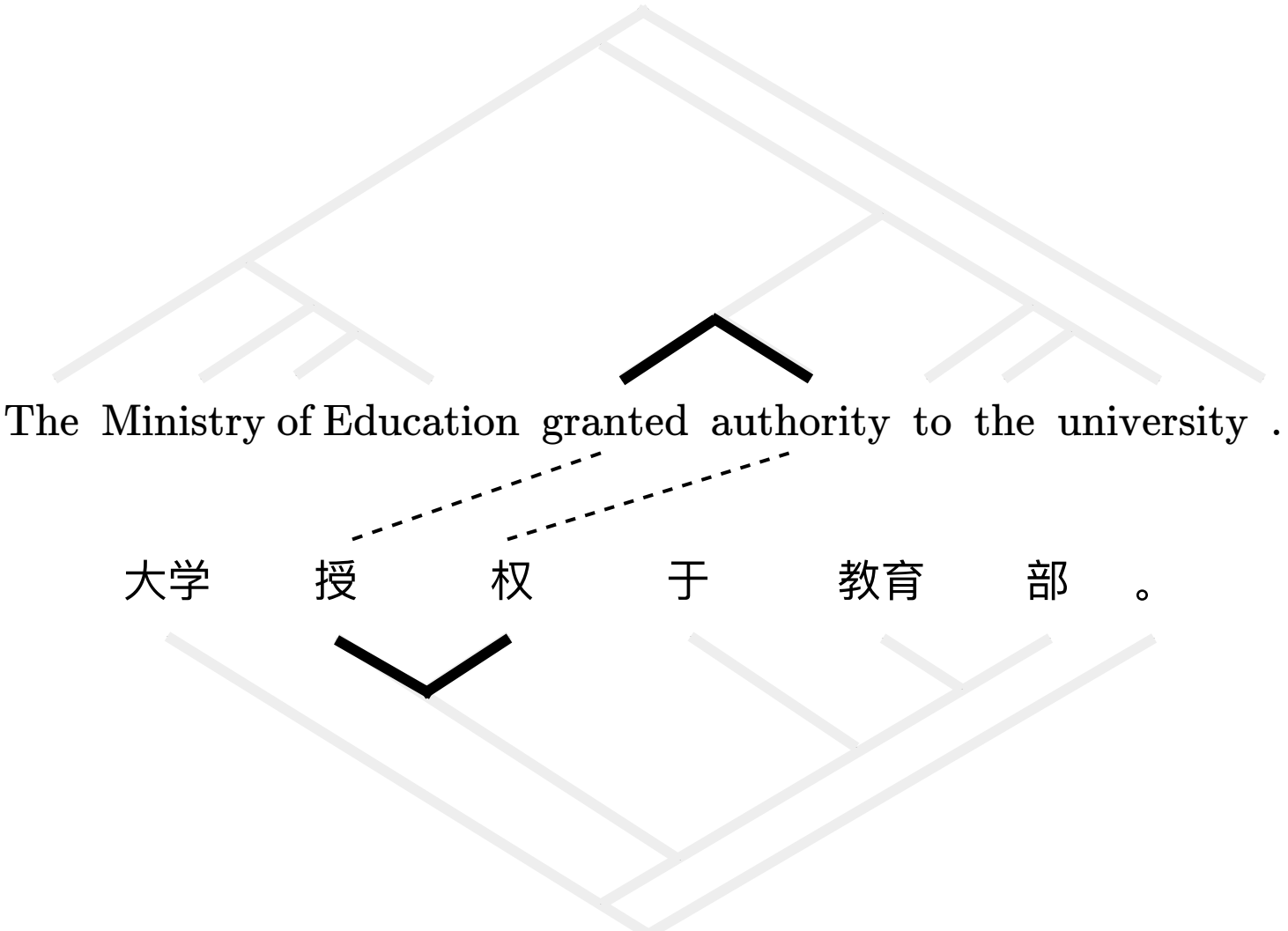
The Authority will be accountable to the Financial Secretary .

