Treebank Transfer

Martin Jansche
jansche@acm.org

Center for Computational Learning Systems
Columbia University

1 Overview

For the vast majority of the world's languages, there are no linguistically annotated data collections. Those can be very costly to produce.

There is a growing interest in methods that would allow us to transfer annotation from one language to another (e.g. work by Yarowsky et al.).

This talk presents a very general technique, based on Bayesian data augmentation, for transferring annotation. The focus is on syntactic annotation, as found in treebanks, and on a novel sampling procedure for distributions over trees.

2 The basic ingredients

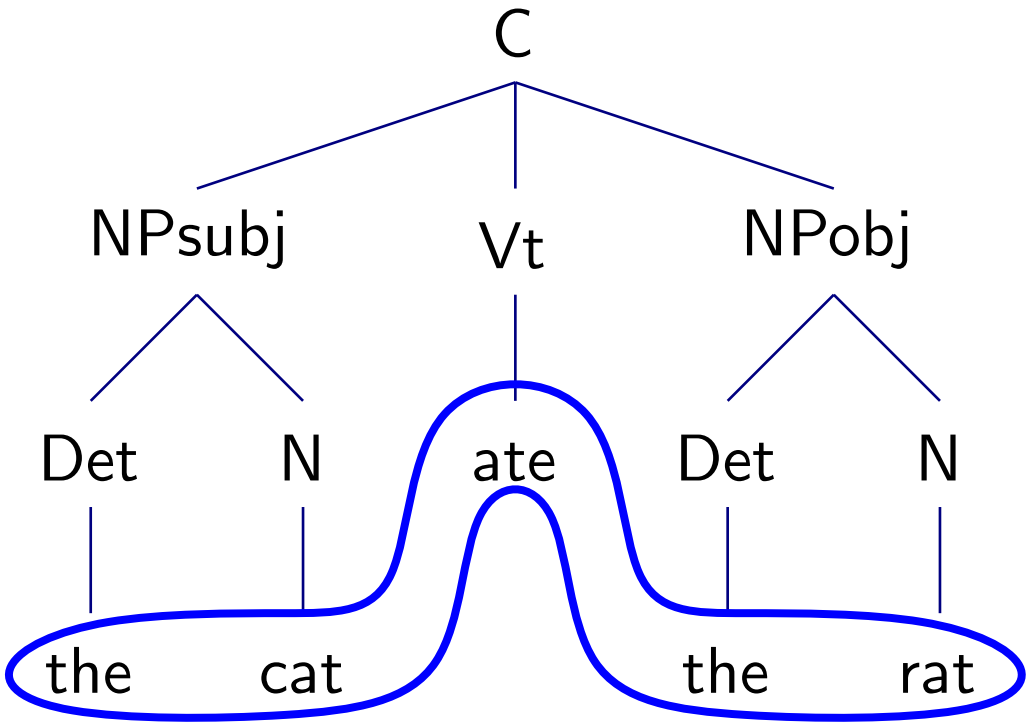
Start with a source-language treebank, i.e., a collection of trees $S_1, \ldots, S_n$. The goal is to generate corresponding target-language trees $T_1, \ldots, T_n$.

Assume that unannotated text for the target language is available, which can be used to build a target-language language model.

Assume a parametric probabilistic mapping from target-language trees to source-language trees. Parameter values of this transfer model need not be known.

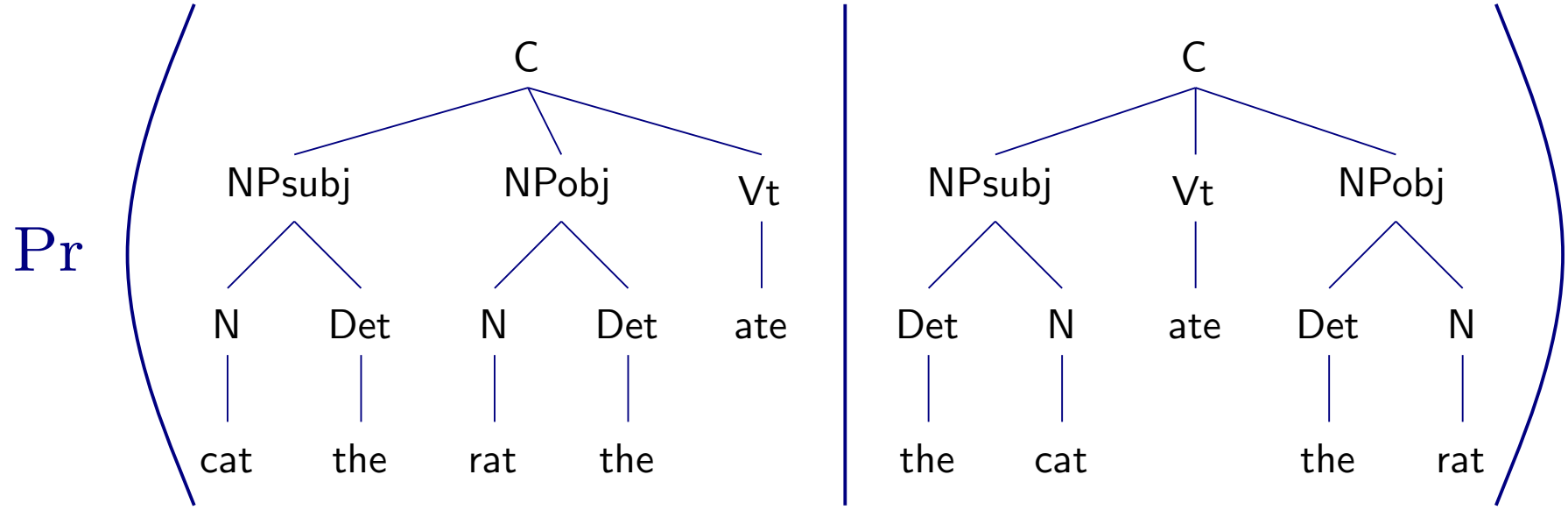
2.1 The target-language language model

Using unannotated target-language text, build a language model for the target language. View it as an impoverished distribution $\Pr(T_i)$ over trees which "forgets" the tree structure and only considers the fringe or terminal yield of a tree.

