Generation in Machine Translation from Deep Syntactic Trees

Keith Hall
Johns Hopkins University

Petr Němec
Charles University in Prague

Outline

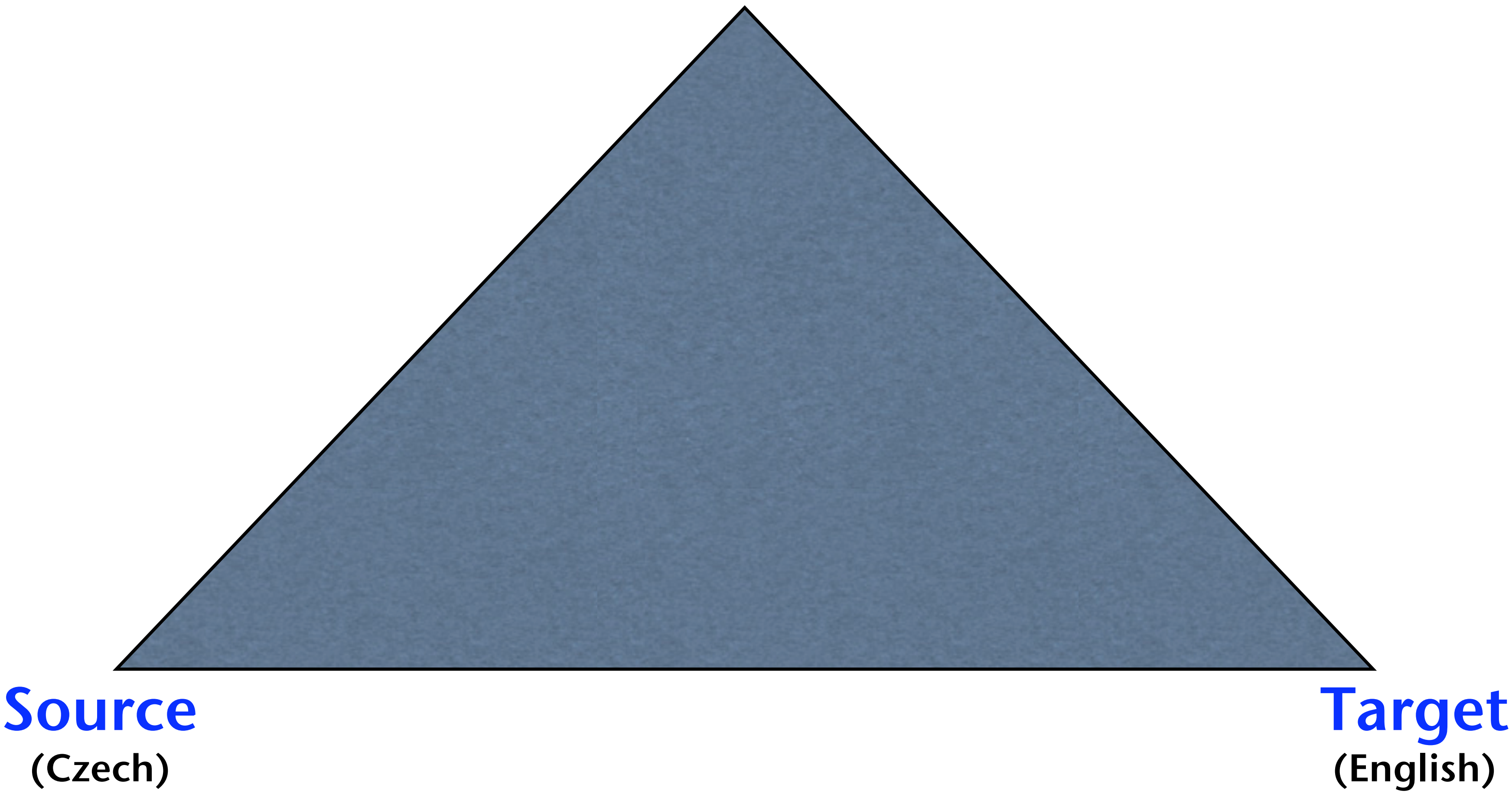
- Transfer-based MT
- Tectogrammatical Representation (TR) (deep syntax)
- Generation from English TR trees
  - process
  - models
- Empirical results

Transfer-based MT

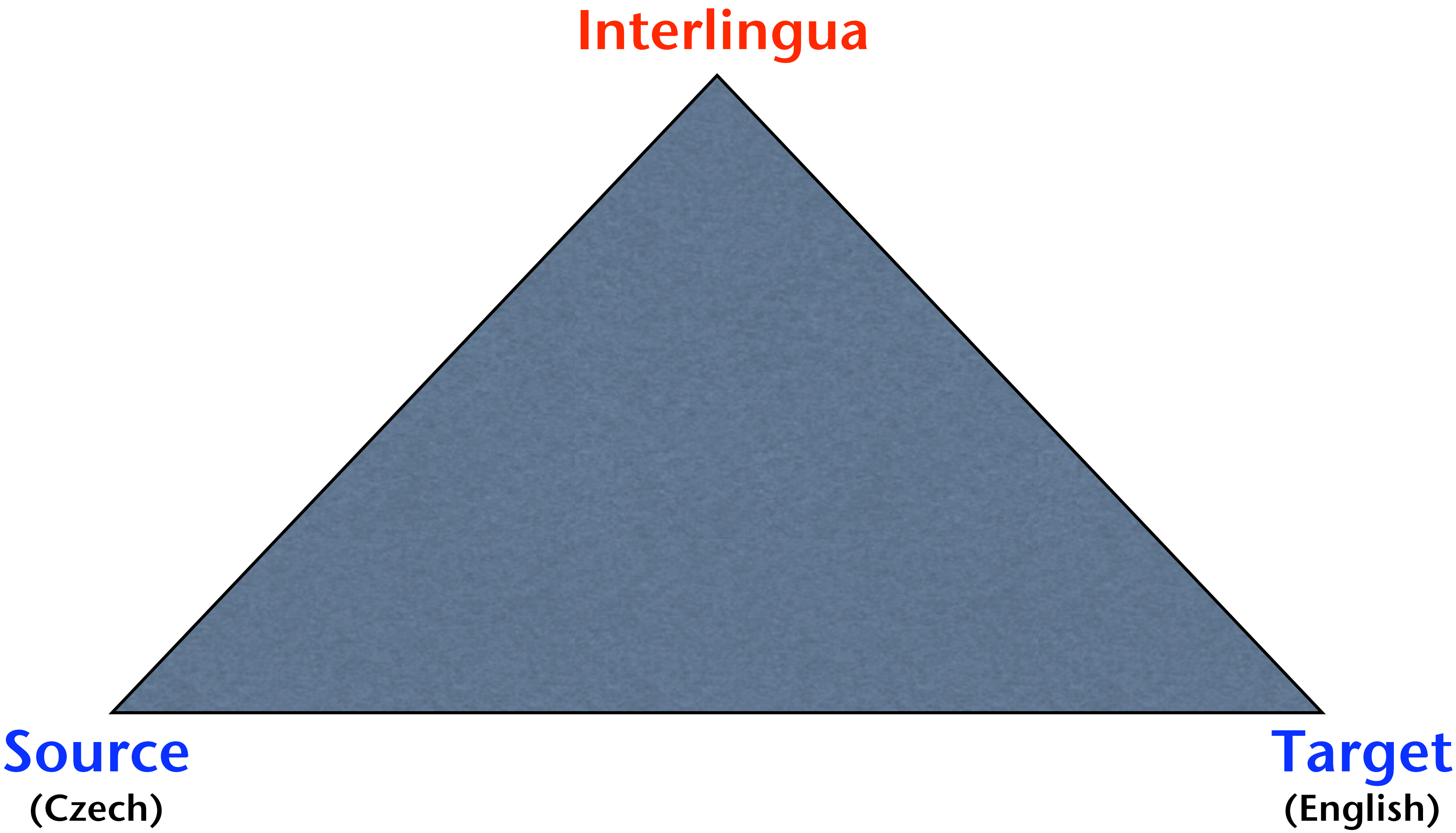
Source
(Czech)

Target
(English)