Limitations with ITG

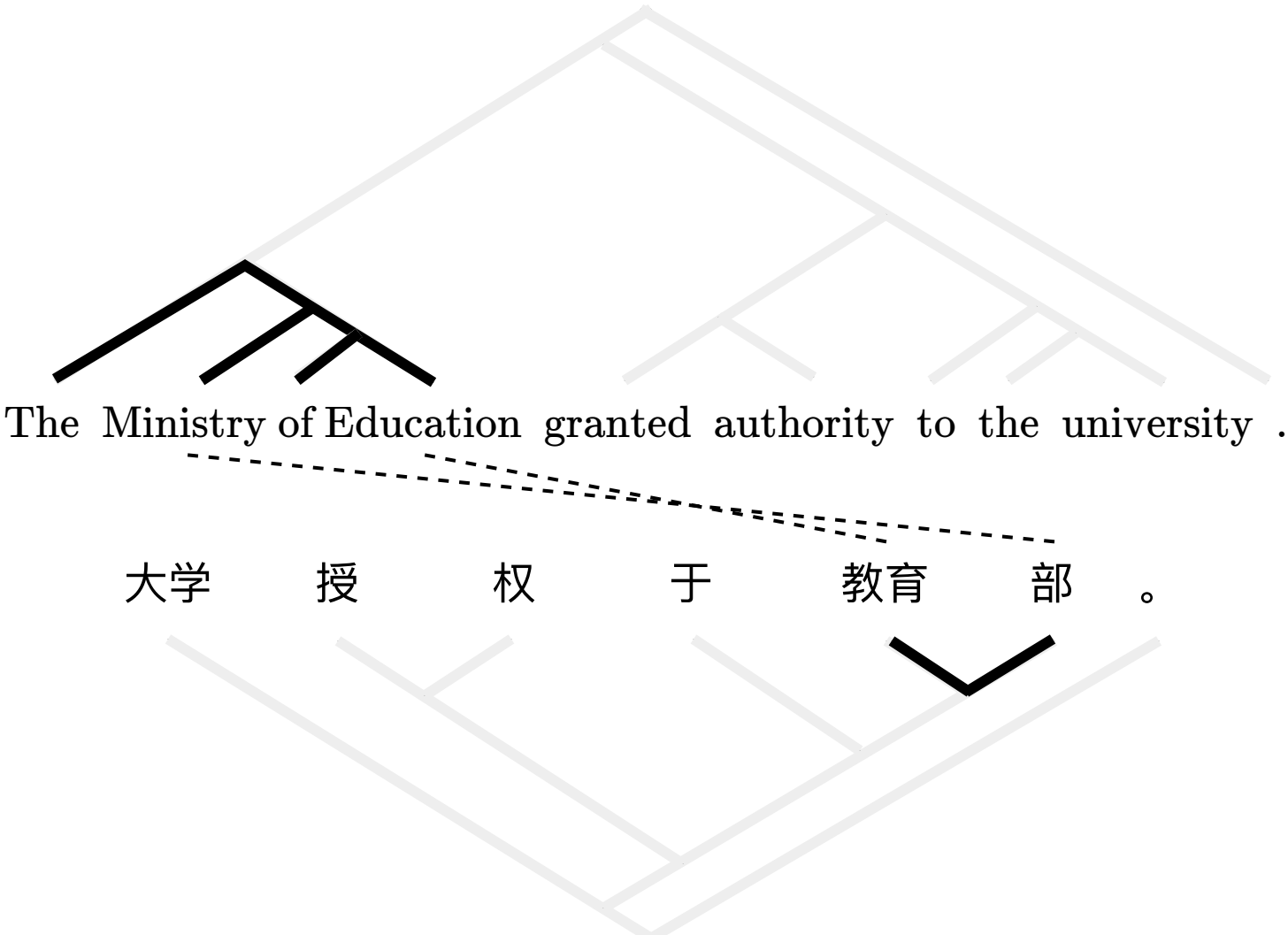


Not possible to find the correct alignment for certain sentence pairs.