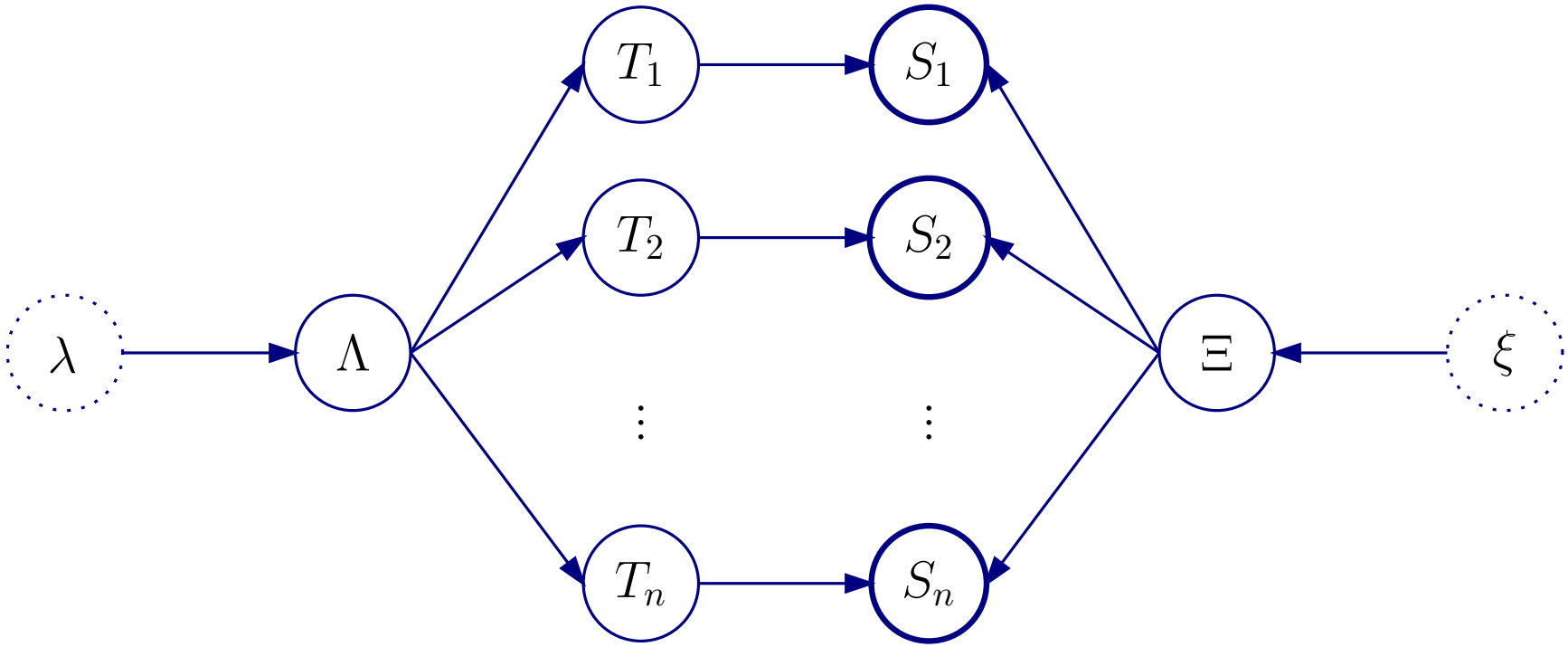


2.2 The transfer model

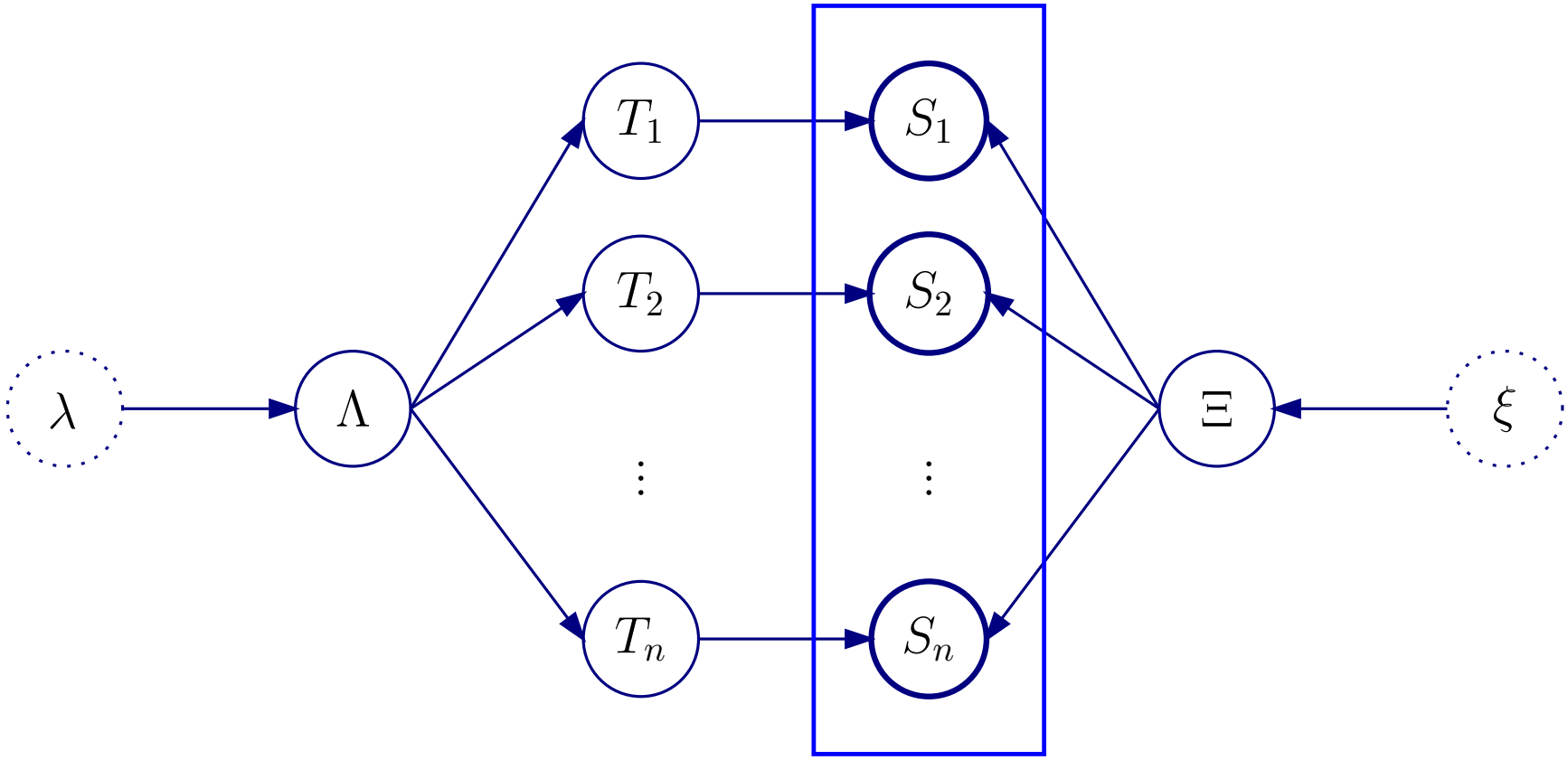
Assume that the parametric form of a probabilistic mapping from target-lg. trees to source-lg. trees has been specified (e.g. in the form of a tree transducer). Let $\Pr(S_i \mid T_i)$ denote the probability of source-language tree S_i corresponding to target-language tree T_i. For example:



3 The probability model

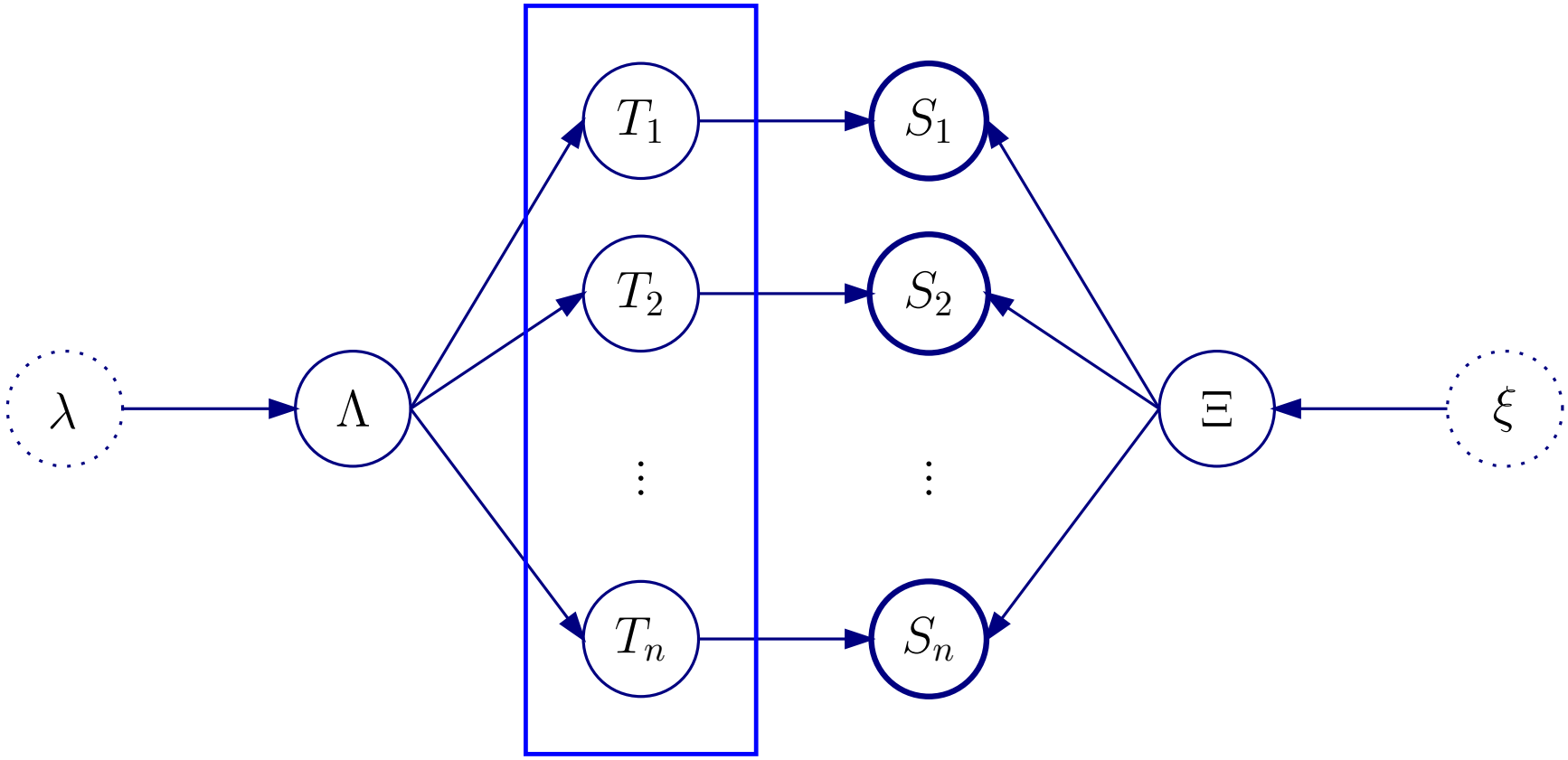


Combine the basic ingredients into a graphical model which specifies a joint distribution over all quantities of interest.