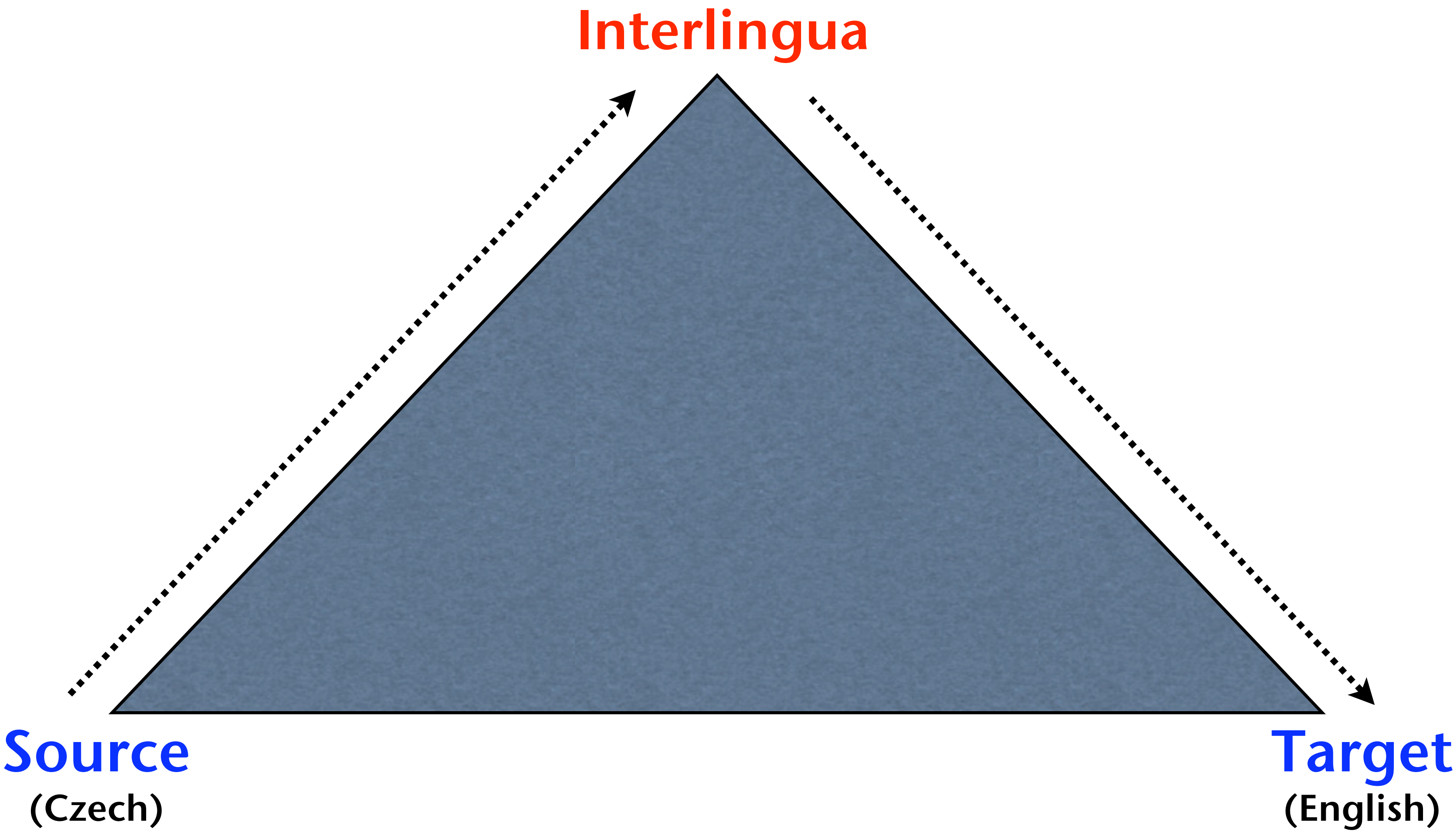
Transfer-based MT



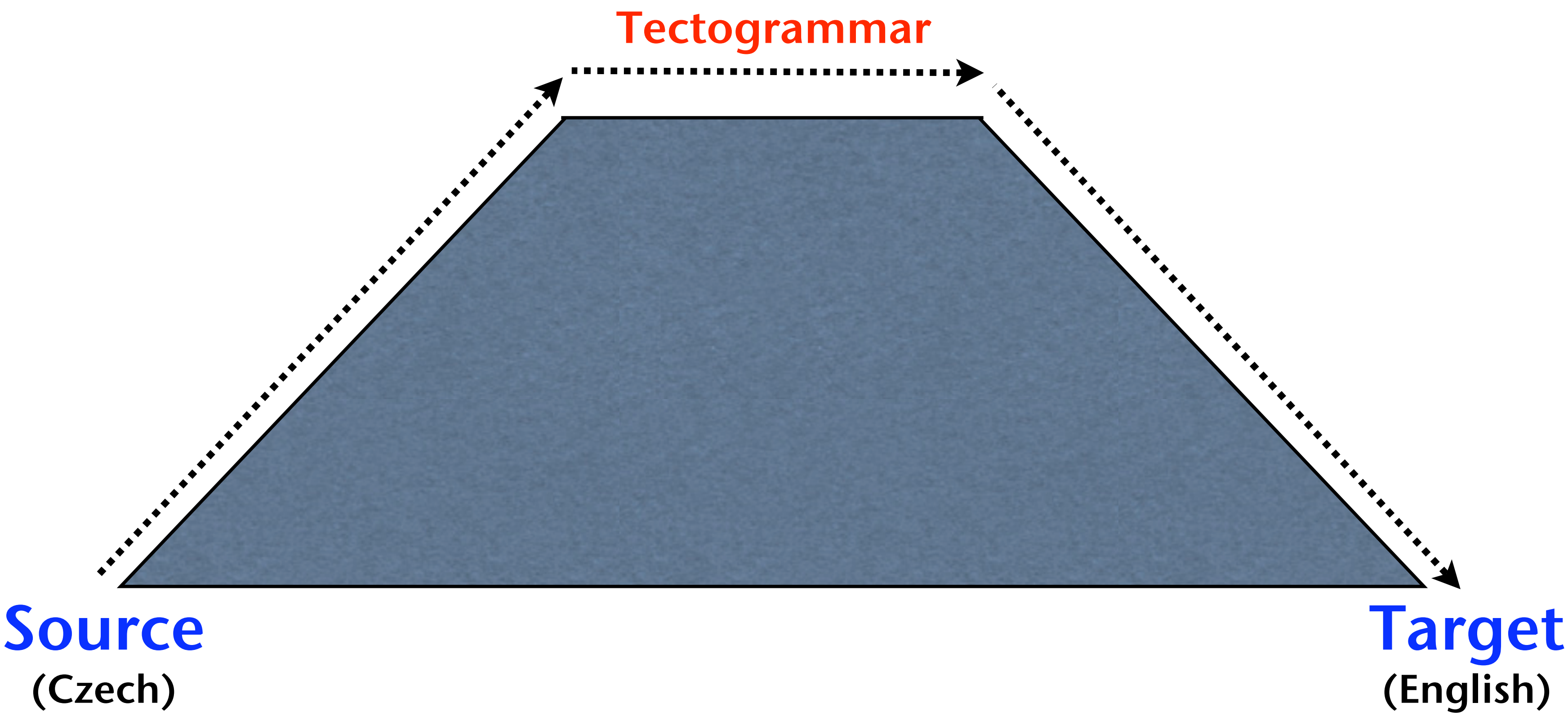
Transfer-based MT



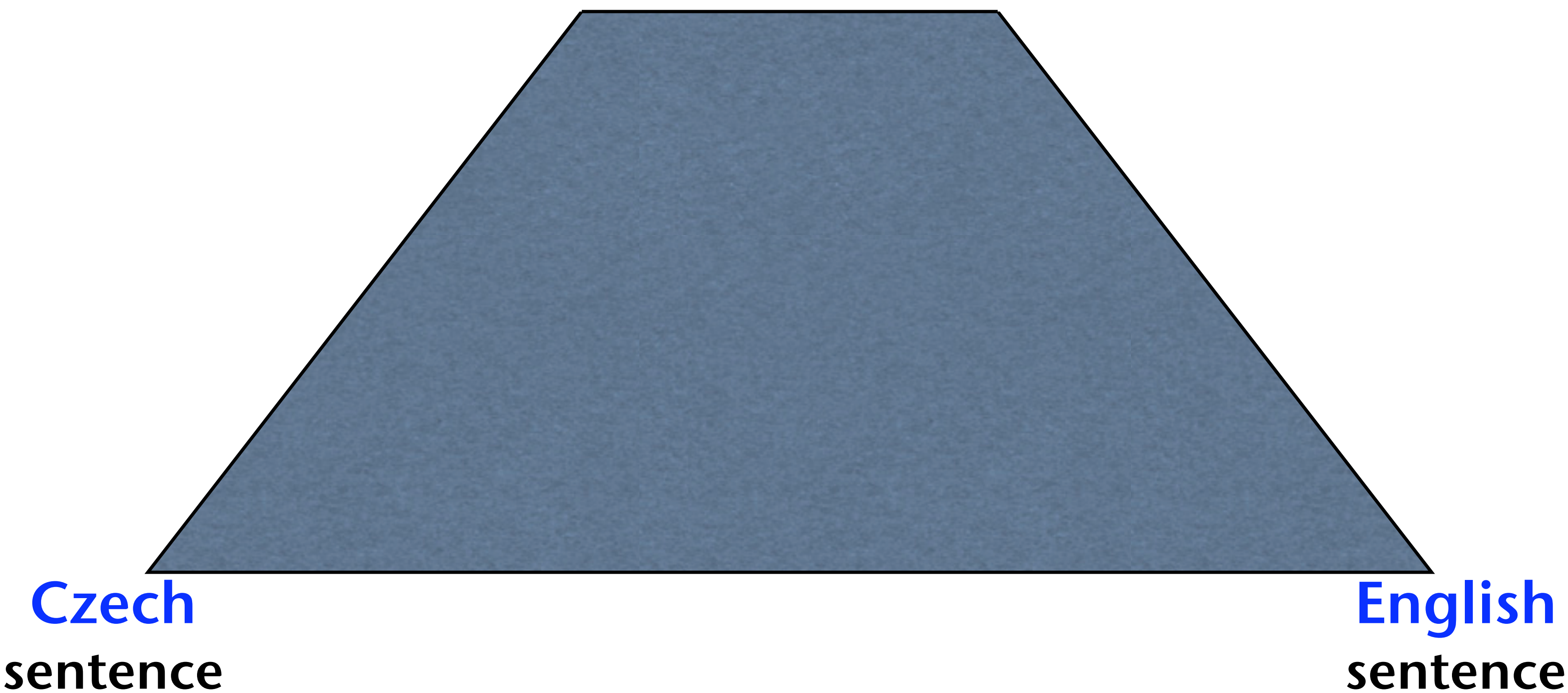
Transfer-based MT



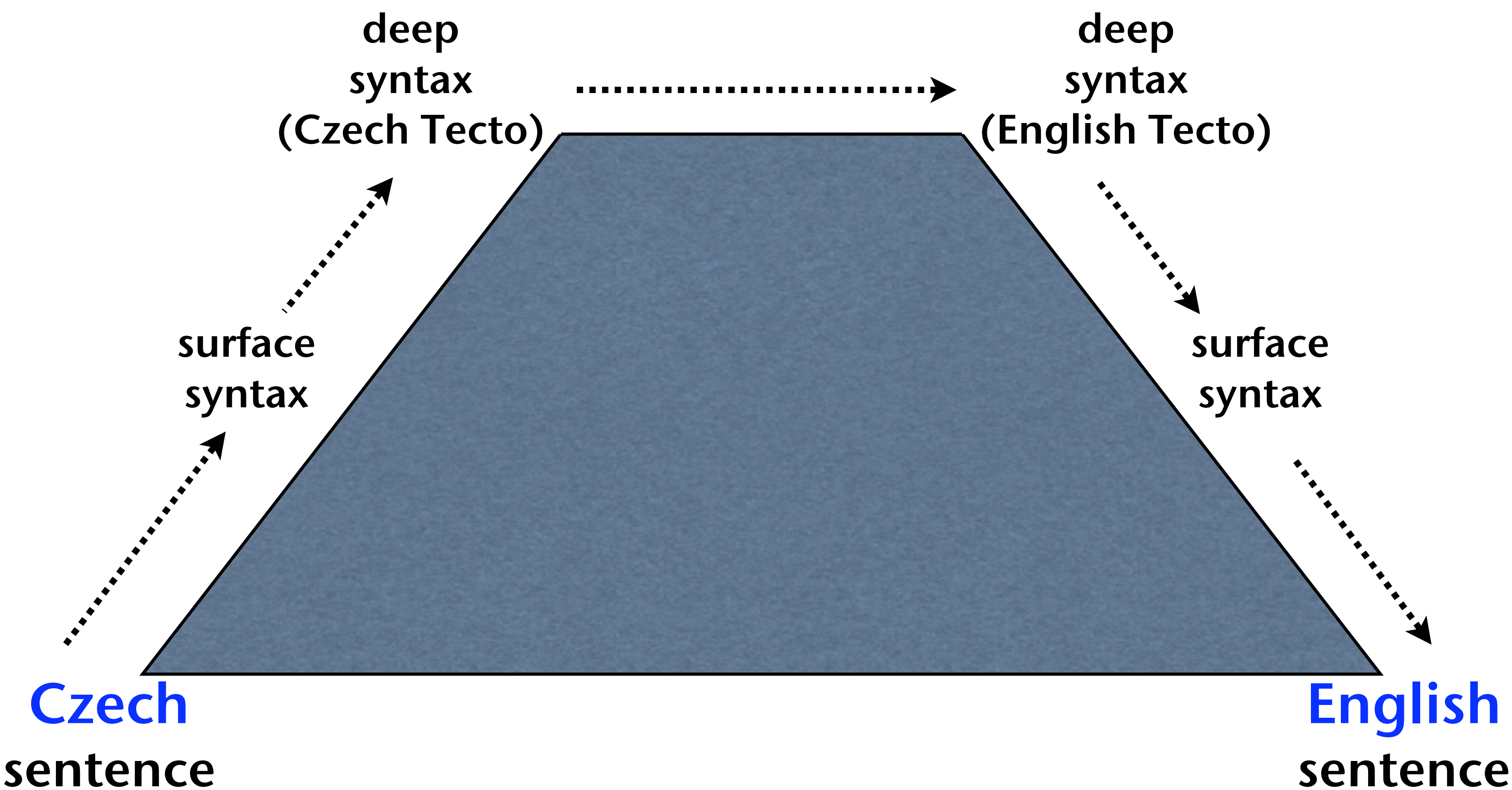
Transfer-based MT



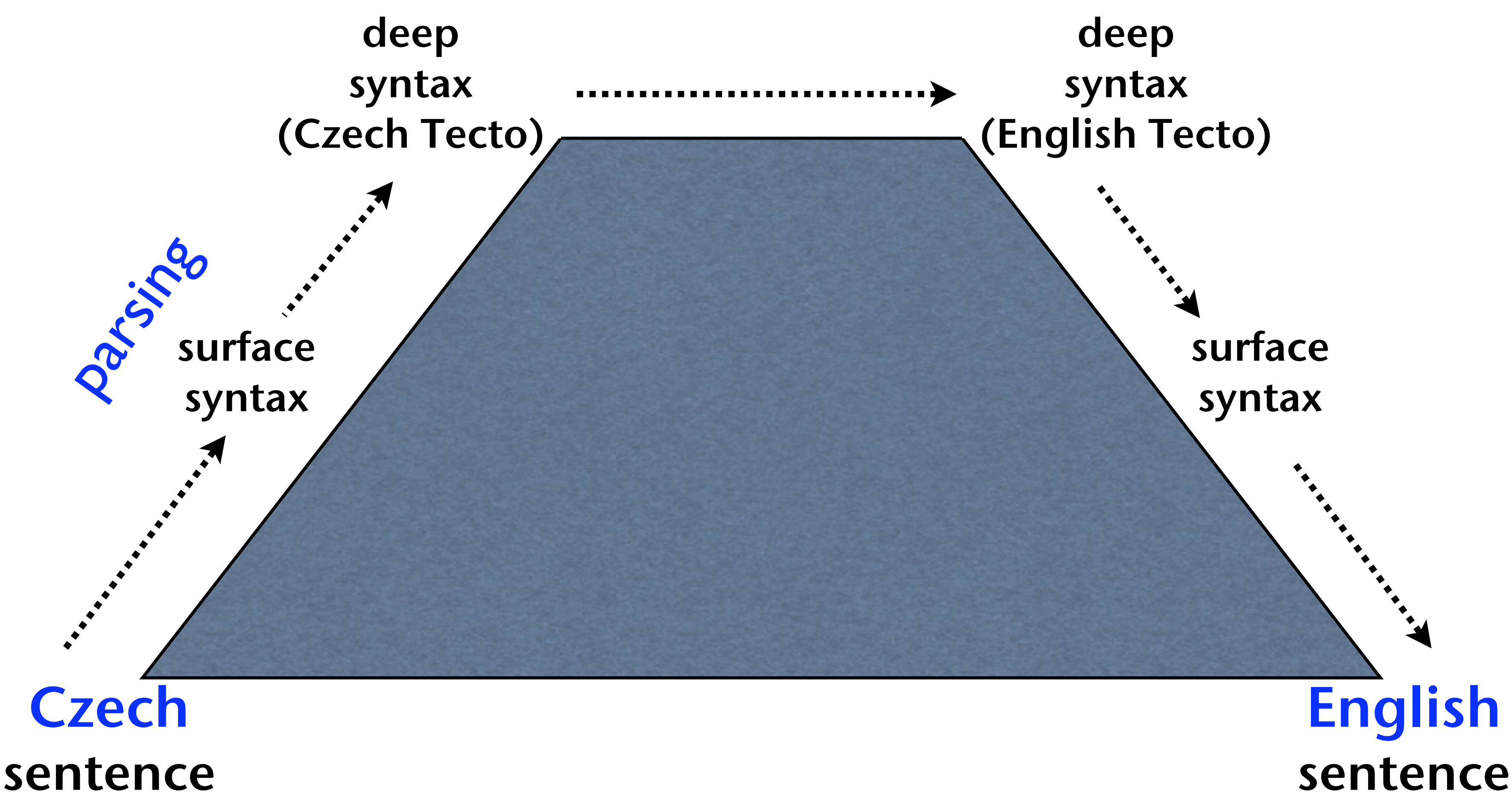
