Quasi-Synchronous Phrase Dependency Grammars for Machine Translation

Kevin Gimpel   Noah A. Smith

Introduction

- MT using dependency grammars on **phrases**
  - Phrases capture local reordering and idiomatic translations
  - Dependencies model long-distance phenomena
  - Tighter integration of phrase-based and dependency-based MT

Introduction

- MT using dependency grammars on **phrases**
  - Phrases capture local reordering and idiomatic translations
  - Dependencies model long-distance phenomena
  - Tighter integration of phrase-based and dependency-based MT

- Uses **quasi-synchronous grammar** (Smith & Eisner 2006)
  - Flexible framework for modeling sentence relationships
  - Naturally models tree-to-tree translation using features instead of hard constraints

ANC opposition sanction Zimbabwe

非国大 反对 制裁 津巴布韦

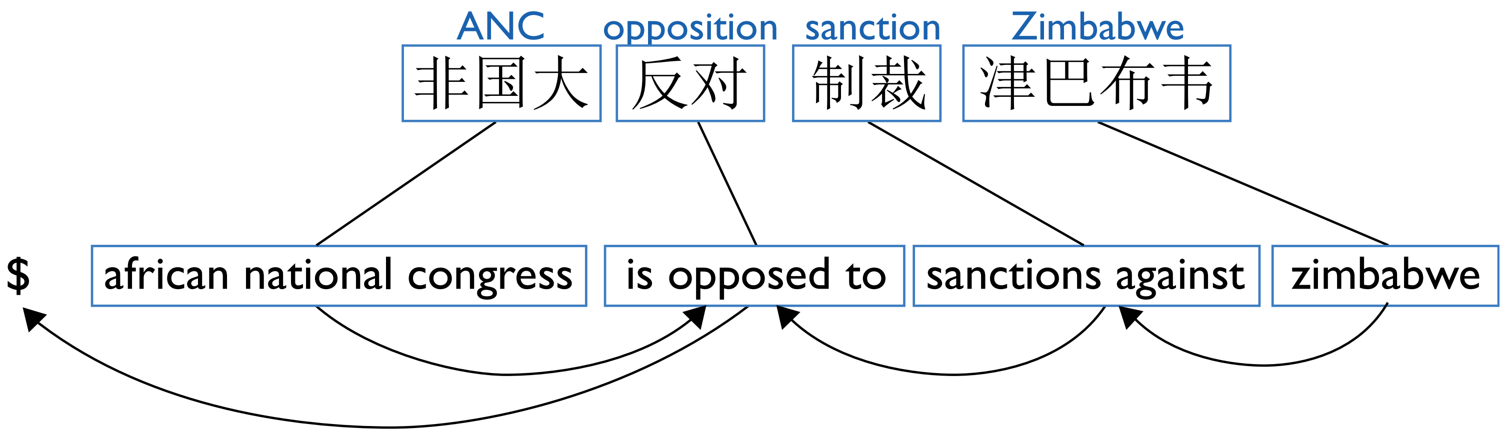
Reference: african national congress opposes sanctions against zimbabwe