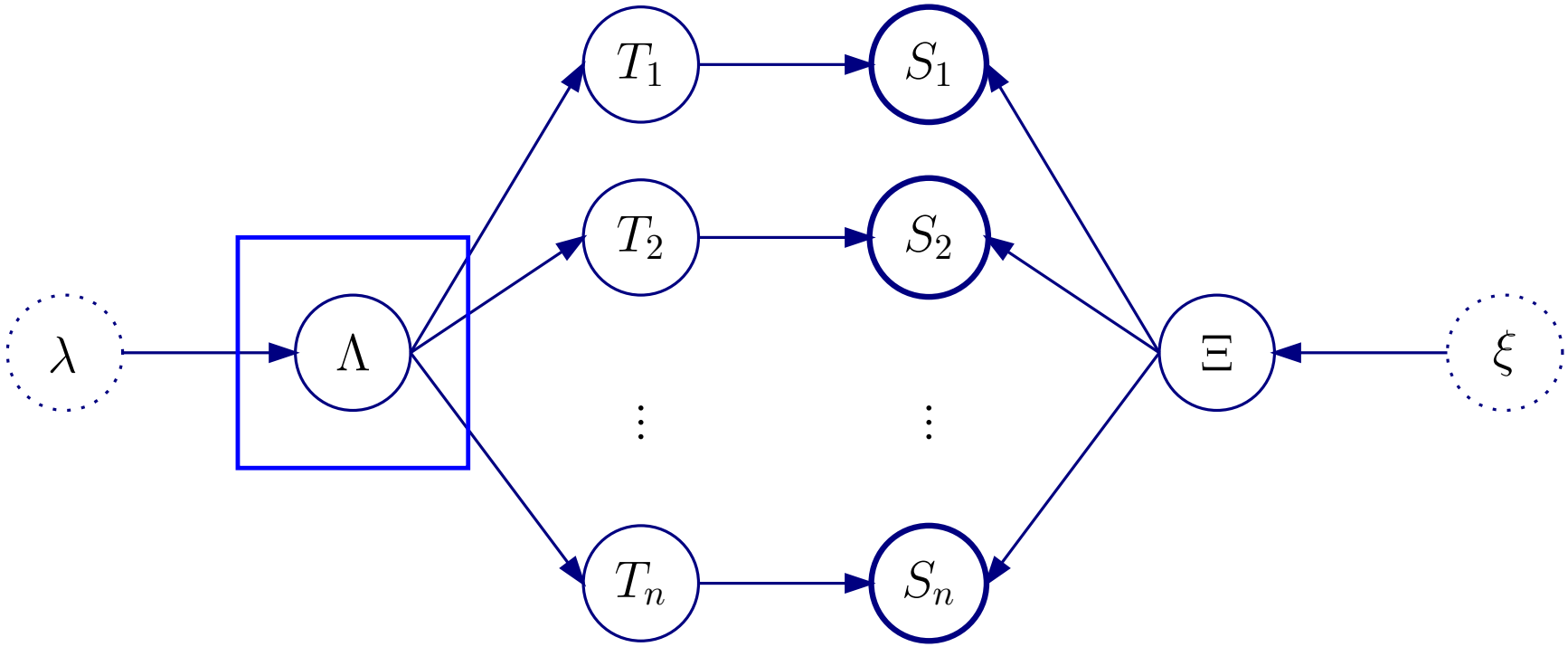
Each S_i is a source-language tree.

Source-language trees are observable and fixed.

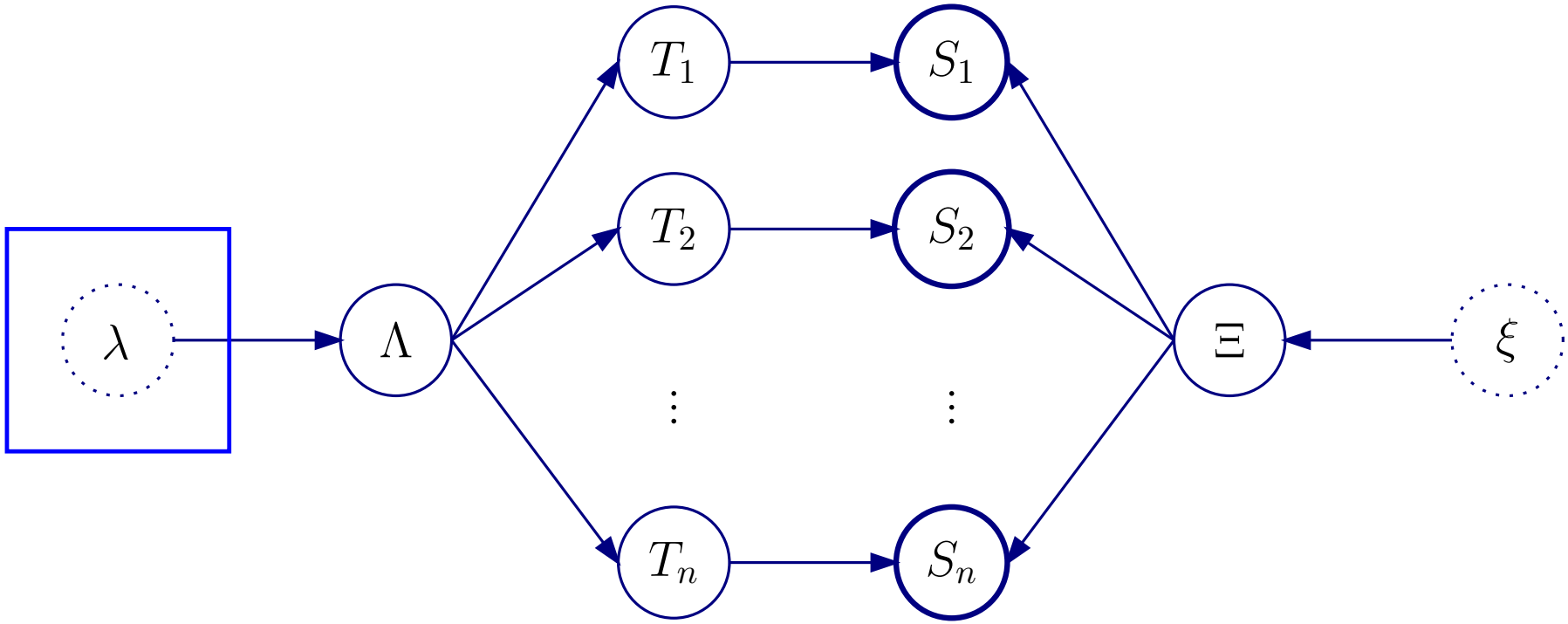


Each T_i is a target-language tree.

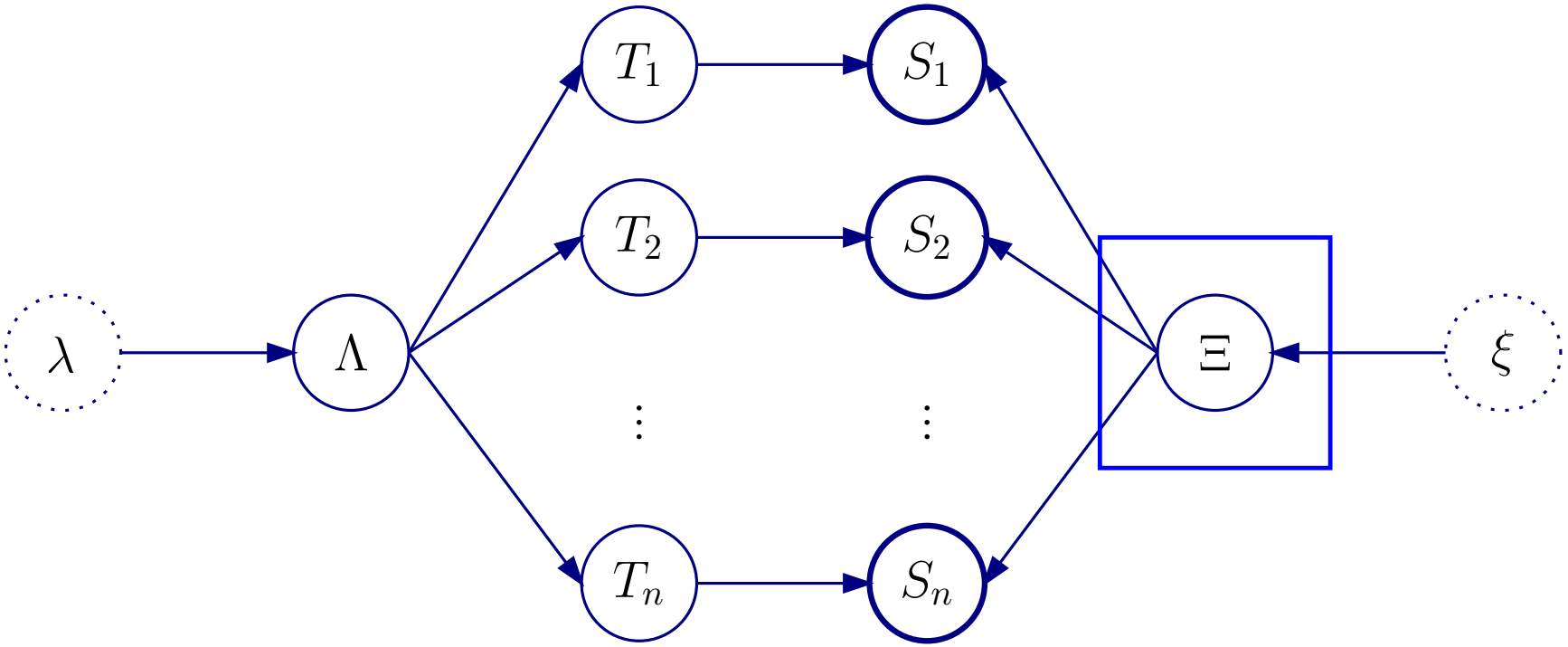
Target-language trees are unobserved.