Tecto Transfer-based MT



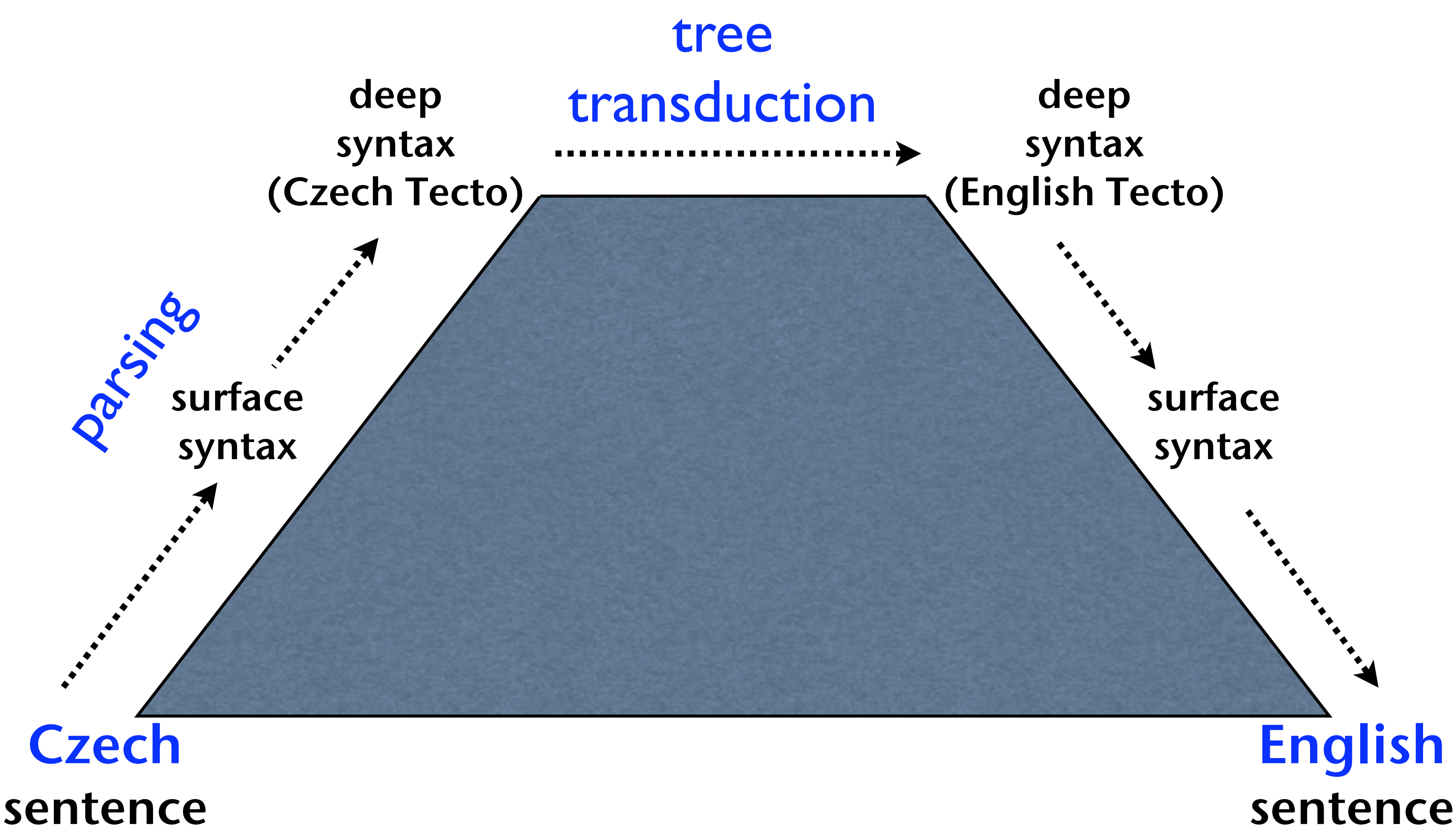
Tecto Transfer-based MT



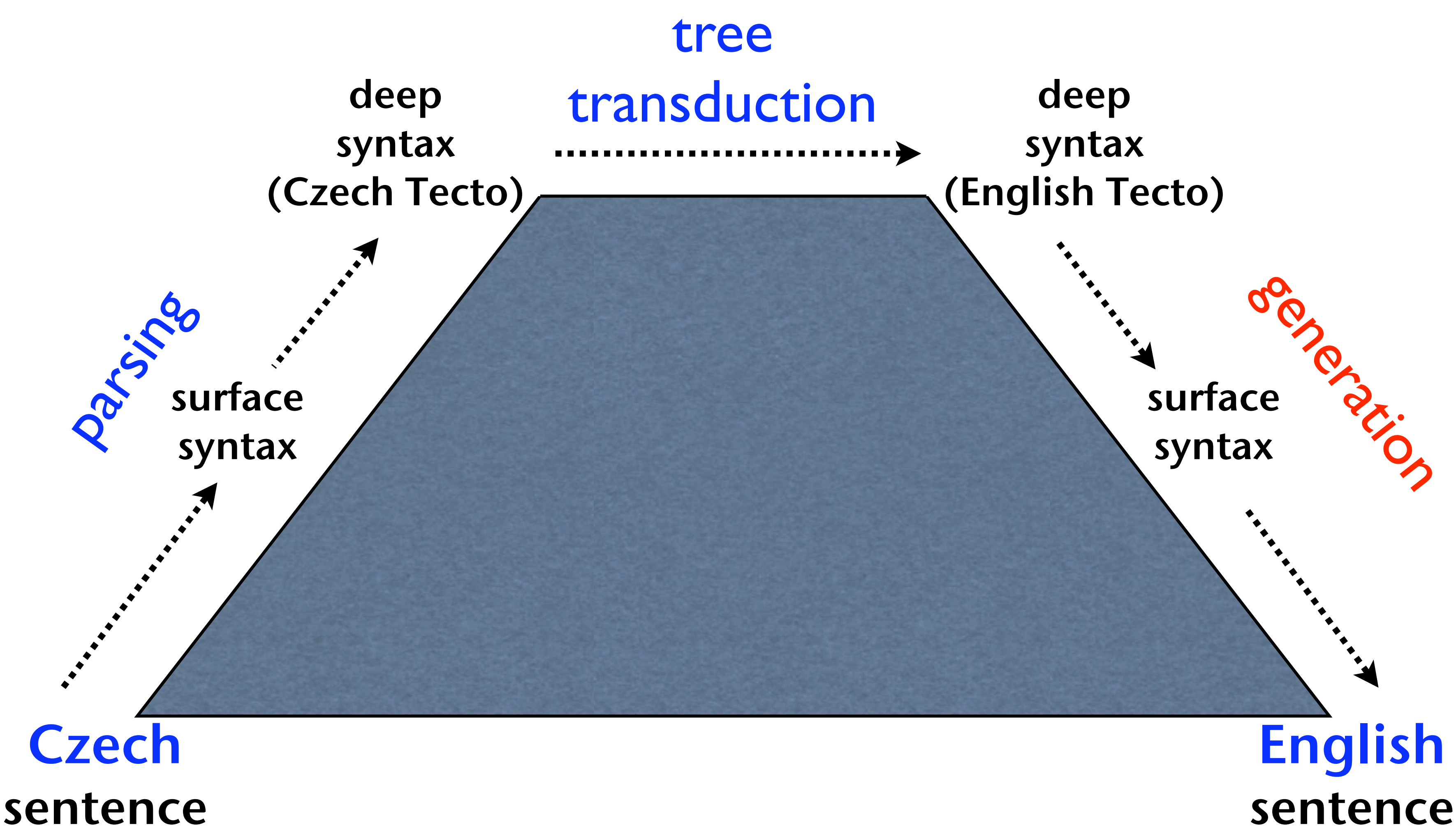
Tecto Transfer-based MT



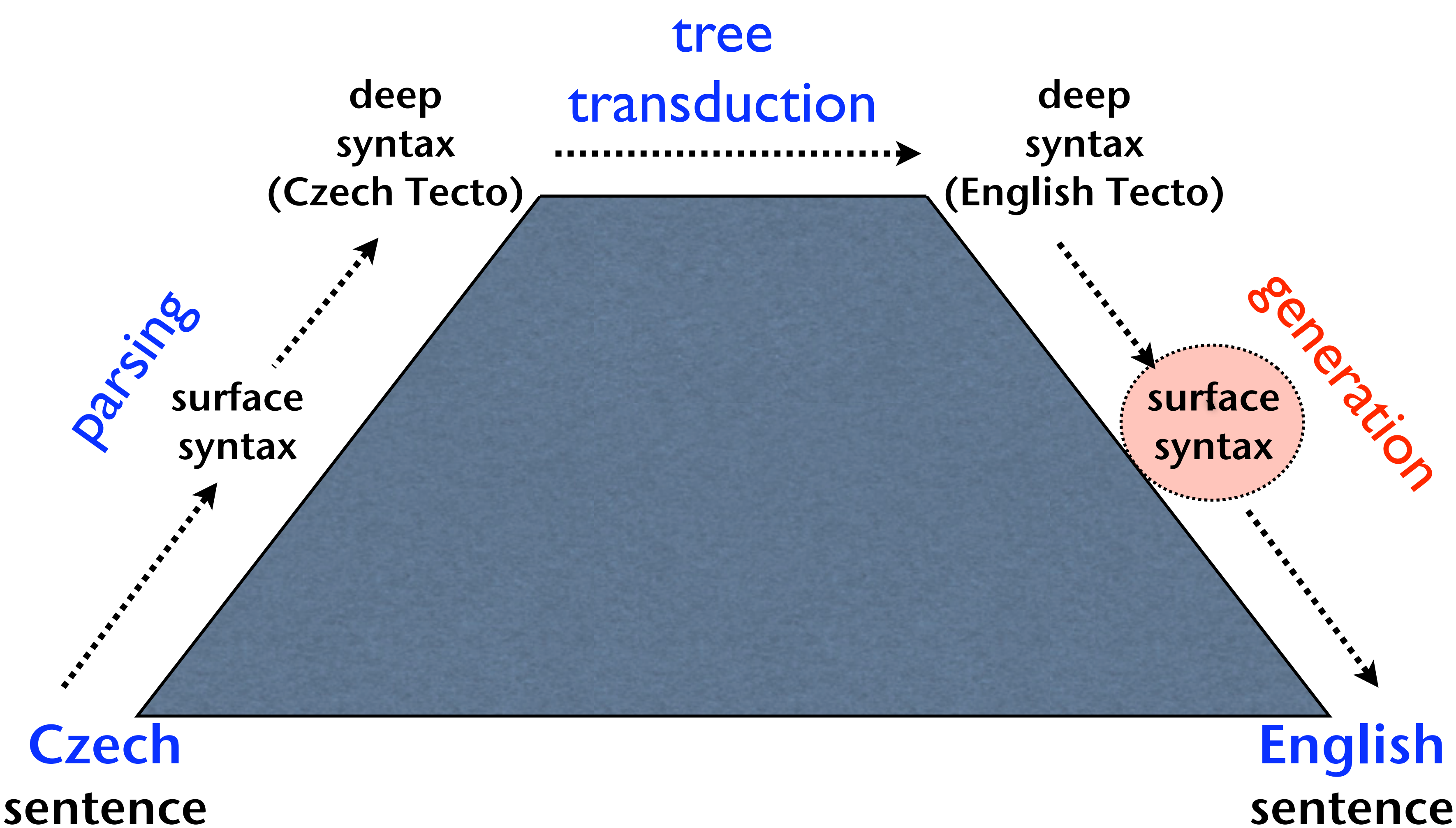
Tecto Transfer-based MT



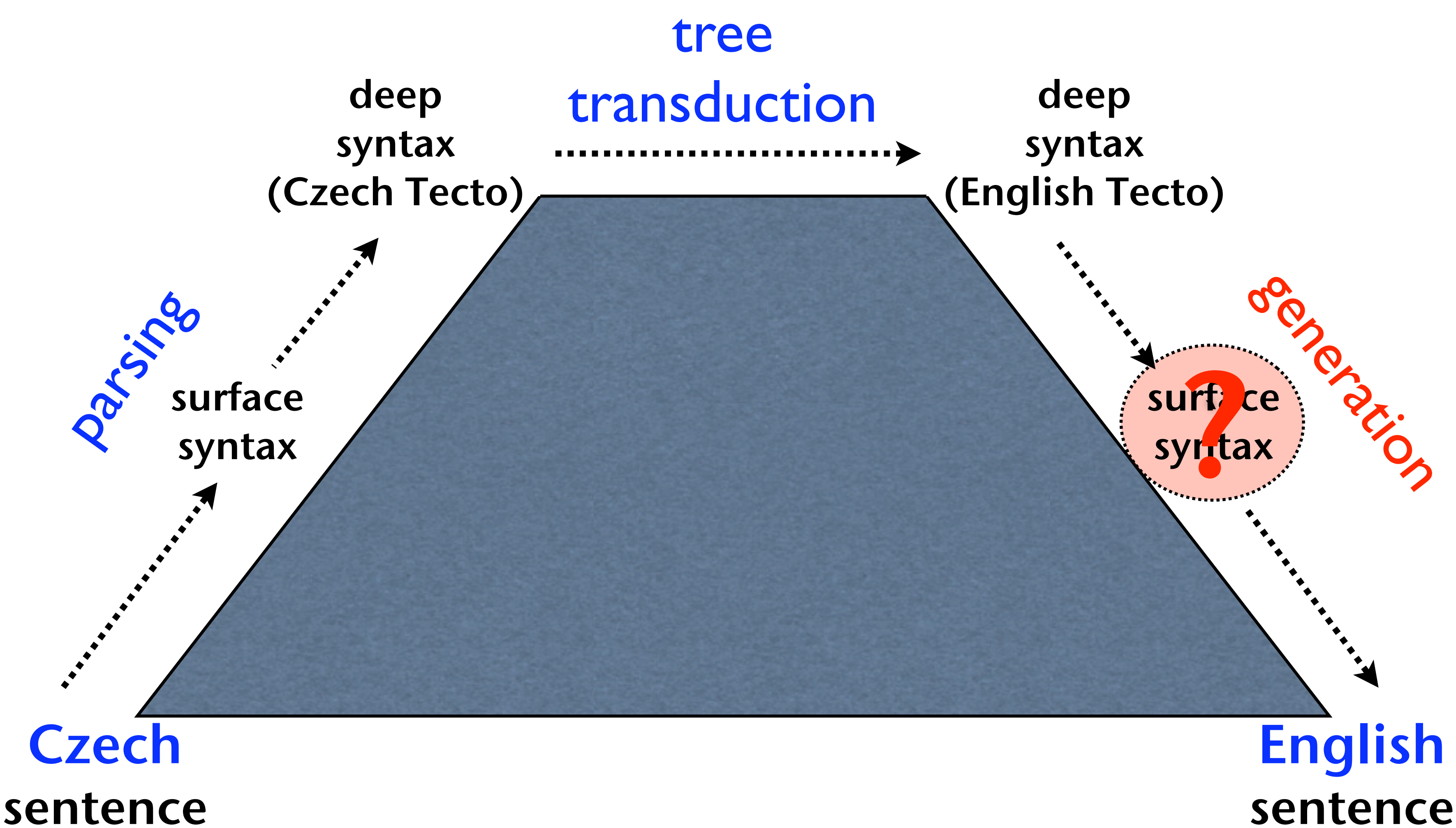
Tecto Transfer-based MT