Limitations with ITG

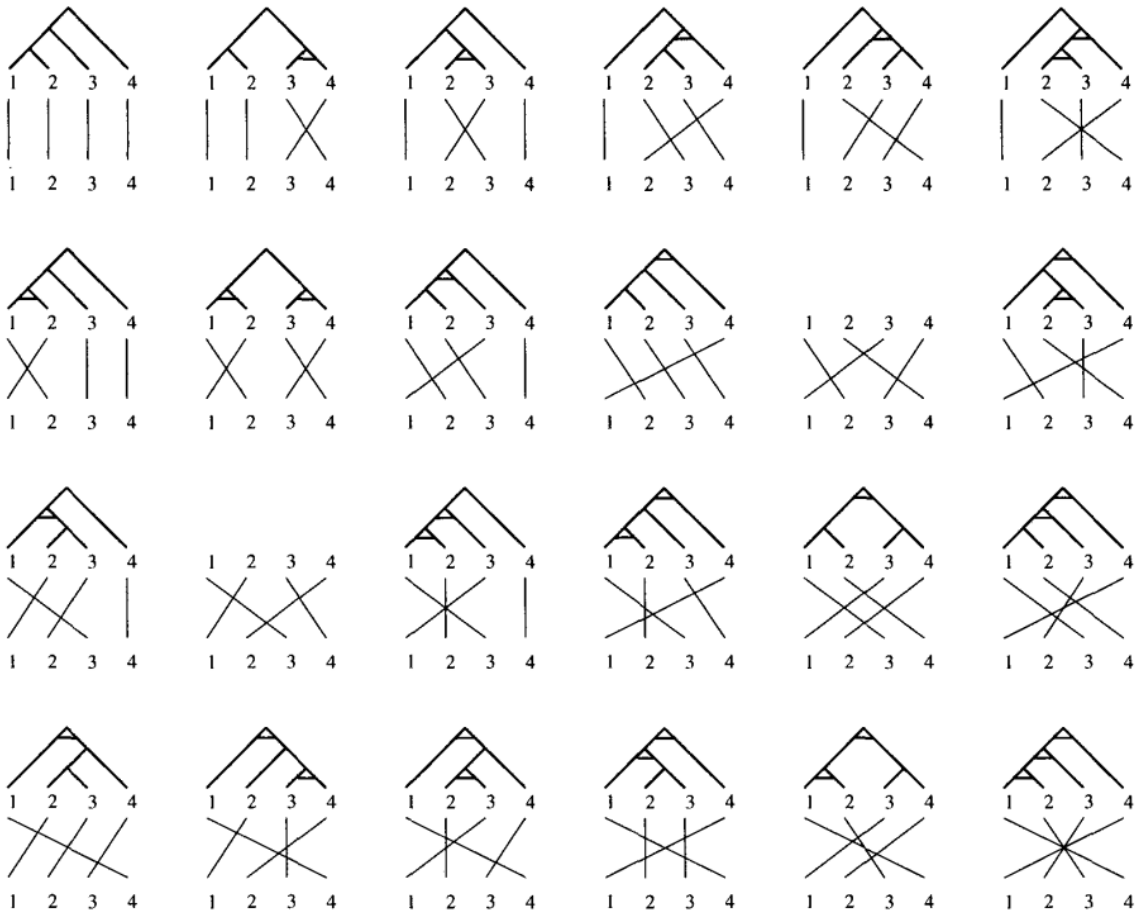


Limitations with ITG



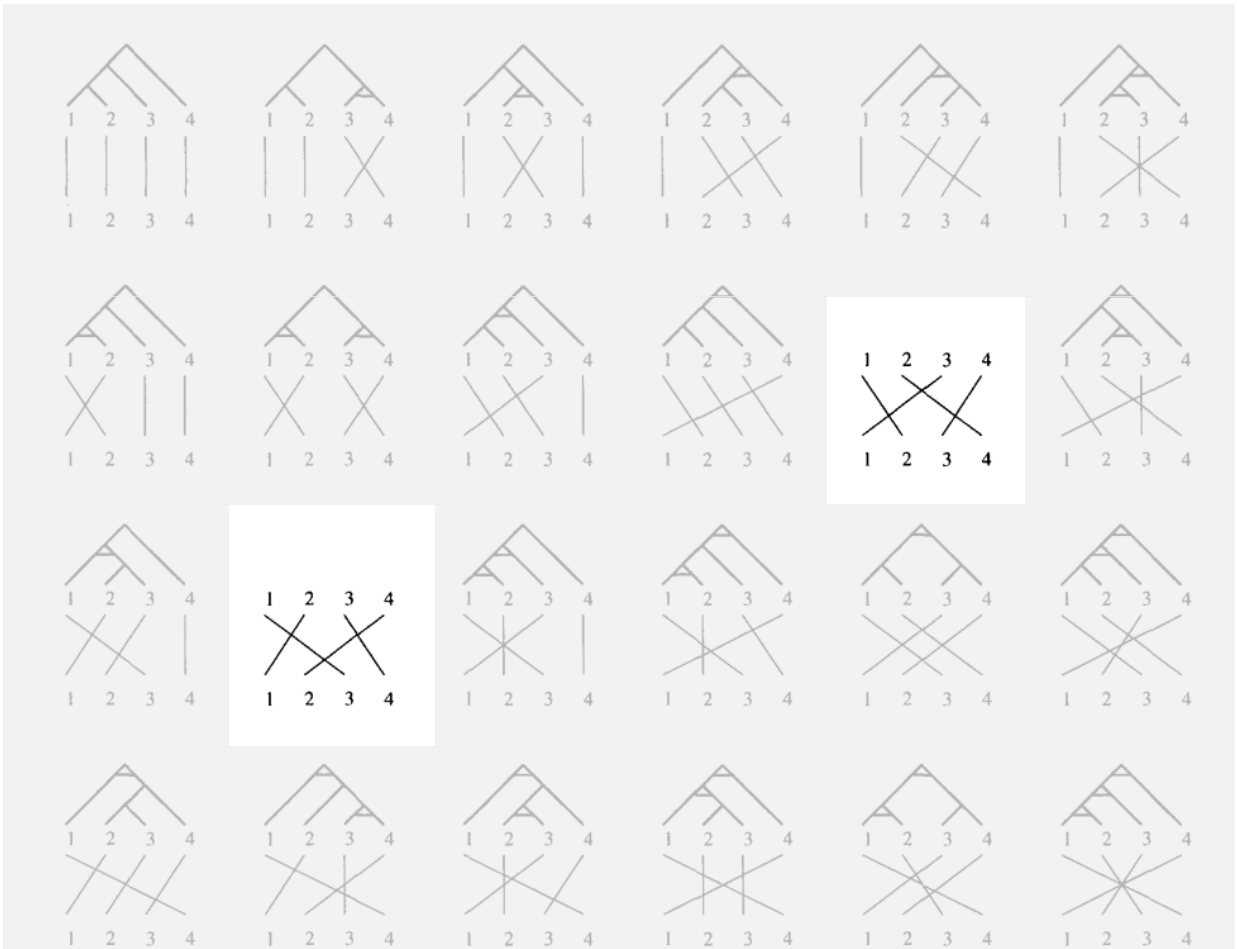
Two different sub-trees are in different larger sub-trees.

Limitations with ITG



Possible word alignments and the corresponding ITG parses (if possible)

Limitations with ITG



2 out of 24 cases where the ITG is unable to handle.

Limitations with ITG

We will come back to this issue shortly…

2 out of 24 cases where the ITG is unable to handle.

Optional

Question

What about parsing the source sentence first?

Optional

Tree Transducer

A Syntax-based Statistical Translation Model

Kenji Yamada and **Kevin Knight**
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
{kyamada,knight}@isi.edu

Abstract

We present a syntax-based statistical translation model. Our model transforms a source-language parse tree into a target-language string by applying stochastic operations at each node. These operations capture linguistic differences such as word order and case marking. Model parameters are estimated in polynomial time using an EM algorithm. The model produces word alignments that are better than those produced by IBM Model 5.