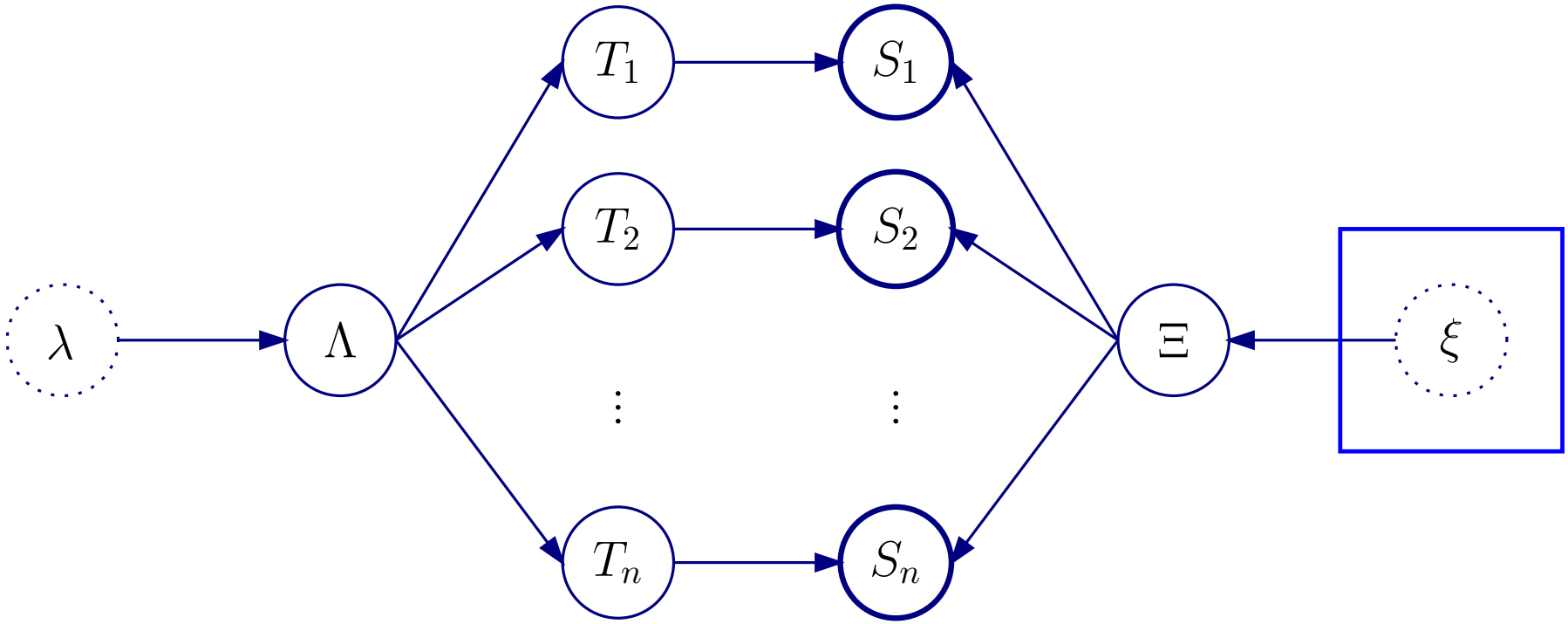
Λ is a vector of target-language language model parameters.



λ is a vector of hyper-parameters for the language model.

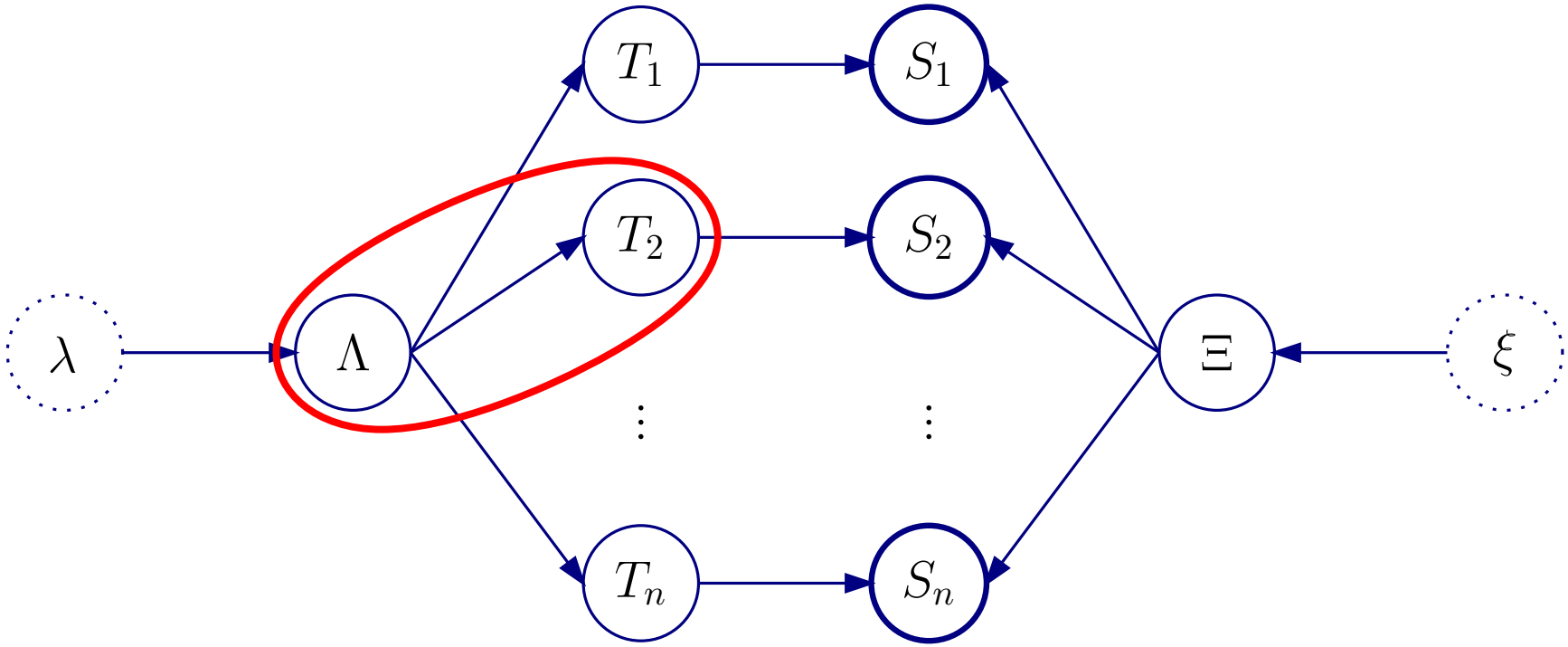


Ξ is a vector of transfer model parameters.



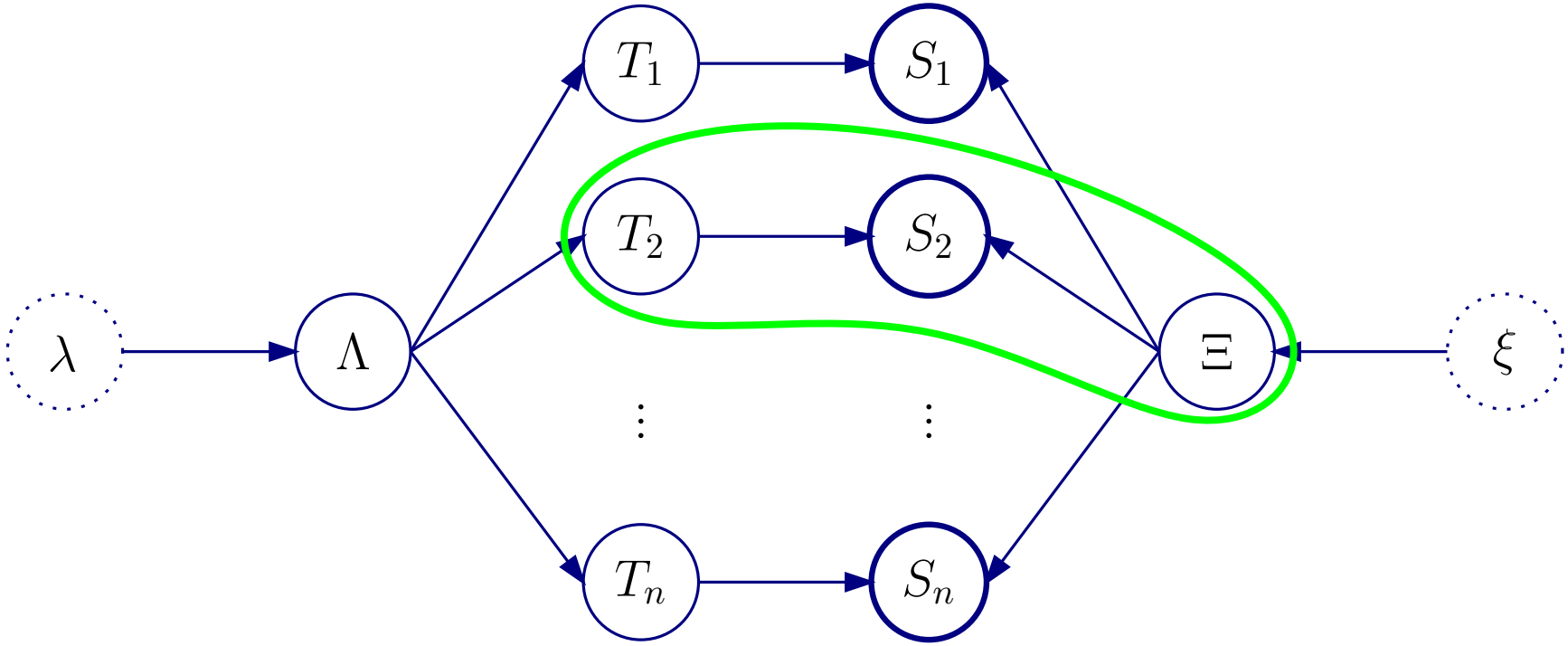
ξ is a vector of hyper-parameters for the transfer model.