Tecto Transfer-based MT



Tecto Transfer-based MT

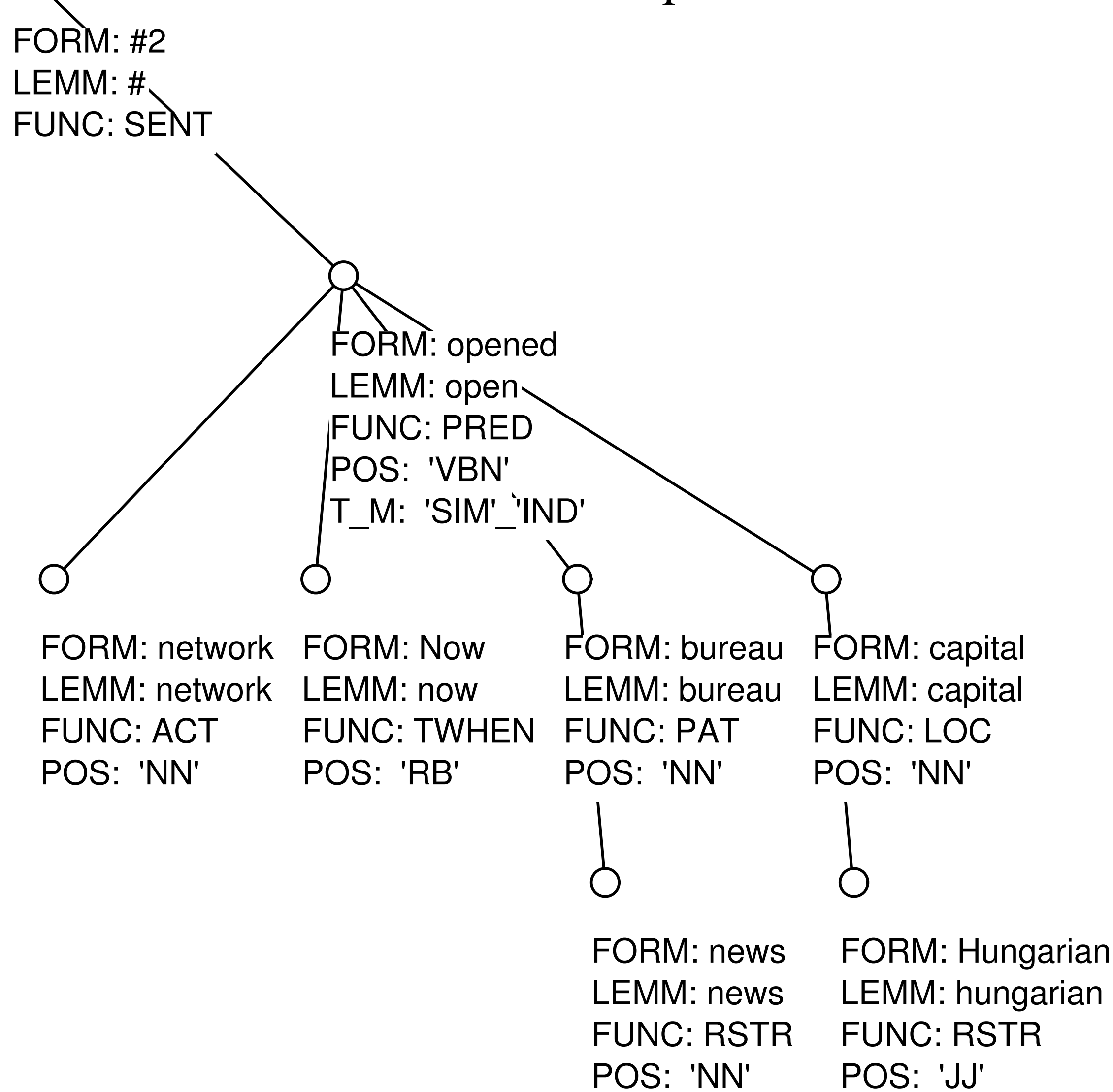


Transfer-based MT

- Allows us to explore deep syntactic representations
- Factored models are clear
- Need not be greedy one-best process
  - although we present one-best generation/results