1 Introduction

A statistical translation model (TM) is a mathematical model in which the process of human-language translation is statistically modeled. Model parameters are automatically estimated using a corpus of translation pairs. TMs have been used for statistical machine translation (Berger et al., 1996), word alignment of a translation corpus (Melamed, 2000), multilingual document retrieval (Franz et al., 1999), automatic dictionary construction (Resnik and Melamed, 1997), and data preparation for word sense disambiguation programs (Brown et al., 1991). Developing a better TM is a fundamental issue for those applications.

Researchers at IBM first described such a statistical TM in (Brown et al., 1988). Their models are based on a string-to-string noisy channel model. The channel converts a sequence of words in one language (such as English) into another (such as French). The channel operations are movements, duplications, and translations, applied to each word independently. The movement is conditioned only on word classes and positions in the string, and the duplication and translati... are conditioned only on the word identi... ematical details are fully described in (... al., 1993).

One criticism of the IBM-style TM is ... does not model structural or syntactic aspec... the language. The TM was only demonstrated ... a structurally similar language pair (English a... French). It has been suspected that a language ... pair with very different word order such as En... glish and Japanese would not be modeled well by these TMs.

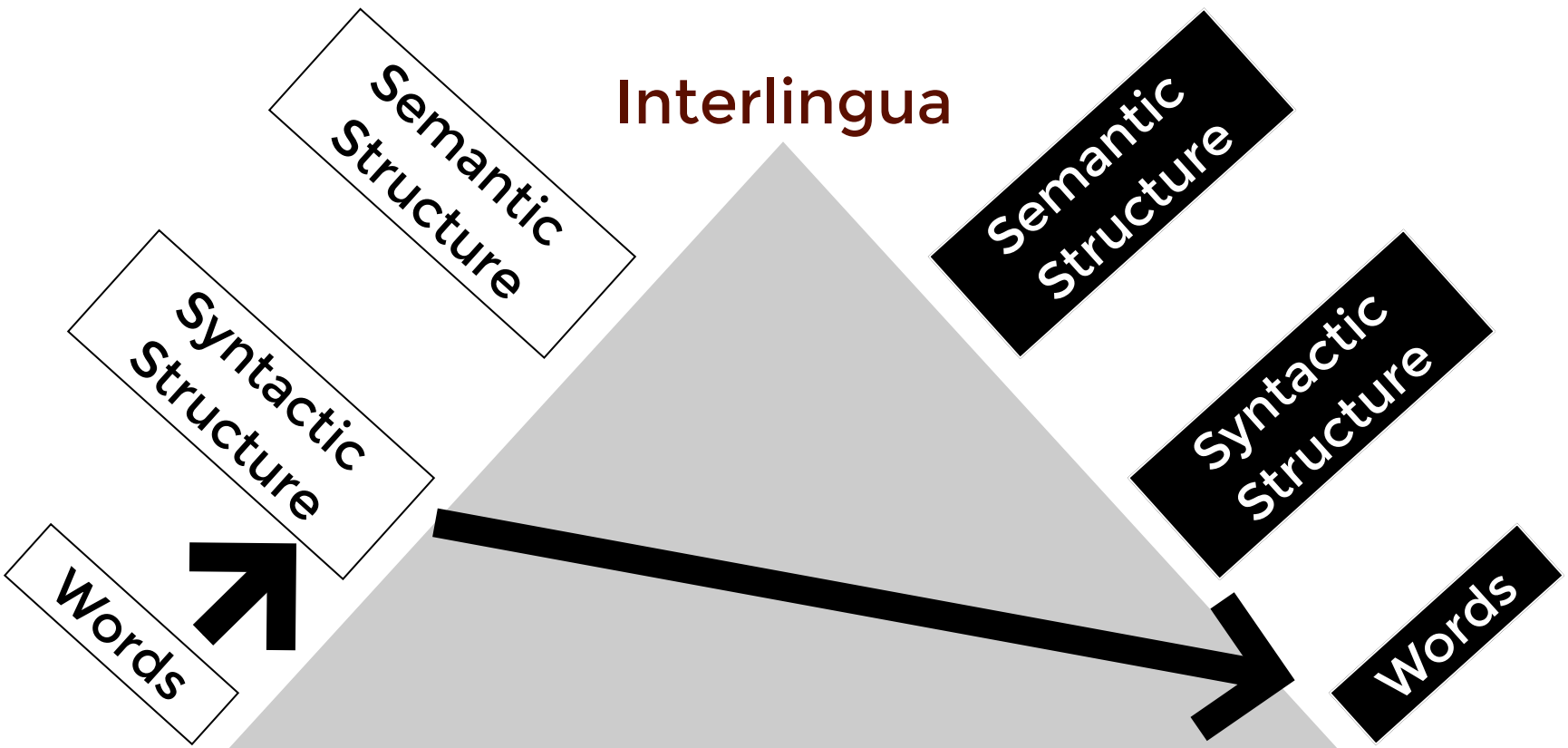
To incorporate structural aspects of the language, our channel model accepts a parse tree as an input, i.e., the input sentence is preprocessed by a syntactic parser. The channel performs operations on each node of the parse tree. The operations are *reordering* child nodes, *inserting* extra words at each node, and *translating* leaf words. Figure 1 shows the overview of the operations of our model. Note that the output of our model is a string, not a parse tree. Therefore, parsing is only needed on the channel input side.

The reorder operation is intended to model translation between languages with different word orders, such as SVO-languages (English or Chinese) and SOV-languages (Japanese or Turkish). The word-insertion operation is intended to capture linguistic differences in specifying syntactic cases. E.g., English and French use structural position to specify case, while Japanese and Korean use case-marker particles.