3.1 The target-language language model



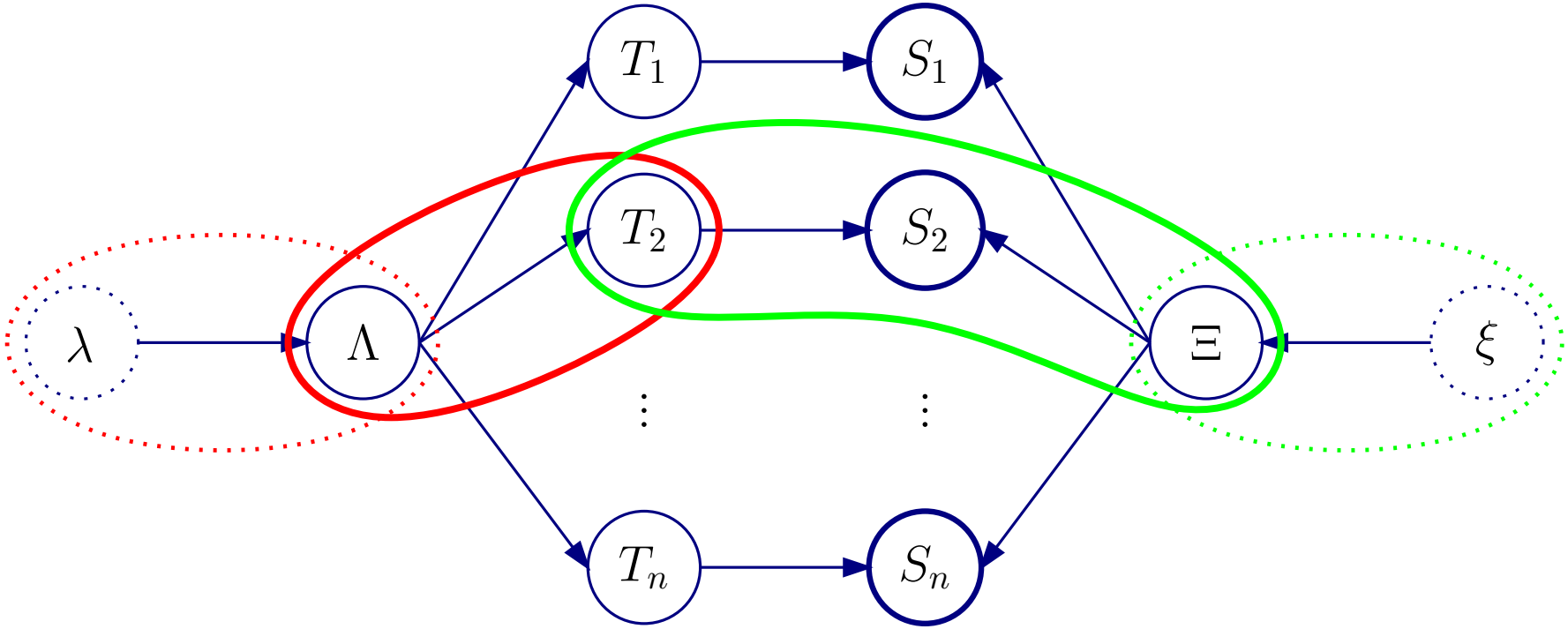
$\Pr(T_i \mid \Lambda)$ is a language model defined over the fringe of T_i. Best to work with simple n-gram models ($n = 2$ for expository purposes).

3.2 The transfer model



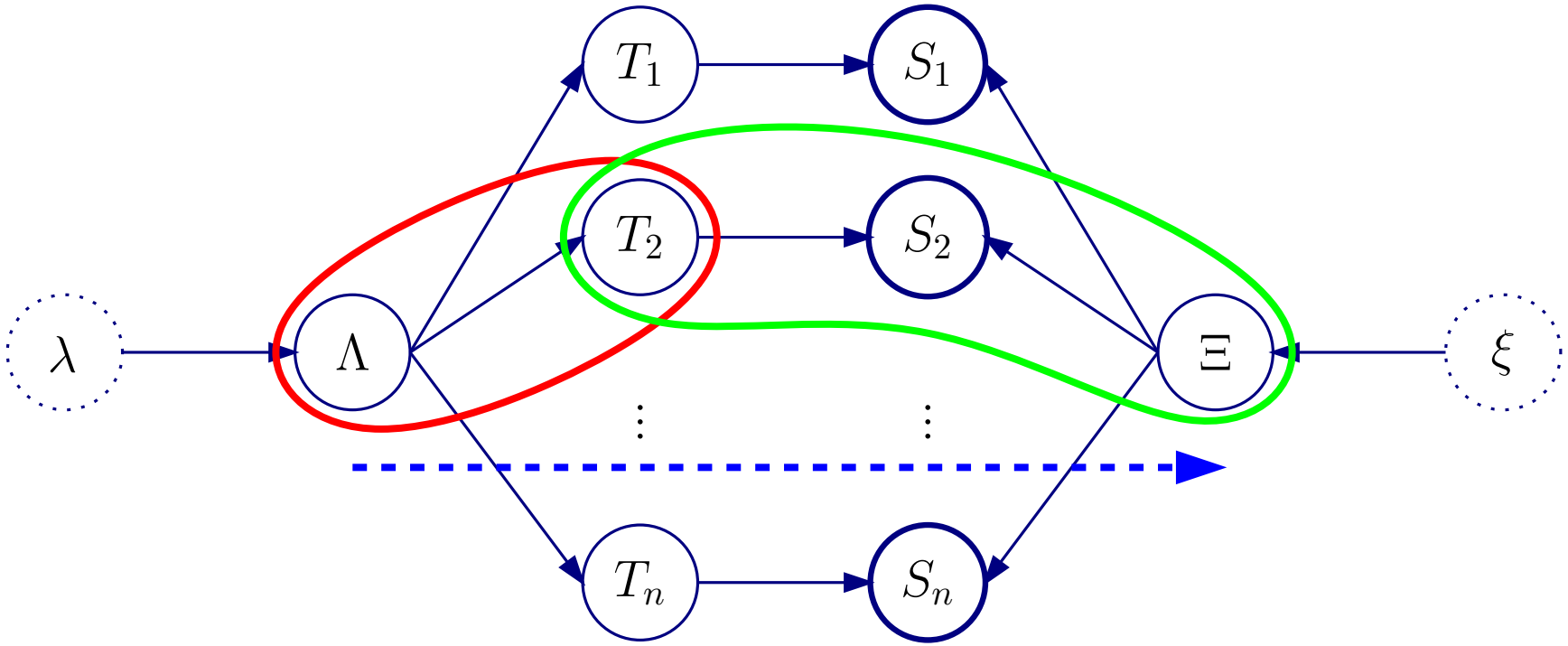
$\Pr(S_i \mid T_i, \Xi)$ is the transfer model applied to the pair (S_i, T_i) of corresponding source-lg. and target-lg. trees.

3.3 The priors for the parameters



The parameter vector Λ of the language language model comes with an informative prior distribution $\Pr(\Lambda \mid \lambda)$. Similarly for the transfer model.

4 Inference and computation



The goal is to transfer information from the structurally weak, but informative language model to the transfer model, and in the process to impute the latent target-language trees.

The goal is

(1) to impute latent target-language trees;

(2) to infer transfer model parameters;

(3) (perhaps) to update language model parameters.

Need to do all of this simultaneously, as parameters and latent data depend on each other.

Inference by Data Augmentation (Tanner & Wong 1987) or Gibbs sampling (Geman & Geman 1984).

Need conditional distributions:

(1) $\Pr(T_i \mid S_i, \Xi, \Lambda)$
(to impute latent target-language trees);

(2) $\Pr(\Xi \mid S_1, \ldots, S_n, T_1, \ldots, T_n, \xi)$
(to infer transfer model parameters);

(3) $\Pr(\Lambda \mid T_1, \ldots, T_n, \lambda)$
(to update language model parameters).

Gibbs sampling proceeds as follows. Initialize Λ and Ξ by drawing each from (a variant of) its respective prior distribution. Then iterate the following steps:

(1) draw each T_i from its posterior distribution given S_i, Λ, and Ξ
(to impute latent target-language trees);

(2) draw Ξ from its posterior distribution given $(S_1, T_1), \ldots, (S_n, T_n)$ and ξ
(to infer transfer model parameters);

(3) draw Λ from its posterior distribution given $T_1, \ldots, T_n$ and λ
(to update language model parameters).

4.1 Language model posterior

For simple n-gram language models, $\Pr(T_i \mid \Lambda)$ takes the form of a product of multivariate Bernoulli distributions.

For example: $\Pr(\#\text{ the cat ate the rat \$} \mid \Lambda) =$

$$
\begin{array}{rccccc}
 & \text{the} & \text{cat} & \text{ate} & \text{rat} & \$ \\
\Pr(& 1 & 0 & 0 & 0 & 0 \mid \Lambda_{\#}) \\
\times \Pr(& 0 & 1 & 0 & 0 & 0 \mid \Lambda_{\text{the}}) \\
\times \Pr(& 0 & 0 & 1 & 0 & 0 \mid \Lambda_{\text{cat}}) \\
\times \Pr(& 1 & 0 & 0 & 0 & 0 \mid \Lambda_{\text{ate}}) \\
\times \Pr(& 0 & 0 & 0 & 1 & 0 \mid \Lambda_{\text{the}}) \\
\times \Pr(& 0 & 0 & 0 & 0 & 1 \mid \Lambda_{\text{rat}})
\end{array}
$$

Therefore

$$\Pr(\Lambda \mid T_1, \ldots, T_n, \lambda)$$

takes the form of a product of Dirichlet distributions (with parameters derived from λ plus n-gram counts along the fringes of the T_i).