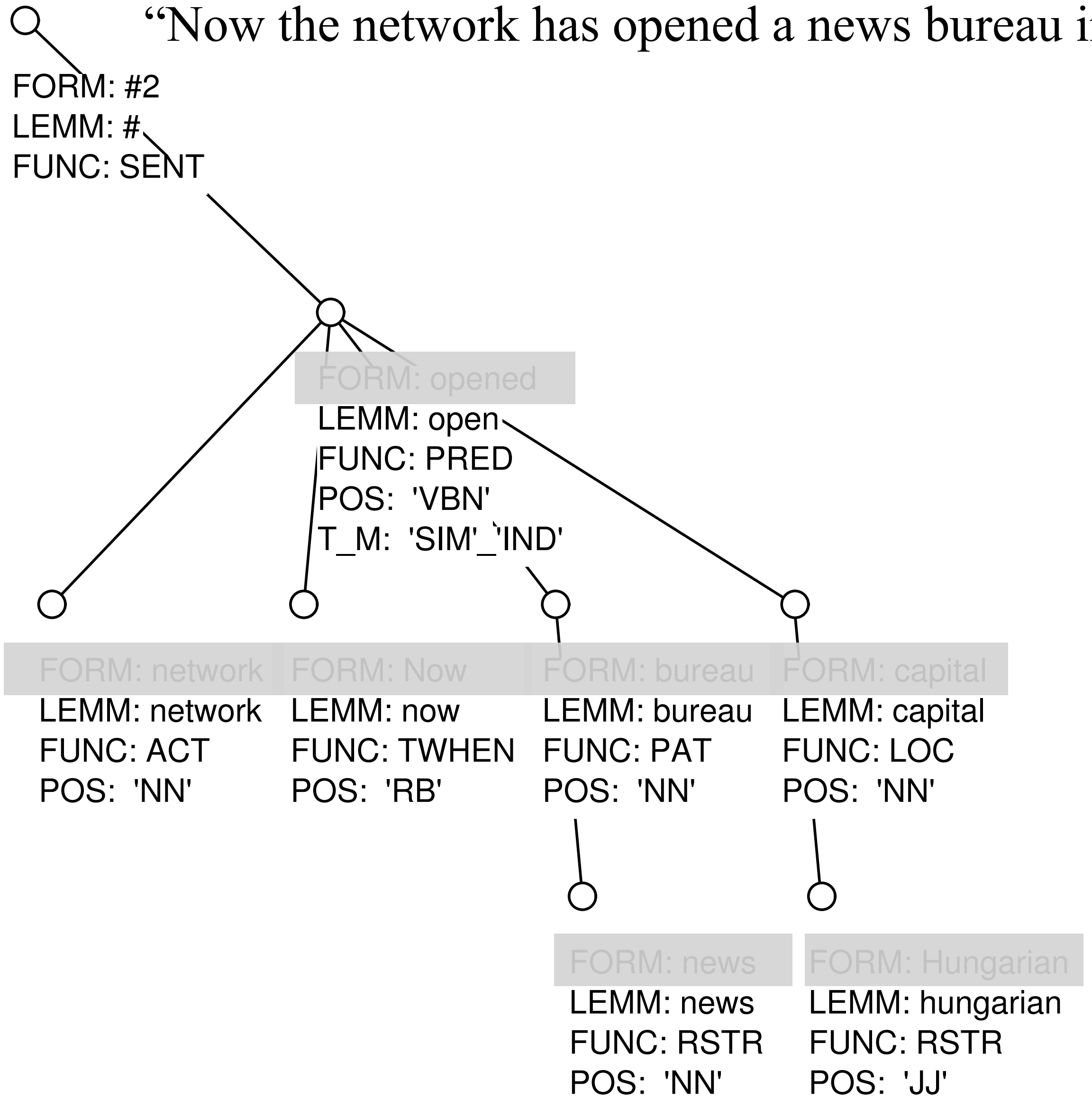
Tectogrammatical Representation

"Now the network has opened a news bureau in the Hungarian capital"



Tectogrammatical Representation

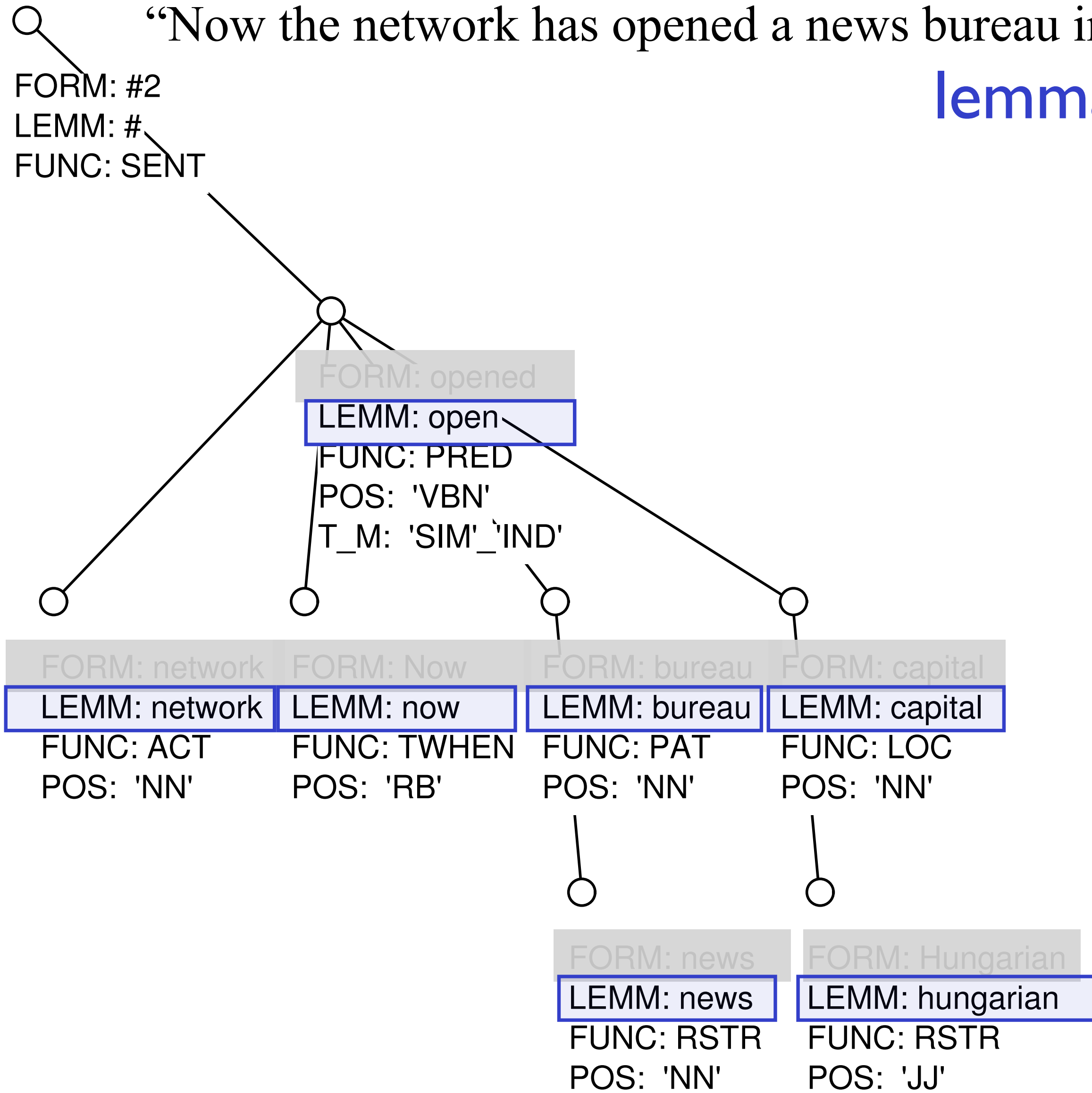
"Now the network has opened a news bureau in the Hungarian capital"



Tectogrammatical Representation

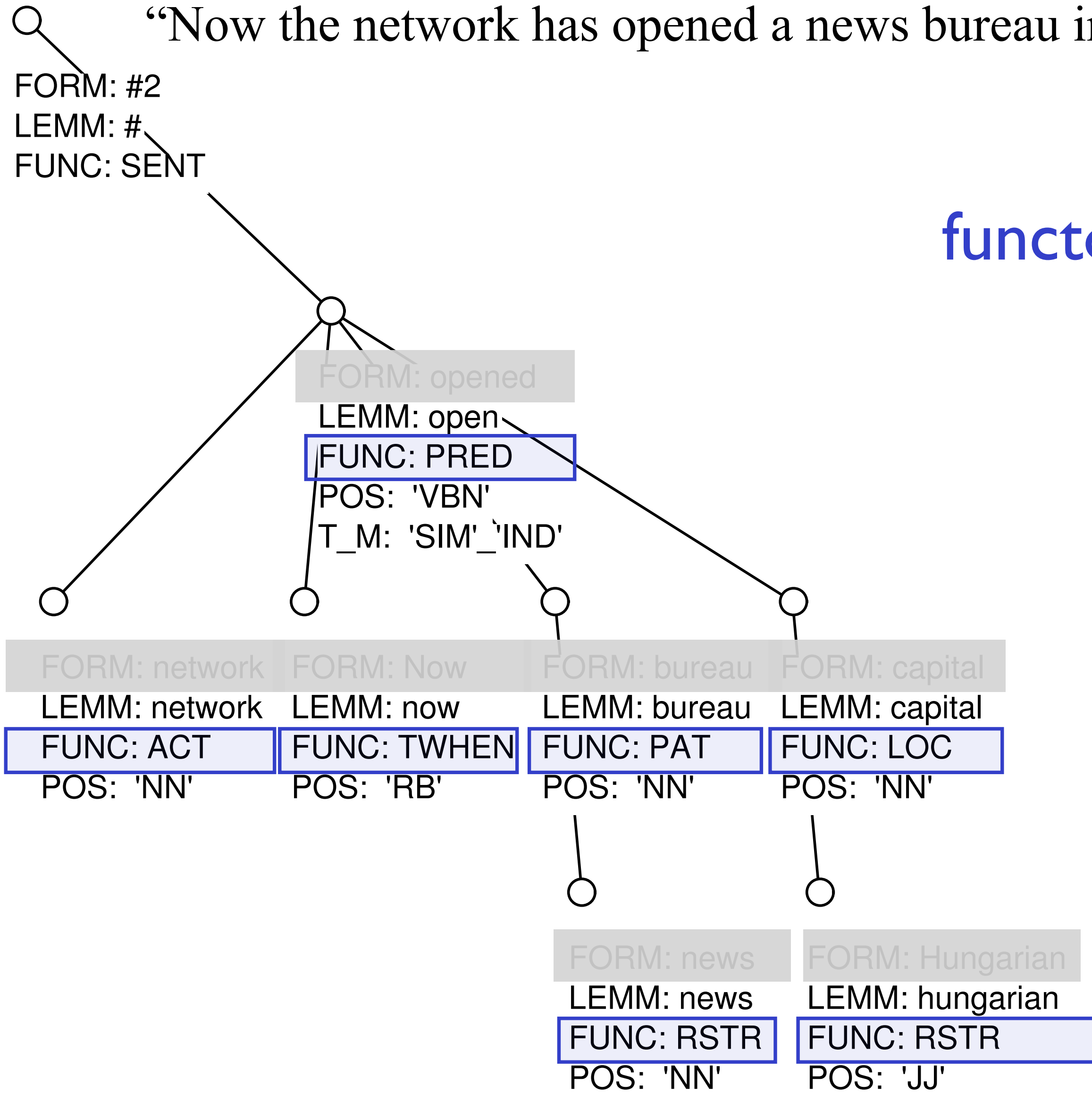
"Now the network has opened a news bureau in the Hungarian capital"