ANC	opposition	sanction	Zimbabwe
非国大	反对	制裁	津巴布韦

Moses: opposition to | sanctions against | zimbabwe | african national congress

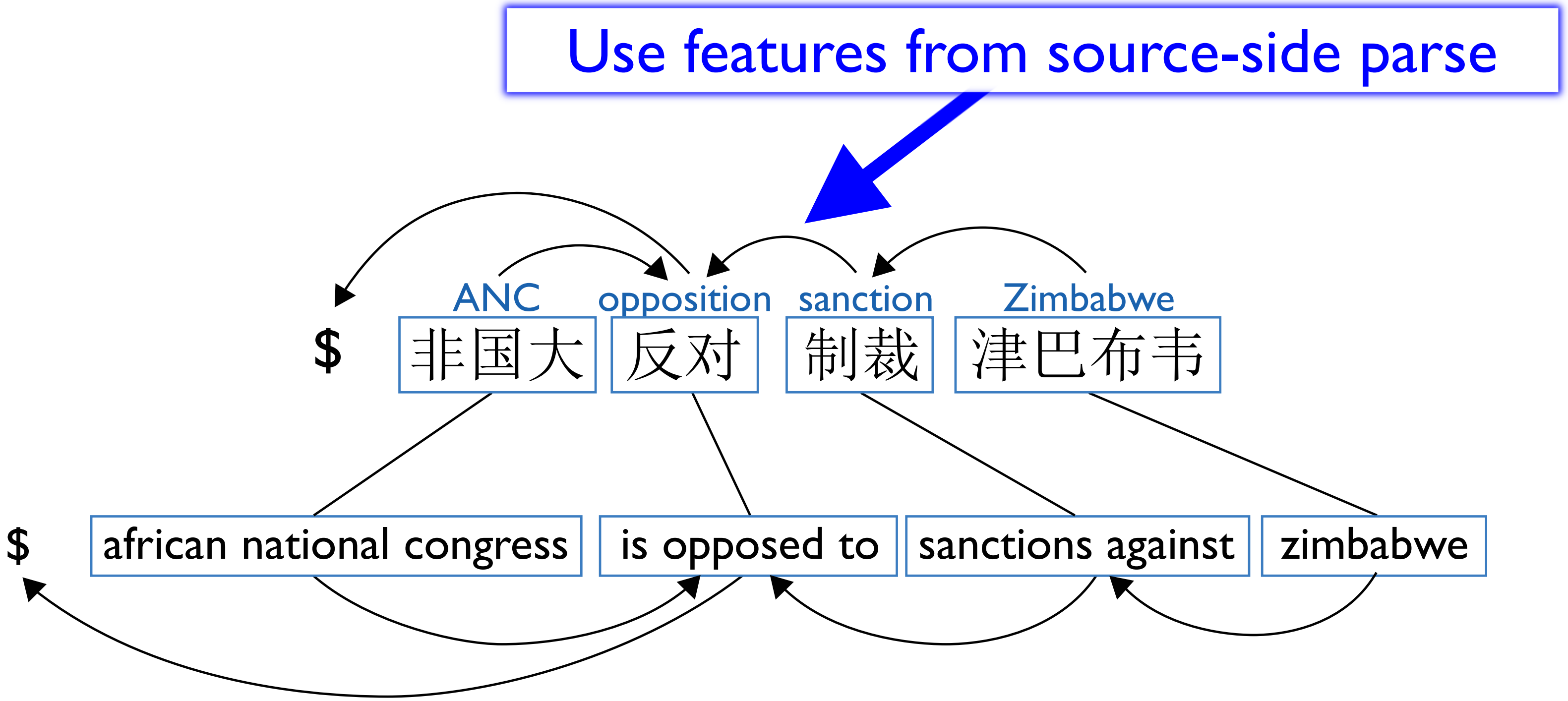
Reference: african national congress opposes sanctions against zimbabwe

Our System:



Reference: african national congress opposes sanctions against zimbabwe

Our System:



Reference: african national congress opposes sanctions against zimbabwe

Related Work

- Ding & Palmer (2005)
- Quirk et al. (2005)
- Shen et al. (2008)
- Galley & Manning (2009), Carreras & Collins (2009)
- Hunter & Resnik (2010): dependency grammars on phrases for MT
- This talk:
  - Formulation using quasi-synchronous grammar
  - New algorithms for feature extraction and decoding
  - Larger-scale experiments and empirical improvements

Two Questions

- How do we score dependency trees on phrases?

- How do we decode?

Two Questions

- How do we score dependency trees on phrases?

- How do we decode?