Generating random variates from a Dirichlet distribution is well understood (Gelman et al. 1995; Devroye 1986).

4.2 Transfer model posterior

We similarly assume that $\Pr(S_i \mid T_i, \Xi)$ involves making one or more independent, discrete choices among fixed sets of options for transforming the trees. Hence this too takes the form of a product of multinomial distributions.

Therefore

$$\Pr(\Xi \mid S_1, \ldots, S_n, T_1, \ldots, T_n, \xi)$$

also takes the form of a product of Dirichlet distributions (with parameters derived from ξ plus counts of correspondences between the (S_i, T_i) pairs).

4.3 Latent tree posterior

Sampling from the latent tree posterior

$$\Pr(T_i \mid S_i, \Xi, \Lambda) \propto \Pr(T_i \mid \Lambda) \times \Pr(S_i \mid T_i, \Xi)$$

is the tricky part, requiring a novel solution.

Notice that rejection sampling using the prior $\Pr(T_i \mid \Lambda)$ as the proposal distribution will not work, because the prior is much too weak.

The likelihood cannot be used directly for rejection sampling because it is not a properly normalized probability distribution.

An important assumption: The likelihood function

$$T_i \mapsto \Pr(S_i \mid T_i, \Xi)$$

can be represented compactly as a packed forest (finite CFG).

This means that it can be normalized efficiently, without explicitly summing over potentially exponentially many latent trees with non-zero likelihood.

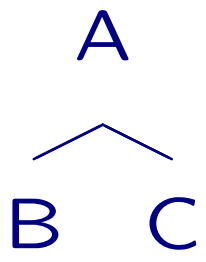
Thus we could use the normalized likelihood as the proposal distribution in a rejection sampling scheme.

But wait, there is more!

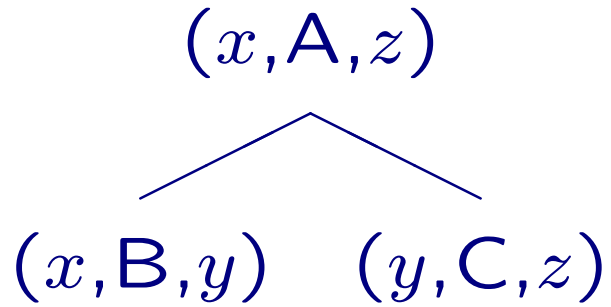
It is also possible to compute a compact representation of the posterior directly, by intersecting the compact representation of the likelihood with the language model prior.

This is a special case of CFG intersection (Bar-Hillel et al. 1961). Related to forest-based generation (Langkilde 2000), which involves searching for the best tree in this kind of intersection (without fully expanding it).