lemma



Tectogrammatical Representation

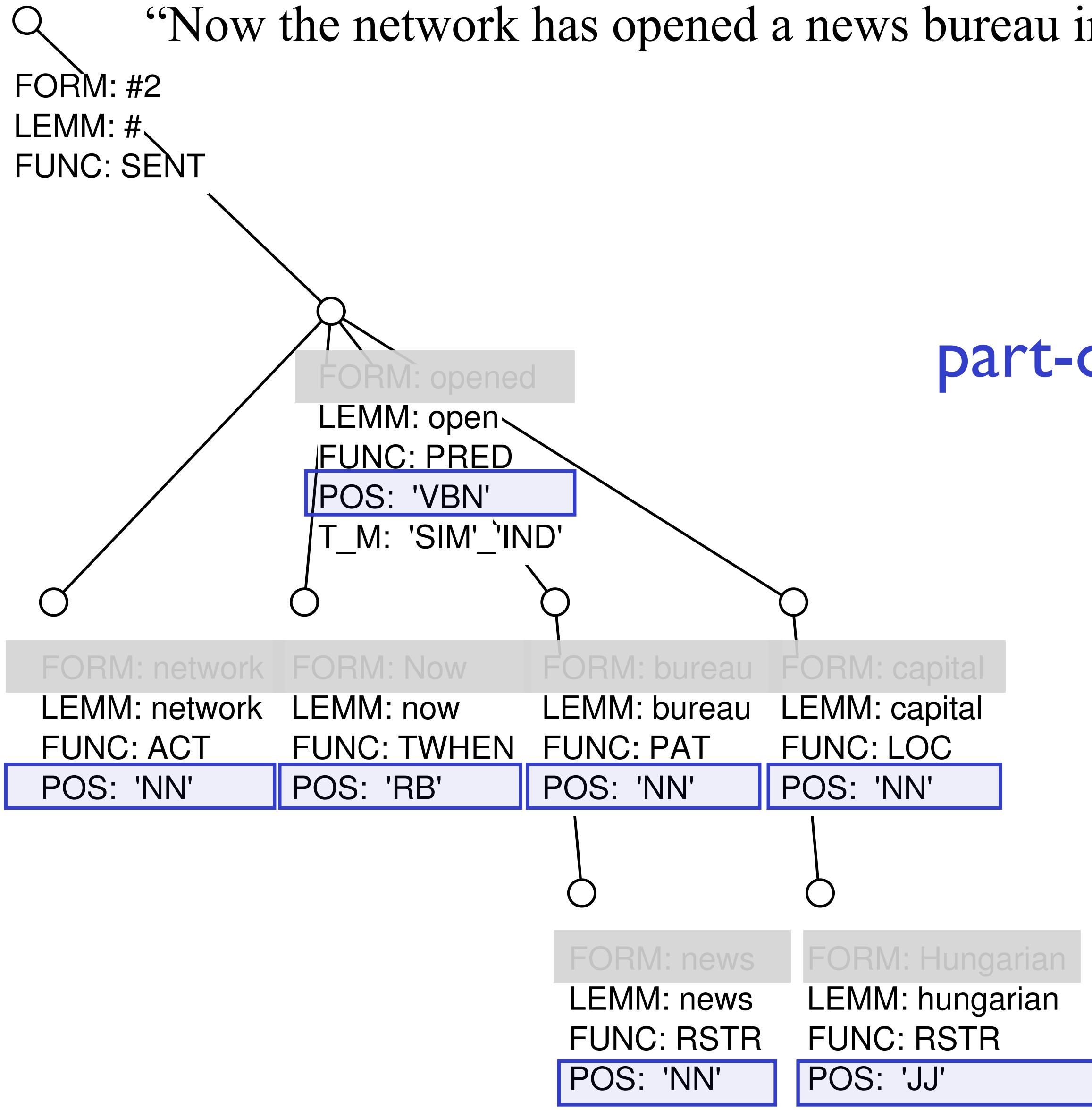
"Now the network has opened a news bureau in the Hungarian capital"



functor

Tectogrammatical Representation

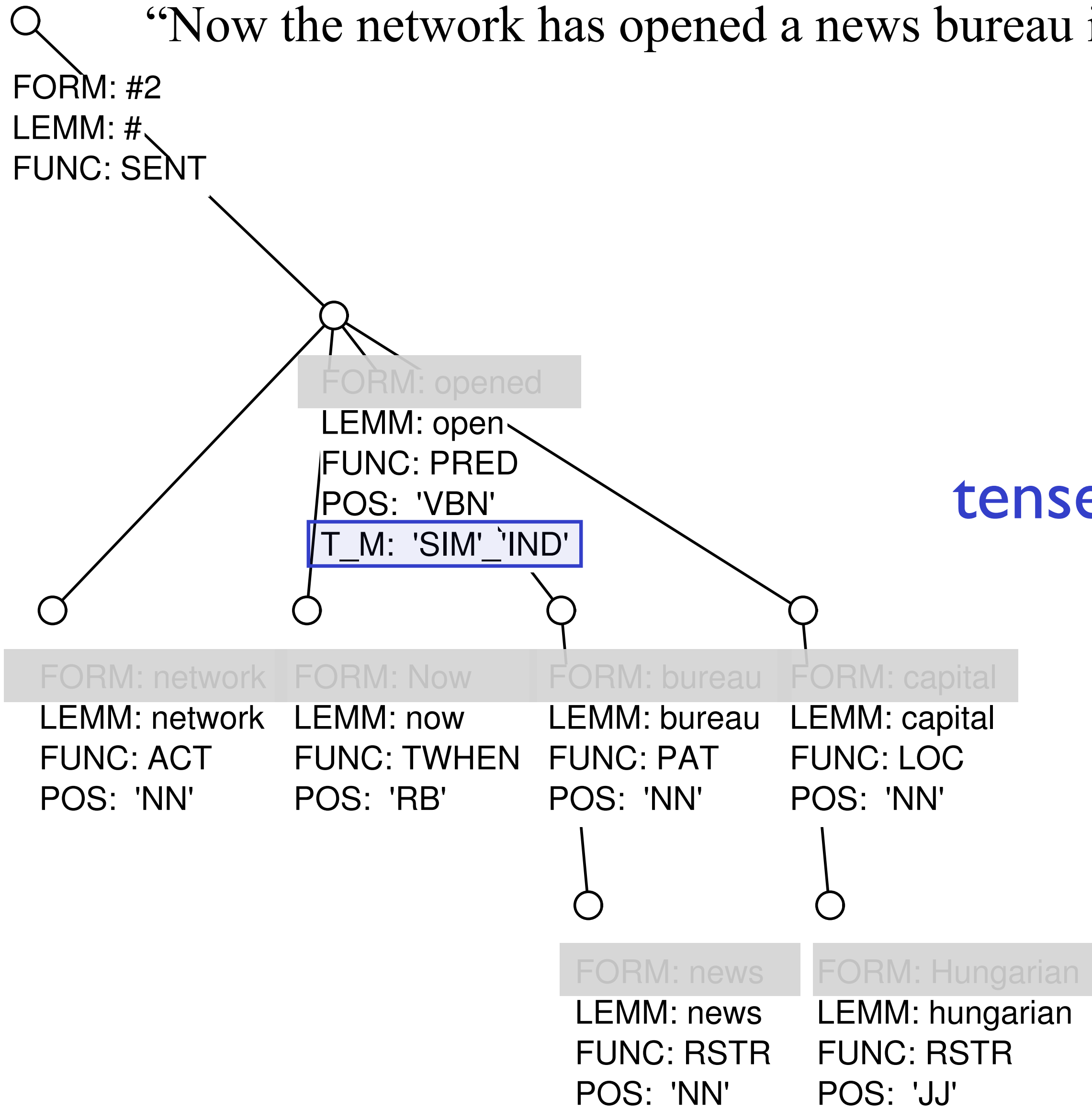
“Now the network has opened a news bureau in the Hungarian capital”



part-of-speech

Tectogrammatical Representation

“Now the network has opened a news bureau in the Hungarian capital”

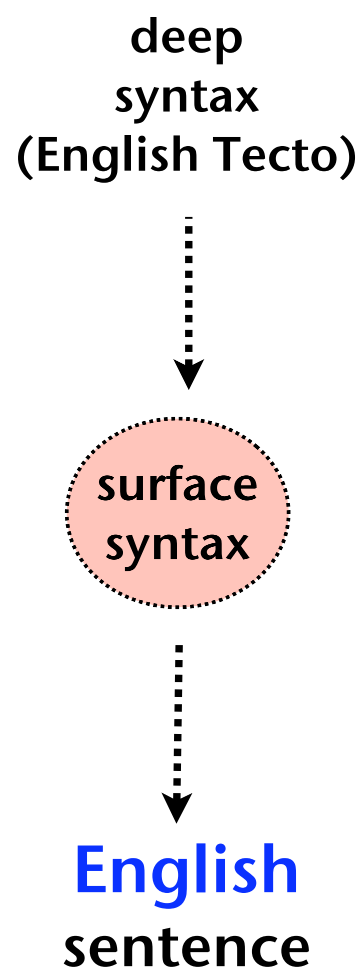


tense & mood

Generation Process

1. Insert syn-semantic (function) words
2. Subtree reordering

- Intermediary surface syntax ?
- Reordering constraints?
    - maximum subtree size
    - coordination



Generation Model

$$\begin{aligned}
&\arg\max_{A,f} P(A, f|T) \\
&= \arg\max_{A,f} P(f|A,T)P(A|T) \\
&\approx \arg\max_{f} P(f|T, \arg\max_{A} P(A|T))
\end{aligned}$$

- tecto nodes: $T = \{t_1, \ldots, t_i, \ldots, t_n\}$
- insertion string: $A = \{a_1, \ldots, a_i, \ldots, a_k\}$

  $n \leq k \leq 2n$
- order mapping: $f : \{A \cup T\} \rightarrow \{1, \ldots, 2n\}$

Generation Model

$$\arg\max_{A,f} P(A, f|T)$$

Insertion

$$= \arg\max_{A,f} P(f|A,T) P(A|T)$$

$$\approx \arg\max_{f} P(f|T, \arg\max_{A} P(A|T))$$

- tecto nodes: $T = \{t_1, \ldots, t_i, \ldots, t_n\}$
- insertion string: $A = \{a_1, \ldots, a_i, \ldots, a_k\}$
  $n \leq k \leq 2n$
- order mapping: $f : \{A \cup T\} \rightarrow \{1, \ldots, 2n\}$