Wang (1998) enhanced the IBM models by introducing phrases, and Och *et al.* (1999) used templates to capture phrasal sequences in a sentence. Both also tried to incorporate structural aspects of the language, however, neither handles

Optional

Tree Transducer



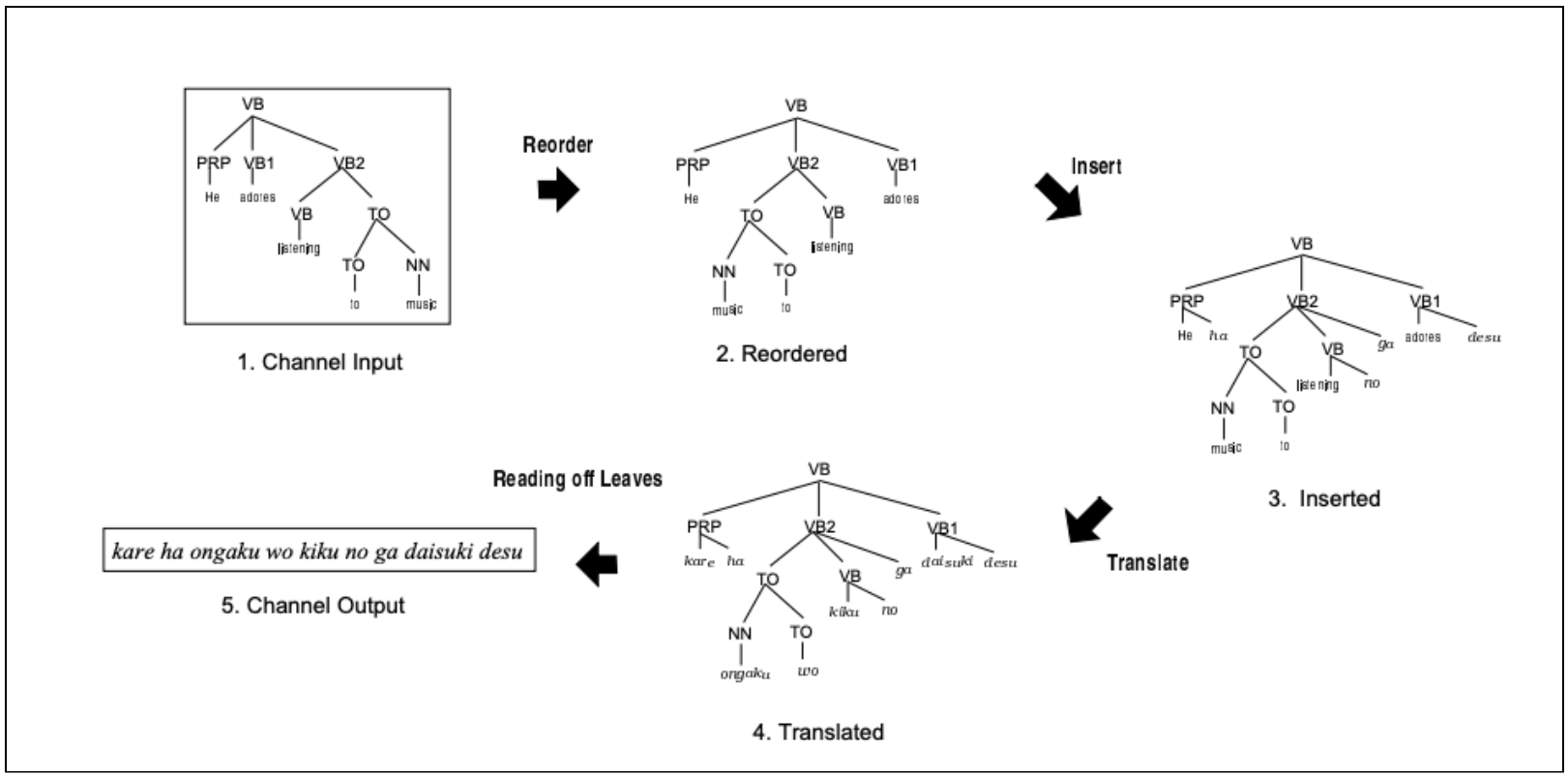
Syntactic Parsing

Channel Operations

Language Generation

Optional

Tree Transducer



Define several operations/actions that can be applied on top of source syntax trees. It incrementally transform the source syntax tree into target sentence.

Question

Can we introduce phrases into the translation process?

Formal Syntax

Hierarchical Phrase-Based Translation

David Chiang*
Information Sciences Institute
University of Southern California

We present a statistical machine translation model that uses hierarchical phrases—phr
that contain subphrases. The model is formally a synchronous context-free grammar but
learned from a parallel text without any syntactic annotations. Thus it can be seen as combinin
fundamental ideas from both syntax-based translation and phrase-based translation. We describe
our system's training and decoding methods in detail, and evaluate it for translation speed and
translation accuracy. Using BLEU as a metric of translation accuracy, we find that our system
performs significantly better than the Alignment Template System, a state-of-the-art phrase-
based system.

1. Introduction

The alignment template translation model (Och and Ney 2004) and related phrase-based models advanced the state of the art in machine translation by expanding the basic unit of translation from words to **phrases**, that is, substrings of potentially unlimited size (but not necessarily phrases in any syntactic theory). These phrases allow a model to learn local reorderings, translations of multiword expressions, or insertions and deletions that are sensitive to local context. This makes them a simple and powerful mechanism for translation.