The idea is to replace a forest node of the form



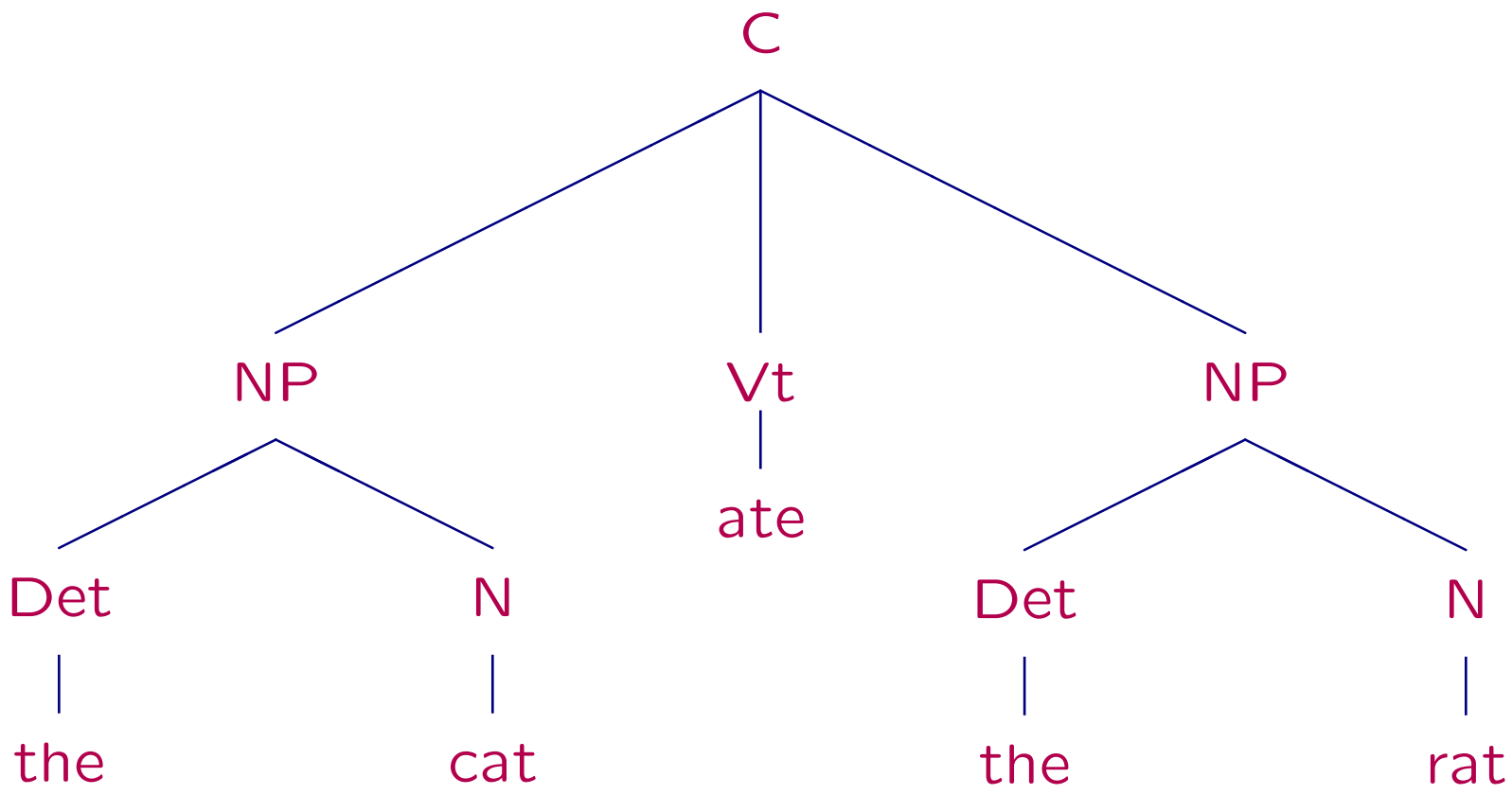
with a node of the form

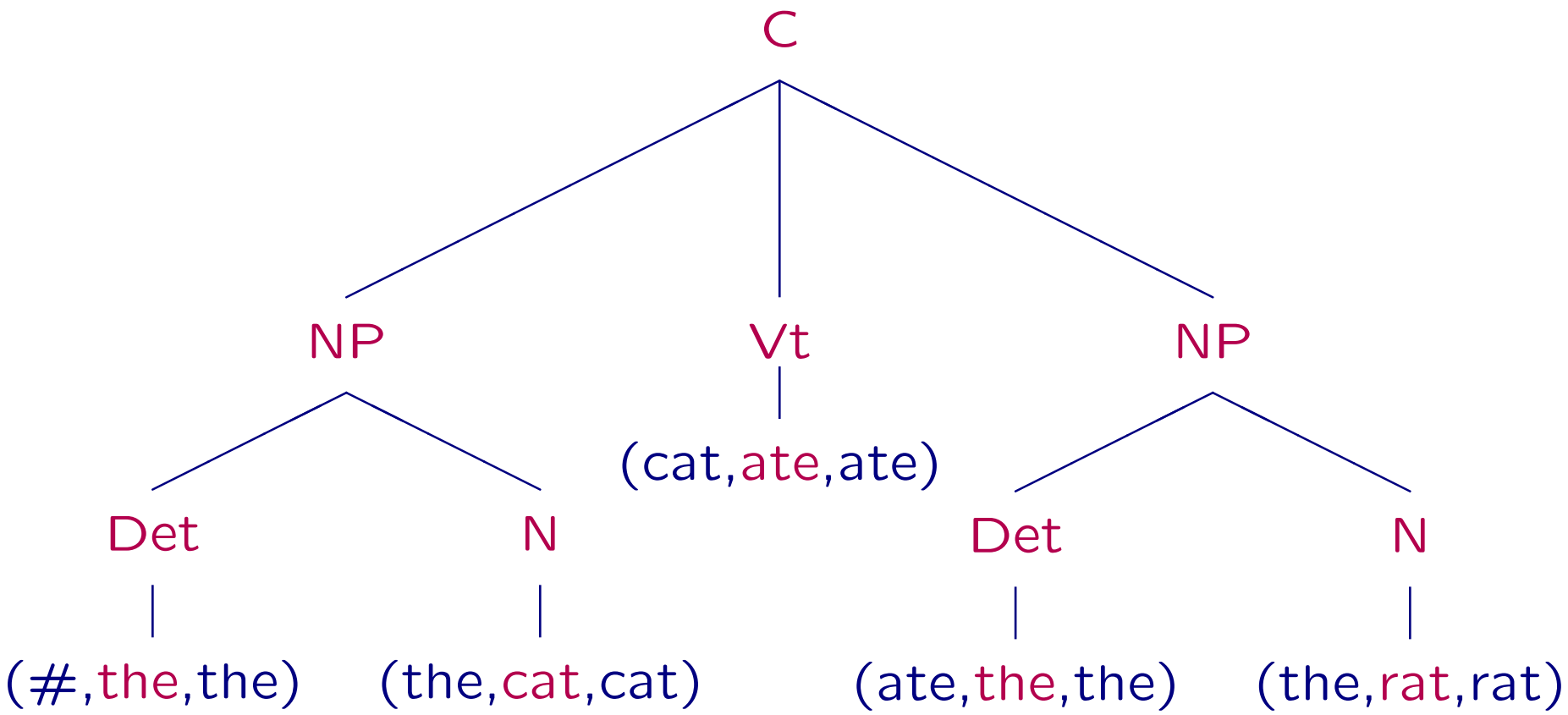


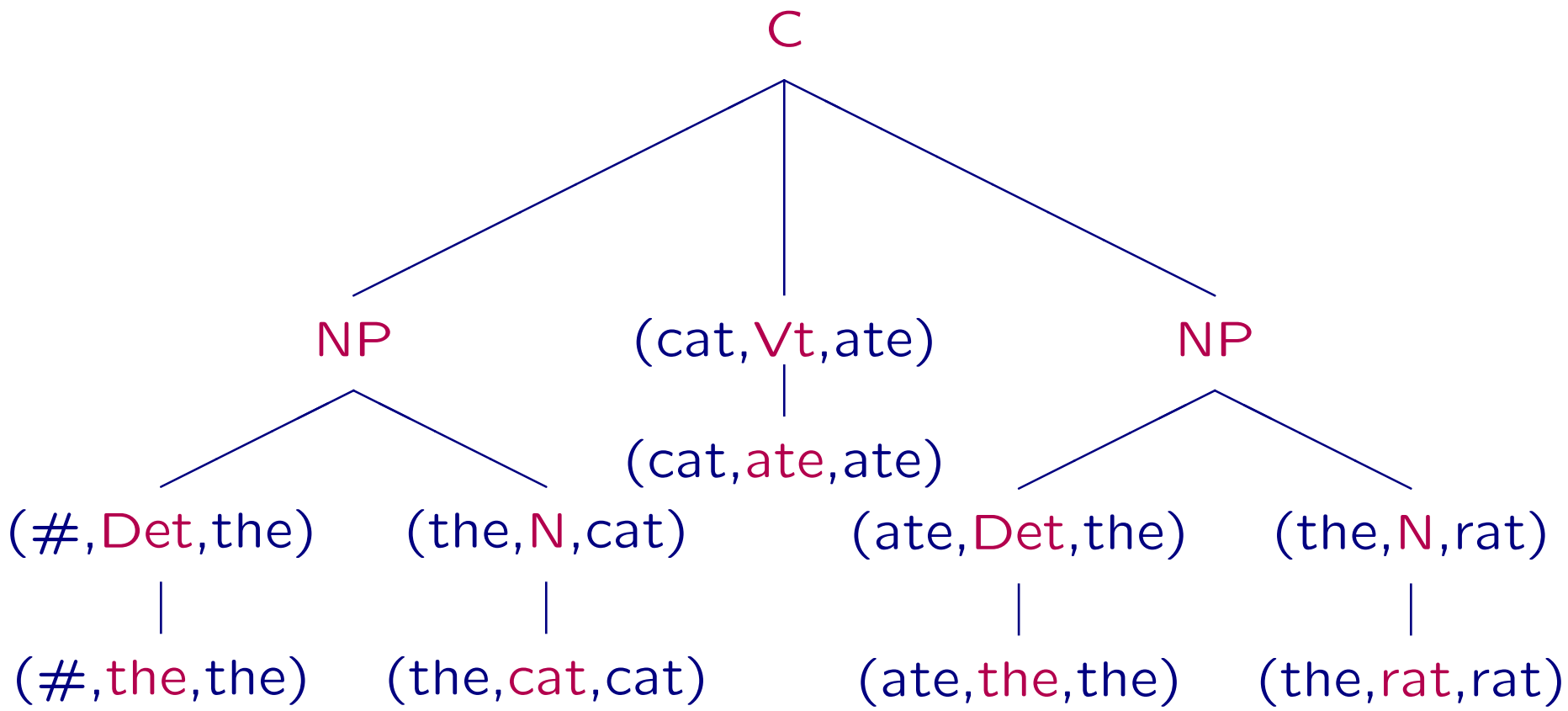
where x, y, and z are terminal symbols.

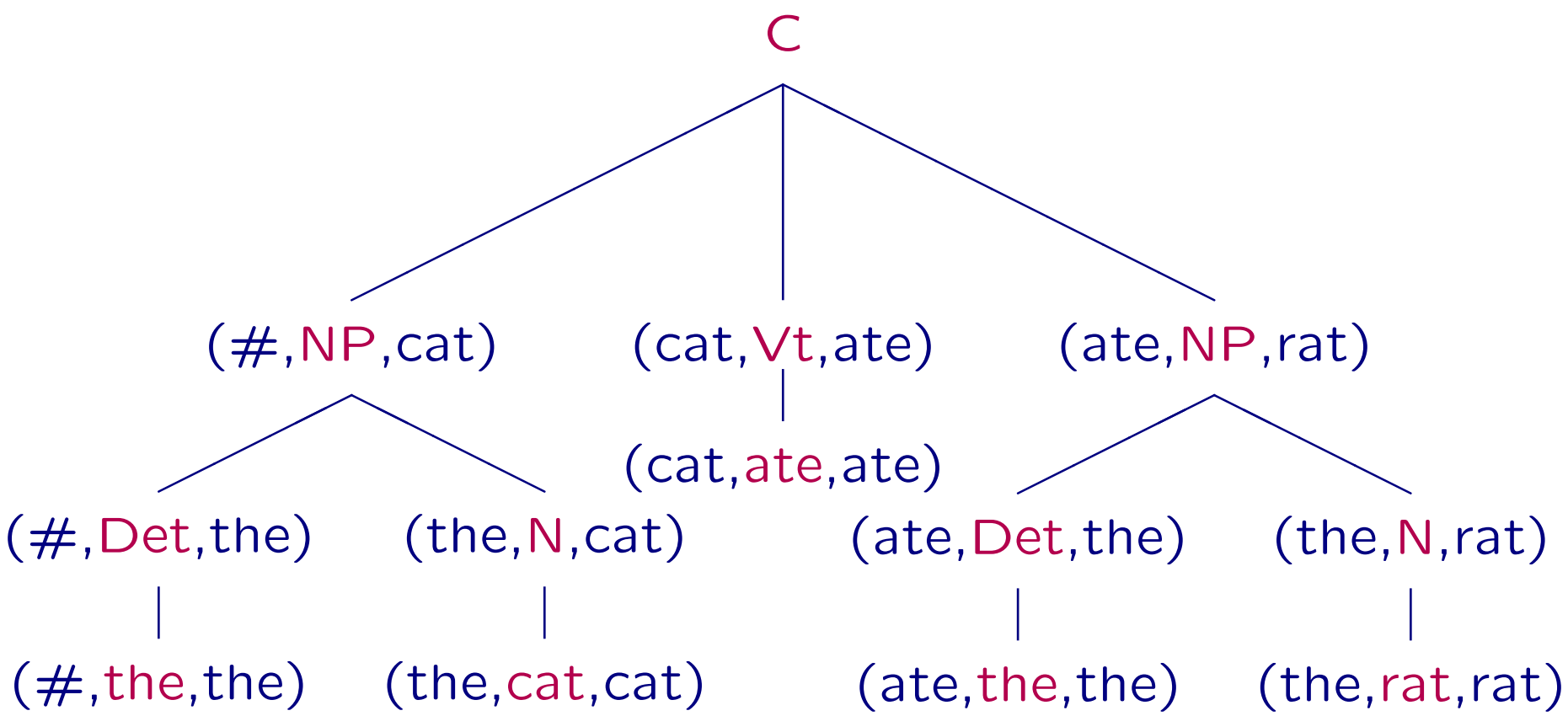
Replace a leaf node y with a triple (x, y, y).