Generation Model

$$\arg\max_{A,f} P(A, f|T)$$ Reordering

$$= \arg\max_{A,f} P(f|A,T) P(A|T)$$

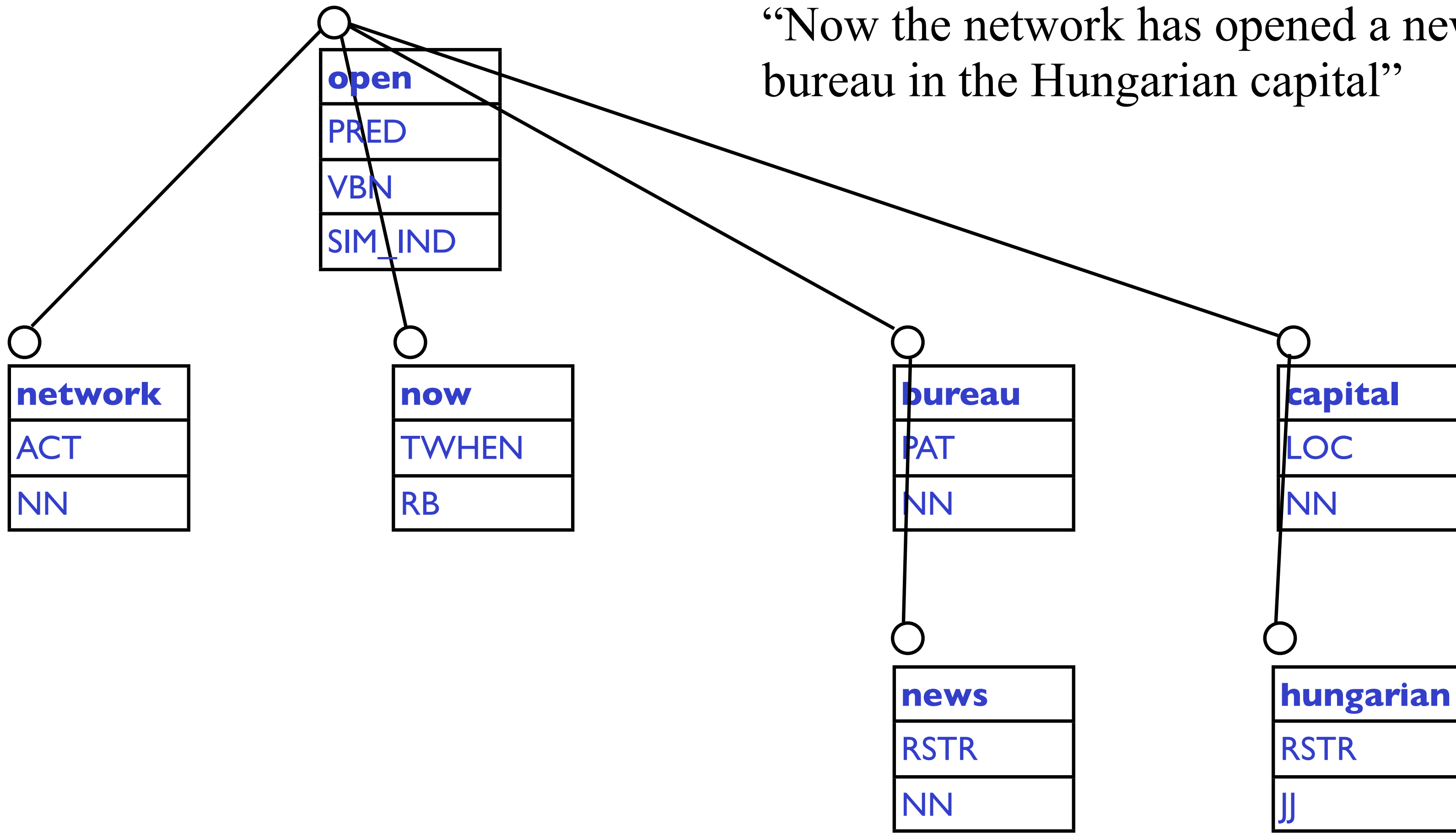
$$\approx \arg\max_{f} P(f|T, \arg\max_{A} P(A|T))$$

- tecto nodes: $T = \{t_1, \ldots, t_i, \ldots, t_n\}$
- insertion string: $A = \{a_1, \ldots, a_i, \ldots, a_k\}$

  $n \leq k \leq 2n$
- order mapping: $f : \{A \cup T\} \rightarrow \{1, \ldots, 2n\}$

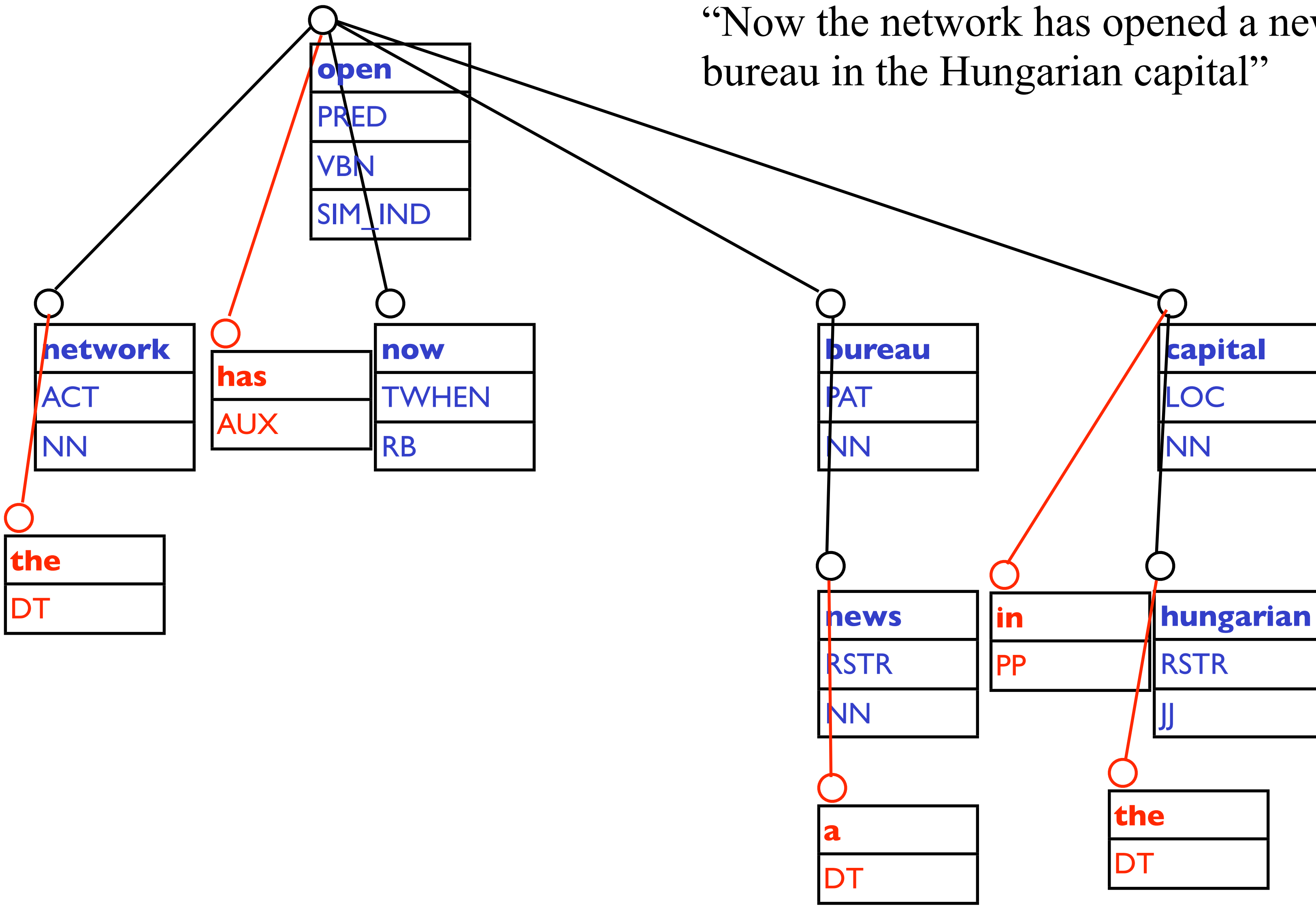
Insertion Process

"Now the network has opened a news bureau in the Hungarian capital"



Insertion Process

"Now the network has opened a news bureau in the Hungarian capital"



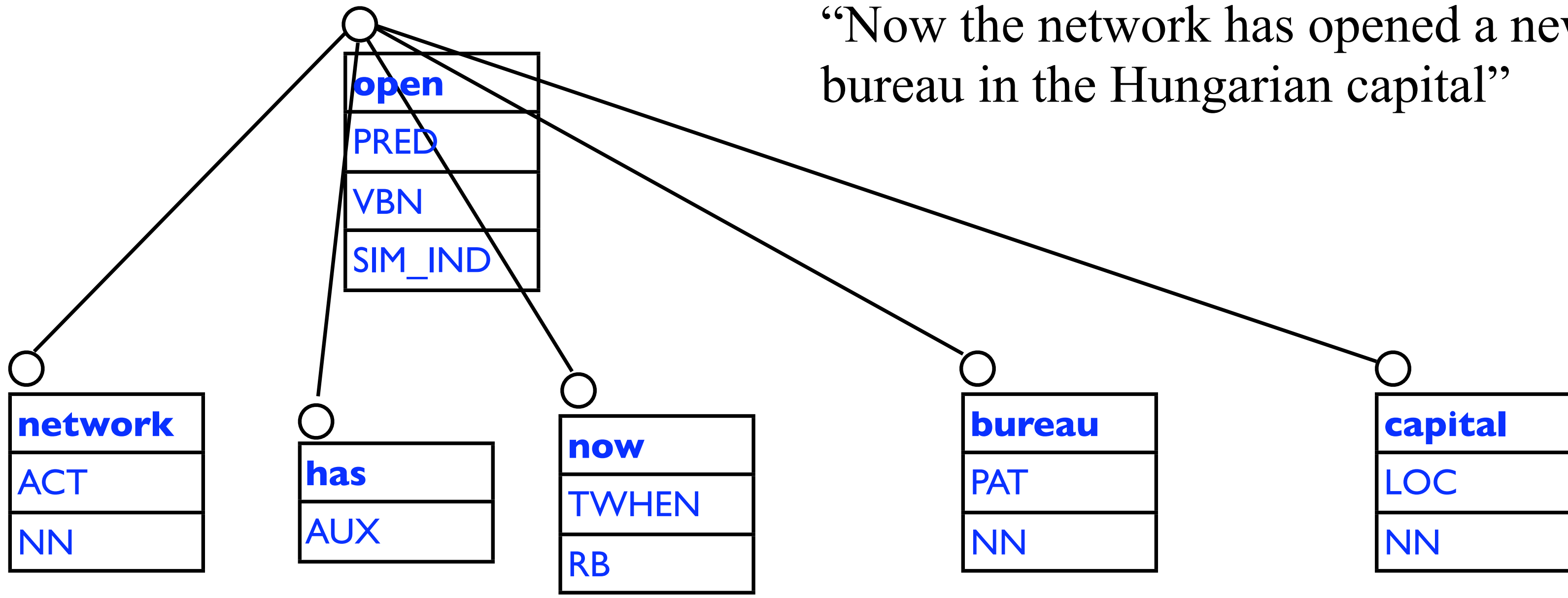
Insertion Model

$$P(A|T)$$
$$= \prod_i P(a_i|a_1, \ldots, a_{i-1}, T)$$
$$\approx \prod_i P(a_i|t_i, t_{g(i)})$$