The basic phrase-based model is an instance of the noisy-channel approach (Brown et al. 1993). Following convention, we call the source language "French" and the target language "English"; the translation of a French sentence f into an English sentence e is modeled as:

$$\arg\max_e P(e \mid f) = \arg\max_e P(e, f) \tag{1}$$

$$= \arg\max_e (P(e) \times P(f \mid e)) \tag{2}$$

The phrase-based translation model $P(f \mid e)$ "encodes" e into f by the following steps:

1. segment e into phrases $\bar{e}_1 \cdots \bar{e}_I$, typically with a uniform distribution over segmentations;

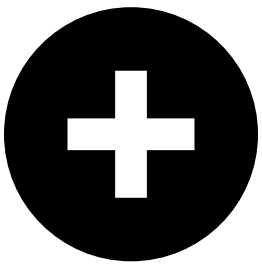
* 4676 Admiralty Way, Suite 1001, Marina del Rey, CA 90292, USA. E-mail: chiang@isi.edu. Much of the research presented here was carried out while the author was at the University of Maryland Institute for Advanced Computer Studies.

Submission received: 1 May 2006; accepted for publication: 3 October 2006.

Hierarchical Phrase based Translation

Formal Syntax

Phrase-based Translation



Synchronous Grammar

It combines the idea of phrase-based translation and synchronous parsing.

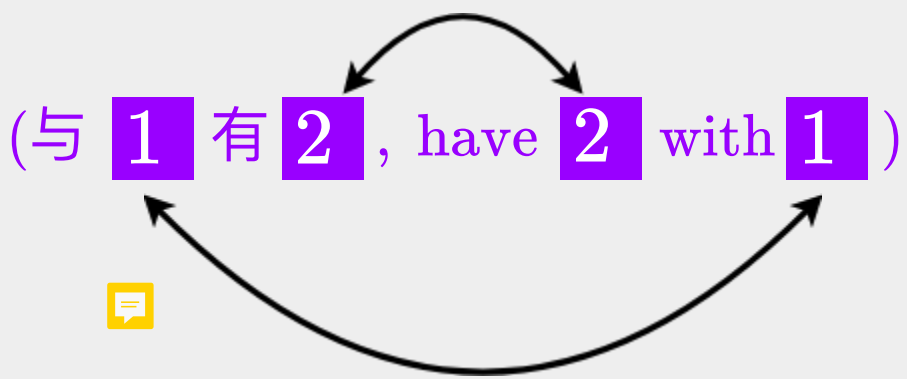
Hierarchical Phrase Pairs

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一

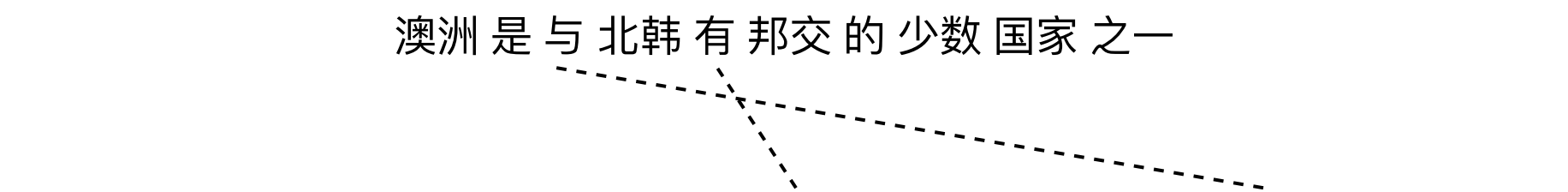
Australia is one of the few countries that have diplomatic relations with North Korea

(北韩, North Korea)

(邦交, diplomatic relations)



Synchronous Grammar



$\mathbf{X} \rightarrow (\text{与 } \mathbf{X}_1 \text{ 有 } \mathbf{X}_2,\ \text{have } \mathbf{X}_2 \text{ with } \mathbf{X}_1)$

$\mathbf{X} \rightarrow (\text{北韩},\ \text{North Korea})$

$\mathbf{X} \rightarrow (\text{邦交},\ \text{diplomatic relations})$

Subscripts indicate the alignment

Use some heuristics similar to what we discussed in phrase-based translation to acquire the synchronous grammar rules

Grammar Rule

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一

Australia is one of the few countries that have diplomatic relations with North Korea