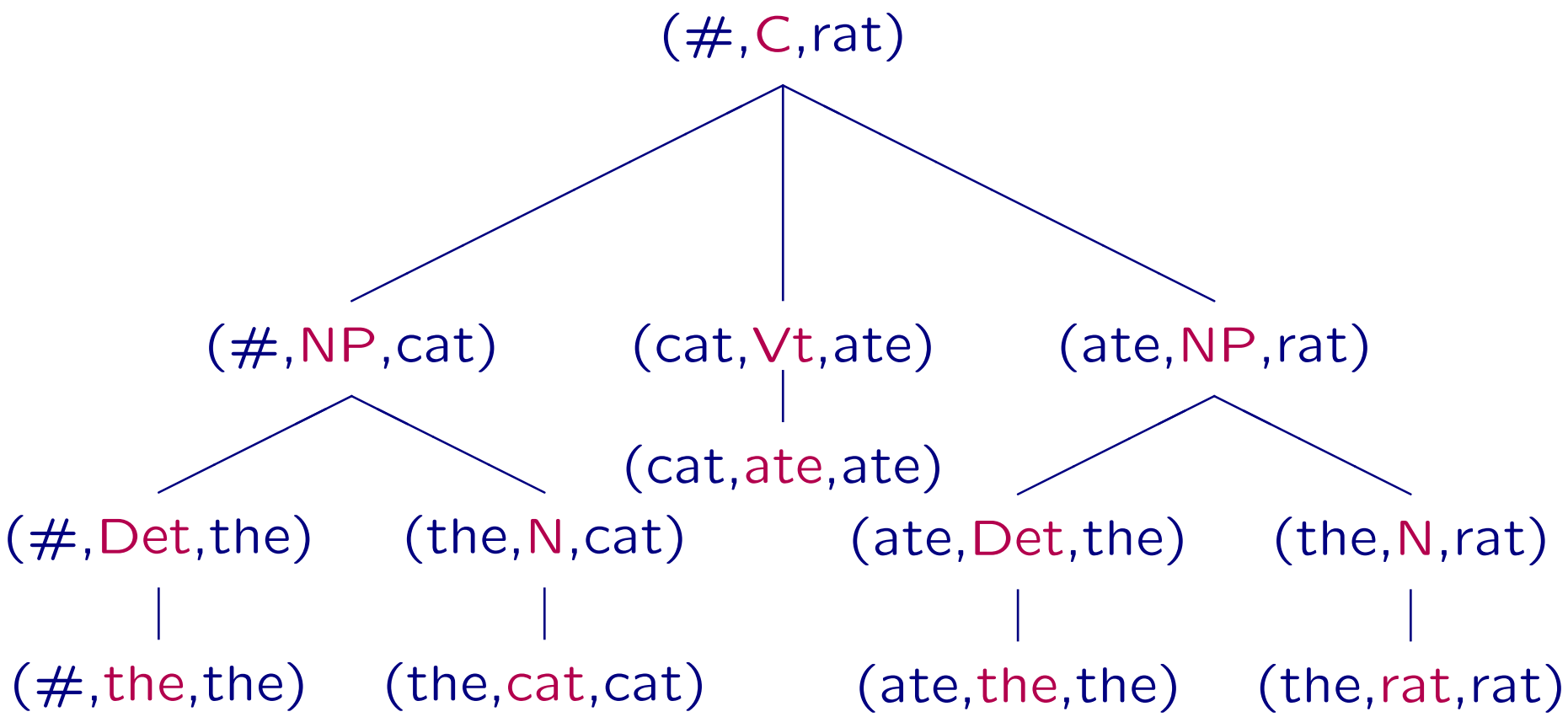
Illustrate this using a singleton forest:

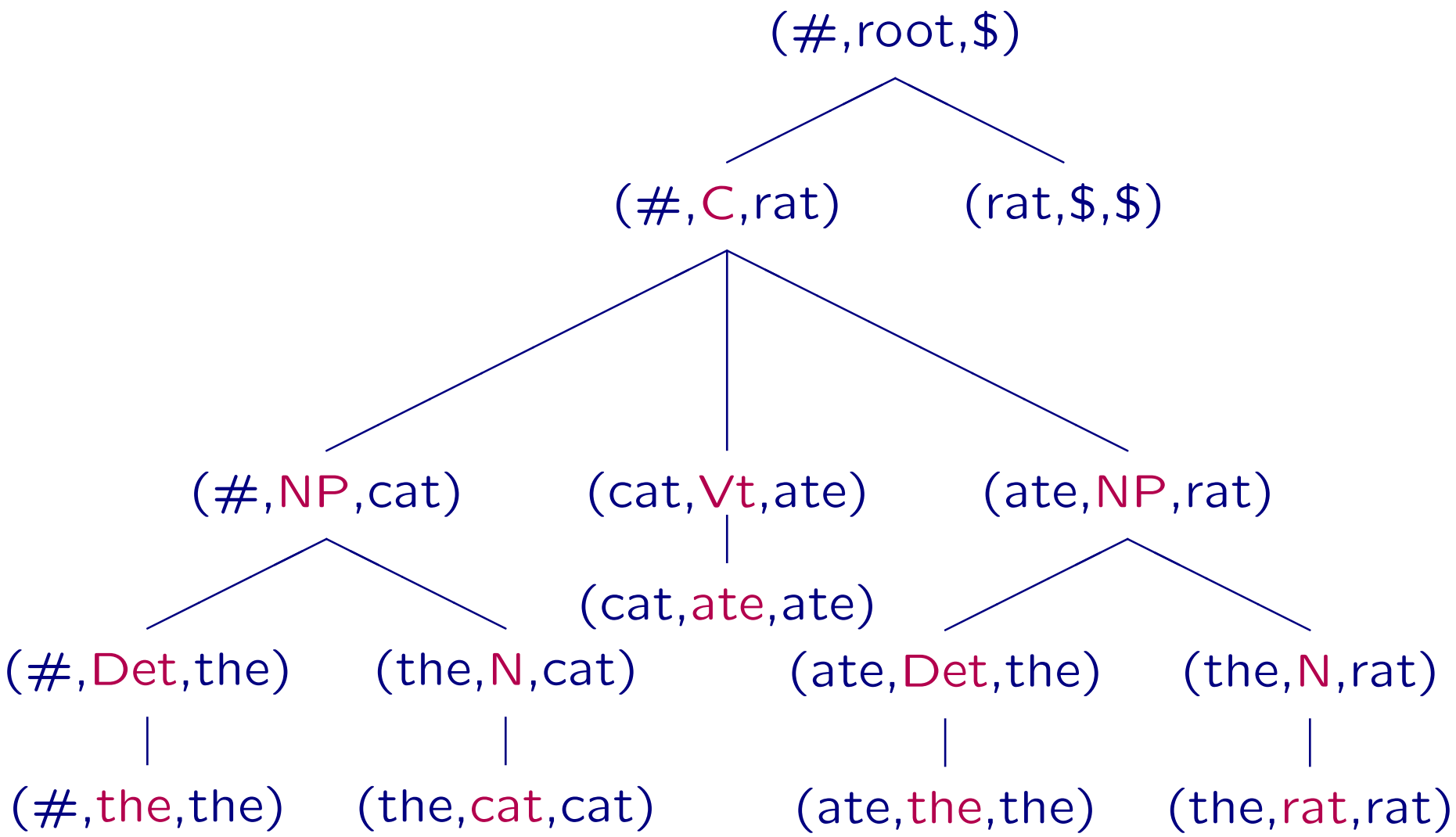



C
NP
Vt
NP
(cat,ate,ate)
Det
N
Det
N
(#,the,the)
(the,cat,cat)
(ate,the,the)
(the,rat,rat)


C
NP
(cat,Vt,ate)
NP
(cat,ate,ate)
(#,Det,the)
(the,N,cat)
(ate,Det,the)
(the,N,rat)
(#,the,the)
(the,cat,cat)
(ate,the,the)
(the,rat,rat)


C
(#,NP,cat)
(cat,Vt,ate)
(ate,NP,rat)
(cat,ate,ate)
(#,Det,the)
(the,N,cat)
(ate,Det,the)
(the,N,rat)
(#,the,the)
(the,cat,cat)
(ate,the,the)
(the,rat,rat)


(#,C,rat)
(#,NP,cat)
(cat,Vt,ate)
(ate,NP,rat)
(cat,ate,ate)
(#,Det,the)
(the,N,cat)
(ate,Det,the)
(the,N,rat)
(#,the,the)
(the,cat,cat)
(ate,the,the)
(the,rat,rat)


(#,root,$)
(#,C,rat)
(rat,$,$)
(#,NP,cat)
(cat,Vt,ate)
(ate,NP,rat)
(cat,ate,ate)
(#,Det,the)
(the,N,cat)
(ate,Det,the)
(the,N,rat)
(#,the,the)
(the,cat,cat)
(ate,the,the)
(the,rat,rat)

When doing this on a non-trivial forest (an example can be found in the paper), this process will split existing forest nodes into multiple nodes. The number of nodes in the forest will grow at most polynomially, i.e., the forest will remain compact.

Combine the weights of the forest (likelihood) with the probabilities from the language model (read off the leaves). Normalize the forest to obtain the latent tree posterior.

5 Conclusions

Treebank transfer can be naturally viewed as a missing data problem. The problem is constrained from two sides:

(1) an informative prior defined over the fringes of target-lg. trees exploits unstructured text and provides strong evidence for or against certain kinds of trees;

(2) at the same times, the latent target-lg. trees must correspond to observed source-lg. trees in a systematic fashion specified by the transfer model.

Because of the uncertainty in the transfer model, and because the prior over latent trees is structurally weak, use Gibbs sampling (as opposed to EM) for a better understanding of remaining uncertainty.

By running many iterations of the Gibbs sampler after convergence, one can simulate the marginal posterior distribution of the latent target-lg. trees and obtain multiple imputations (as opposed to finding only a single instantiation of each T_i).

The method presented here is very general, and many assumptions can in fact be relaxed. All that is required is that the latent tree likelihood under the transfer model have a compact representation, and that composition/intersection with the prior on trees preserve compactness. Then efficient direct sampling from the latent tree posterior is possible.

Acknowledgments

This work was partly supported by Columbia University's Fu Foundation School of Engineering and Applied Science. All opinions expressed here are those of the author.

Thanks to Steve Abney, the participants of the 2005 Johns Hopkins workshop on Arabic dialect parsing, and the anonymous reviewers for helpful discussions.

Thanks to Bob Carpenter for presenting this talk on my behalf.