- Insertion is dependent on local context:
  - tecto node (includes: lemma, functor, POS)
  - parent node
- Three independent models:
  - articles
  - prepositions and subordinating conjunctions
  - modals (deterministic, given functor)

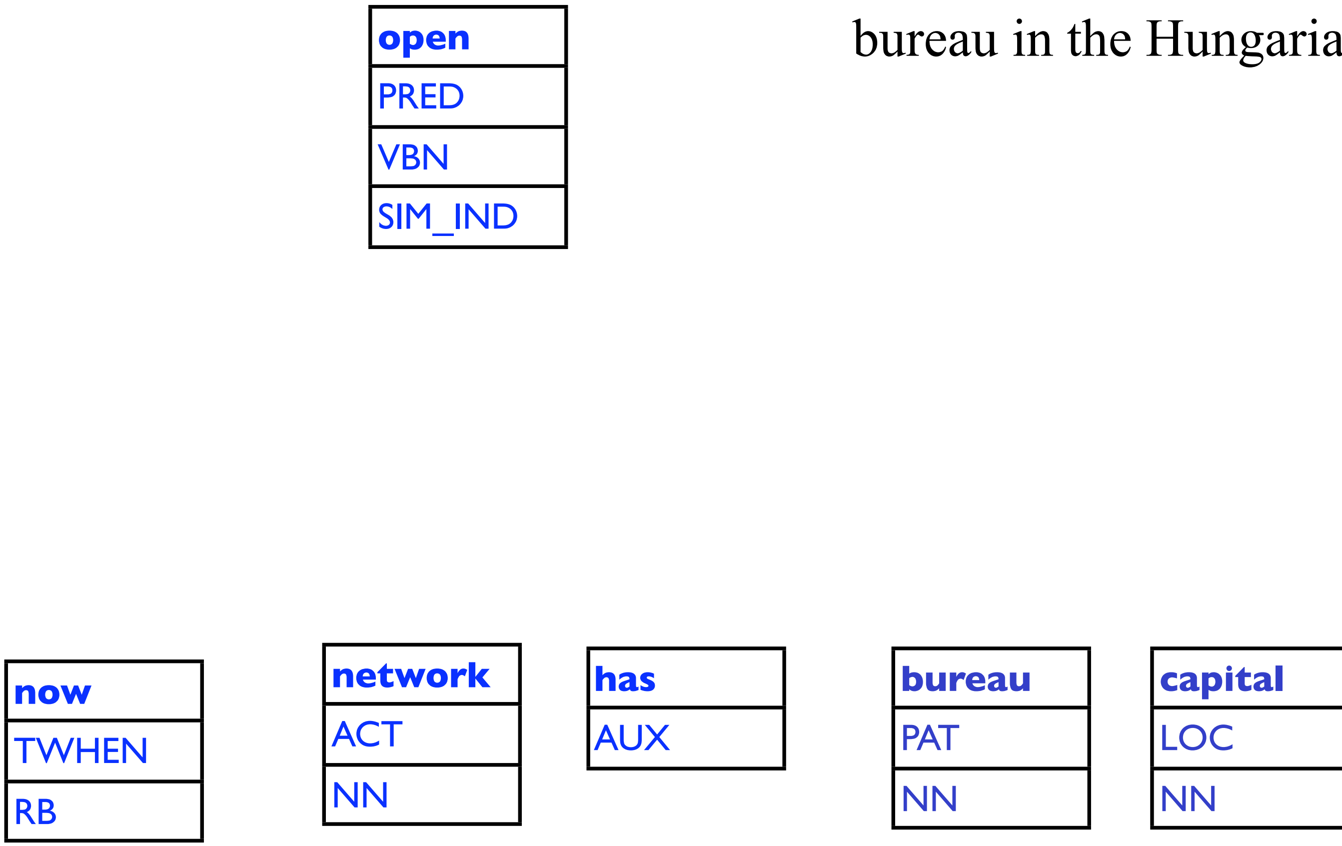
Reordering Process

“Now the network has opened a news bureau in the Hungarian capital”



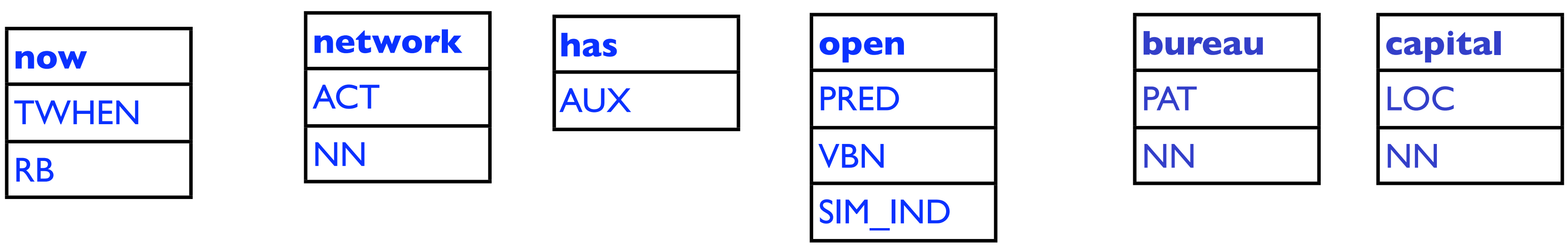
Reordering Process

"Now the network has opened a news bureau in the Hungarian capital"



Reordering Process

"Now the network has opened a news bureau in the Hungarian capital"



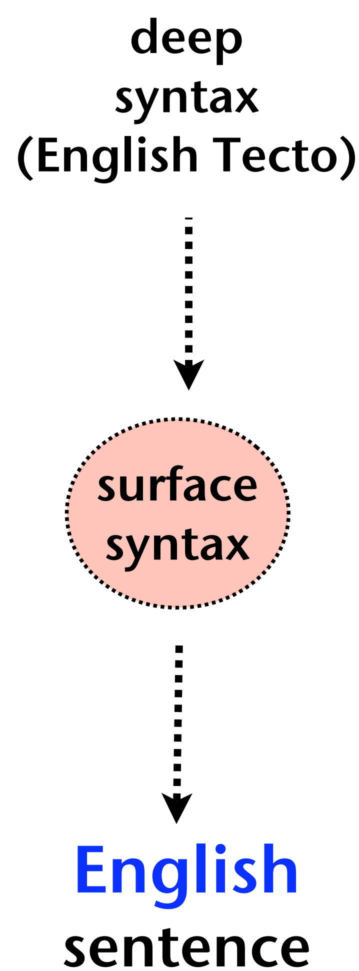
Surface Order Model

1. child order: $P(c_i \prec c_{i+1} | c_i, c_{i+1}, g)$
   $= (c_i \prec c_{i+1} | f_i, t_i, f_{i+1}, t_{i+1}, f_g, t_g)$
2. gov. position: $P(c_i \prec g \prec c_{i+1} | c_i, c_{i+1}, g)$
   $= P(c_i \prec g \prec c_{i+1} | f_i, t_i, f_{i+1}, t_{i+1}, t_g, f_g)$

- Greedy procedure
  (there is an alternative DP solution)
- Factored models can be estimated separately
- Constraint on reorderings: maximum 5 children
- Features: functors & POS tags

Intermediate Syntax

- Insertion from Tectogrammatical Trees
- Convert deep functors to syntactic functions
  - P(VERB | PRED)
  - P(SBJ | ACT)
- Reordering based on syntactic features
  - should be a closer match to surface-syntax transfer



Evaluation

- Training
  - ~50k WSJ treebank automatically converted
- Training & Eval: PCEDT Corpus 1.0:
  - Penn WSJ treebank translated to Czech
    4 retranslations back to English
  - ~ 20k sentences of automatic TR
  - ~ 500 sentences of manual TR
- History based modes
  - smoothed via linear-backoff EM-smoothing

Evaluation: Insertion

Model	Manual Data				Synthetic Data			
	Ins. Rules		No Rules		Ins. Rules		No Rules	
Model	Articles	Prep & SC	Articles	Prep & SC	Articles	Prep & SC	Articles	Prep & SC
Baseline	N/A	N/A	77.93	76.78	N/A	N/A	78.00	78.40
w/o g. functor	87.29	89.65	86.25	**89.31**	**88.07**	91.83	**87.34**	91.06
w/o g. lemma	86.77	89.48	85.68	89.02	87.53	90.95	86.55	91.16
w/o g. POS	87.29	89.45	86.10	89.14	87.68	**91.86**	86.89	**92.07**
w/o functor	86.10	85.02	84.86	84.56	86.01	85.60	84.79	85.65
w/o lemma	81.34	89.02	80.88	88.91	81.28	91.03	81.42	91.33
w/o POS	84.81	88.01	84.01	87.29	85.53	91.08	84.69	90.98
All Features	**87.49**	**89.68**	**86.45**	89.28	87.87	91.83	87.24	92.02

- Manual data - hand annotated
- Synthetic data - automatically produced (matches training data)
- "Rules" - Small set of deterministic rules
  - applied if no majority prediction (all < .5)

Article Insertion

% Errors	Reference→Hypothesis
41	the → NULL
19	a/an → NULL
16	NULL → the
11	a/an → the
11	the → a/an
2	NULL → a/an

- Conservative model
  - 60% of the error is do to NULL insertion
- Assume equivalence of 'a' and 'an'

Evaluation: Reordering

Model	Manual Data				Synthetic Data			
	Coord. Rules		No Rules		Coord. Rules		No Rules	
	All	Interior	All	Interior	All	Interior	All	Interior
Baseline	N/A	N/A	68.43	21.67	N/A	N/A	69.00	21.42
w/o g. functor	**94.51**	**86.44**	**92.42**	**81.27**	94.90	87.25	93.37	83.42
w/o g. tag	93.43	83.75	90.89	77.50	93.82	84.56	91.64	79.12
w/o c. functors	91.38	78.70	89.71	74.57	91.91	79.79	90.41	76.04
w/o c. tags	88.85	72.44	82.29	57.36	88.91	72.29	83.04	57.60
All Features	94.43	86.24	92.01	80.26	**95.21**	**88.04**	**93.37**	**83.42**

- Evaluation based on Hajič et al. 2002
  - Percentage of correct subtrees (no credit for partial order)
- Reordering correct trees (no insertion errors)

Evaluation: Full

Model	Manual	Synthetic
TR w/ Rules	**.4614**	**.4777**
TR w/o Rules	.4532	.4657
AR	.2337	.2451

- Morphological insertion by Morphg (Carroll)
- BLEU score against original + 4 retranslations
  - "bound" on performance of MT system using this generation component
- AR - intermediate syntax
  - lost information in mapping (valency ordering!)

Related work

- Amalgam (Corston-Oliver et al. '02)
  - Generation from a logical form
  - Assumes more information than impoverished TR
- Halogen (Langkilde-Geary '02)
  - *minimally specified* results closest to ours

Conclusions

- Simple generative models capable of recovering knowledge from deep structure
  - limited history, simple smoothing
- Greedy decoding procedure is fast, but joint decoder would likely help
  - insertion/reordering not conditionally independent