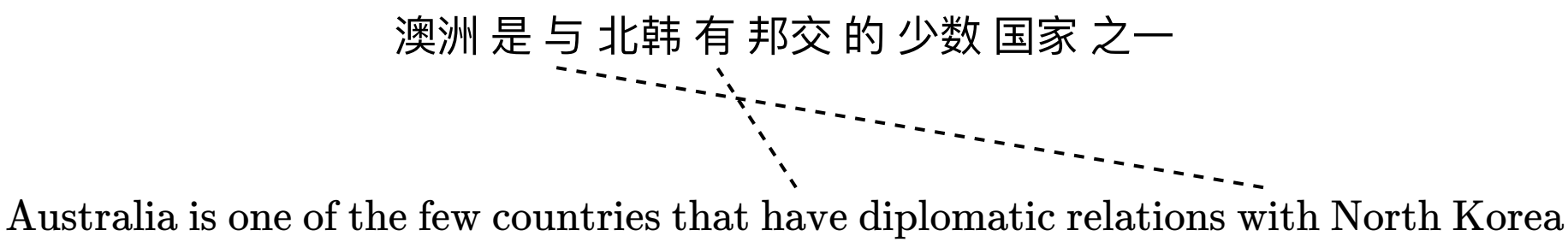
A non-terminal symbol

$$\mathbf{X} \rightarrow (\gamma,\ \alpha)$$

A sequence of *source* terminal symbols intermixed with non-terminal symbols

A sequence of *target* terminal symbols intermixed with non-terminal symbols

Grammar Rule



$$\mathbf{S} \rightarrow (\mathbf{X}_1,\ \mathbf{X}_1) \qquad \text{A}$$

$$\mathbf{S} \rightarrow (\mathbf{S}_1\mathbf{X}_2,\ \mathbf{S}_1\mathbf{X}_2) \qquad \text{B}$$

Two special rules considered (auxiliary rules for completion of derivations)

Synchronous Derivation

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一

Australia is one of the few countries that have diplomatic relations with North Korea

$$\mathbf{X} \rightarrow (\text{与 } \mathbf{X}_1 \text{ 有 } \mathbf{X}_2,\ \text{have } \mathbf{X}_2 \text{ with } \mathbf{X}_1) \quad ❶$$

$$\mathbf{X} \rightarrow (\mathbf{X}_1 \text{ 的 } \mathbf{X}_2,\ \text{the } \mathbf{X}_2 \text{ that } \mathbf{X}_1) \quad ❷$$

$$\mathbf{X} \rightarrow (\mathbf{X}_1 \text{ 之一},\ \text{one of } \mathbf{X}_1) \quad ❸$$

$$\mathbf{X} \rightarrow (\text{澳洲},\ \text{Australia}) \quad ❹$$

$$\mathbf{X} \rightarrow (\text{北韩},\ \text{North Korea}) \quad ❺$$

$$\mathbf{X} \rightarrow (\text{是},\ \text{is}) \quad ❻$$

$$\mathbf{X} \rightarrow (\text{邦交},\ \text{diplomatic relations}) \quad ❼$$

$$\mathbf{X} \rightarrow (\text{少数 国家},\ \text{few countries}) \quad ❽$$

Can you work out the derivation for this sentence pair based on the above rules (and the two special rules)?

Synchronous Derivation

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一

Australia is one of the few countries that have diplomatic relations with North Korea

$(\mathbf{S}_1,\ \mathbf{S}_1)$

$\xRightarrow[\mathbf{S}_1]{(B)}$ $(\mathbf{S}_2\mathbf{X}_3,\ \mathbf{S}_2\mathbf{X}_3)$

Rewrite $\mathbf{S}_1$ with rule (B)

$\xRightarrow[\mathbf{S}_2]{(B)}$ $(\mathbf{S}_4\mathbf{X}_5\mathbf{X}_3,\ \mathbf{S}_4\mathbf{X}_5\mathbf{X}_3)$

$\xRightarrow[\mathbf{S}_4]{(A)}$ $(\mathbf{X}_6\mathbf{X}_5\mathbf{X}_3,\ \mathbf{X}_6\mathbf{X}_5\mathbf{X}_3)$

$\xRightarrow[\mathbf{X}_6]{(4)}$ (澳洲 $\mathbf{X}_5\mathbf{X}_3$, Australia $\mathbf{X}_5\mathbf{X}_3$)

$\xRightarrow[\mathbf{X}_5]{(6)}$ (澳洲 是 $\mathbf{X}_3$, Australia is $\mathbf{X}_3$)

$\xRightarrow[\mathbf{X}_3]{(3)}$ (澳洲 是 $\mathbf{X}_7$ 之一, Australia is one of $\mathbf{X}_7$)

Synchronous Derivation

$\xrightarrow[\mathbf{X}_7]{(2)}$ (澳洲 是 $\mathbf{X}_8$ 的 $\mathbf{X}_9$ 之一, Australia is one of the $\mathbf{X}_9$ that $\mathbf{X}_8$)

$\xrightarrow[\mathbf{X}_8]{(1)}$ (澳洲 是 与 $\mathbf{X}_{10}$ 有 $\mathbf{X}_{11}$ 的 $\mathbf{X}_9$ 之一,

Australia is one of the $\mathbf{X}_9$ that have $\mathbf{X}_{11}$ with $\mathbf{X}_{10}$)

$\xrightarrow[\mathbf{X}_{10}]{(5)}$ (澳洲 是 与 北韩 有 $\mathbf{X}_{11}$ 的 $\mathbf{X}_9$ 之一,

Australia is one of the $\mathbf{X}_9$ that have $\mathbf{X}_{11}$ with North Korea)

$\xrightarrow[\mathbf{X}_{11}]{(7)}$ (澳洲 是 与 北韩 有 邦交 的 $\mathbf{X}_9$ 之一,

Australia is one of the $\mathbf{X}_9$ that have diplomatic relations with North Korea)

$\xrightarrow[\mathbf{X}_9]{(8)}$ (澳洲 是 与 北韩 有 邦交 的 少数 国家 之一,

Australia is one of the few countries that have diplomatic relations with North Korea)

Weighted Grammar Rule

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一

Australia is one of the few countries that have diplomatic relations with North Korea

$$\mathbf{X} \rightarrow (\text{与 } \mathbf{X}_1 \text{ 有 } \mathbf{X}_2,\ \text{have } \mathbf{X}_2 \text{ with } \mathbf{X}_1)\ +7.2$$

$$\mathbf{X} \rightarrow (\text{北韩},\ \text{North Korea})\ -0.7$$

$$\mathbf{X} \rightarrow (\text{邦交},\ \text{diplomatic relations})\ +3.9$$

We are interested in a weighted synchronous context-free grammar (SCFG)

From where can we get the weights?

Score of Derivation

$$score(D) = \underbrace{\sum_{\mathbf{X} \to \langle \gamma, \alpha \rangle \in D} score(\mathbf{X} \to \langle \gamma, \alpha \rangle)}_{\text{Grammar rules}}$$

$$+ \quad \underbrace{\log q(e)}_{\text{Language model}}$$

Optional

This decoding algorithm is more complicated (especially the "top-k" decoding algorithm).

Score of Derivation

$$score(D) = \lambda_{\text{TM}} \times \underbrace{\sum_{\mathbf{X} \to \langle \gamma, \alpha \rangle \in D} score(\mathbf{X} \to \langle \gamma, \alpha \rangle)}_{\text{Grammar rules}}$$

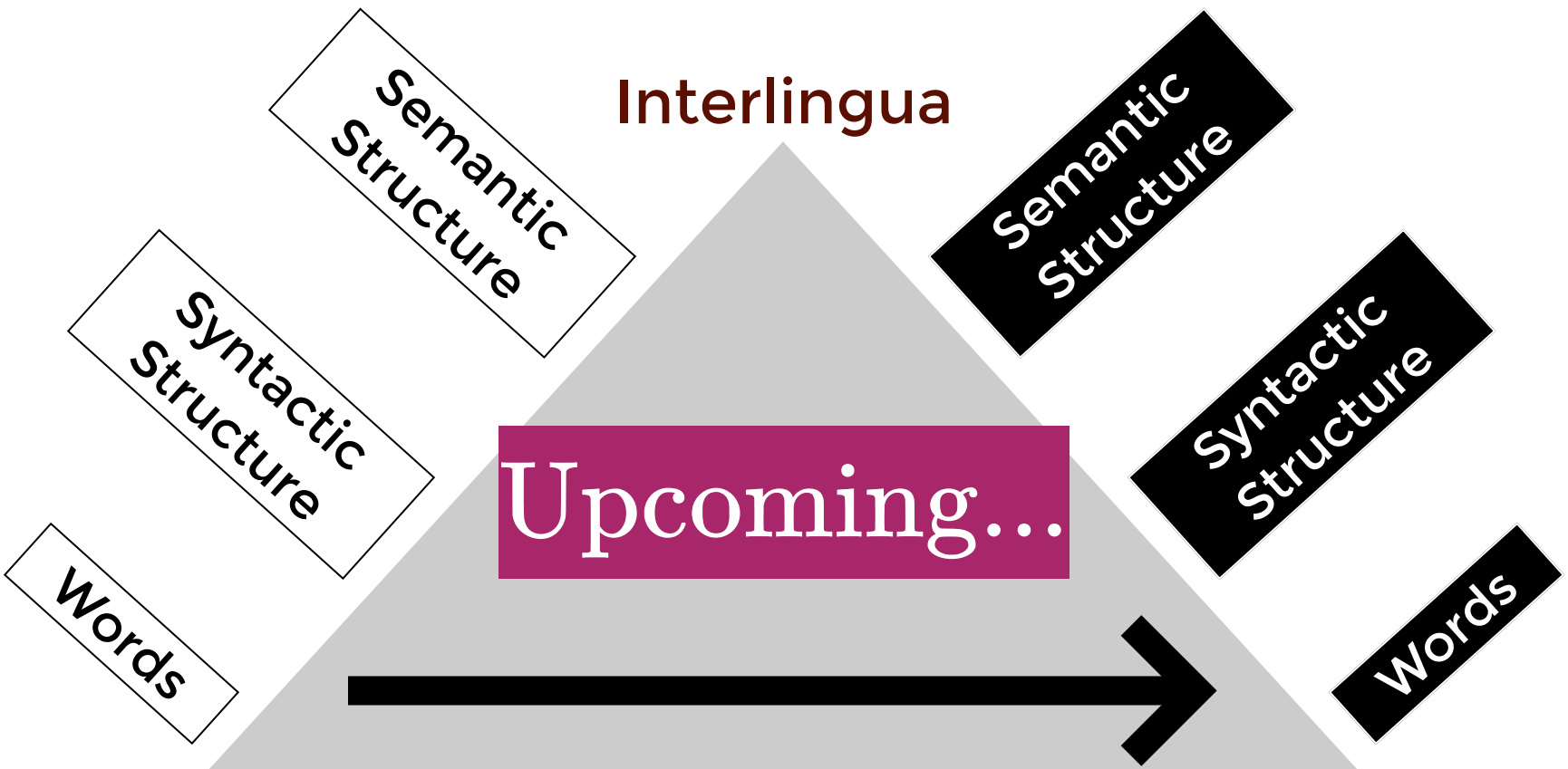
$$+ \quad \lambda_{\text{LM}} \times \underbrace{\log q(e)}_{\text{Language model}}$$

Similarly, we can weight different components differently using hyperparameters. Tune these hyperparameters on the development set, using a similar procedure as phrase-based translation.

Question

Can we make use of word/context embeddings that we learned earlier?

Machine Translation



Text-to-text problem with neural networks