Two Questions

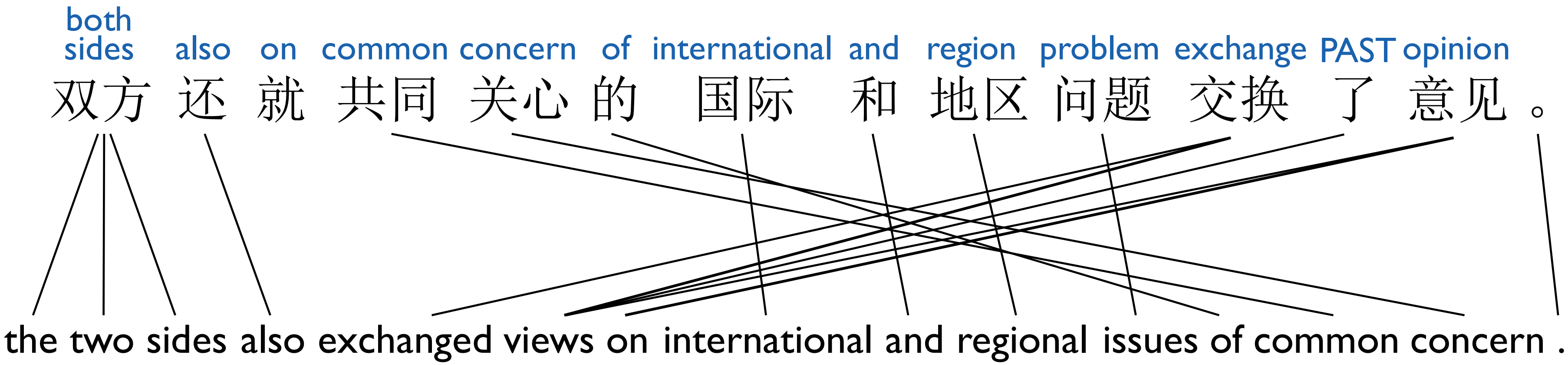
- How do we score dependency trees on phrases?
  - Parse the target side of the parallel corpus
  - Use a heuristic to extract parent-child phrase dependencies
  - Compute relative frequency estimates of P(child phrase | parent phrase, direction)

- How do we decode?

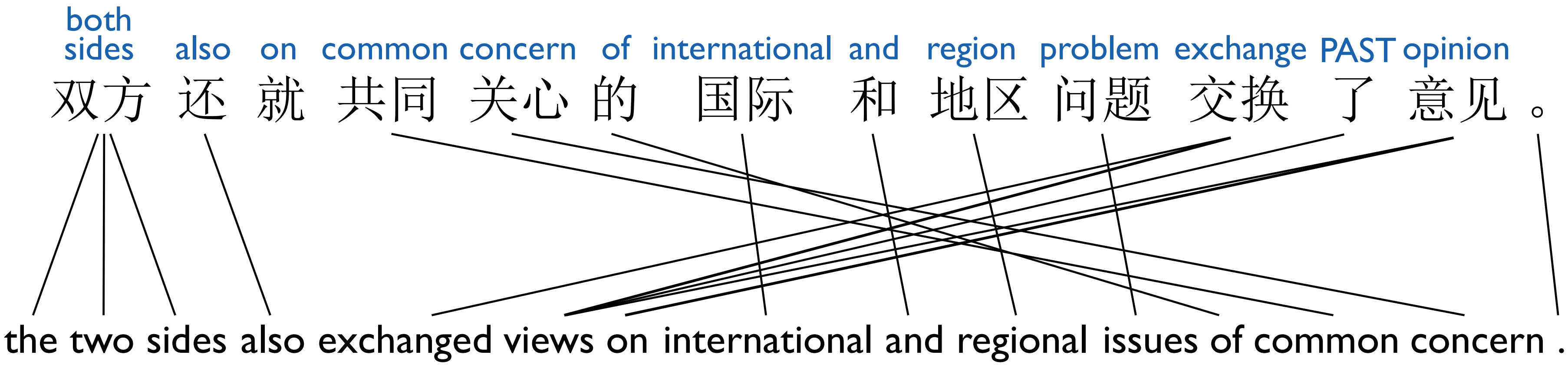
both sides also on common concern of international and region problem exchange PAST opinion

双方 还 就 共同 关心 的 国际 和 地区 问题 交换 了 意见 。

the two sides also exchanged views on international and regional issues of common concern .



双方	the two sides	国际	international
还	also	国际 和	international and
还 就	also	国际 和 地区	international and regional
就 共同	common	和	and
就 共同 关心	common concern	和 地区	and regional
共同	common	和 地区 问题	and regional issues
共同 关心	common concern	地区	regional
共同 关心 的	of common concern	地区 问题	regional issues
关心	concern	问题	issues
的	of	交换 了 意见	exchanged views on
		。	.



the two sides
also
also
common
common concern
common
common concern
of common concern
concern
of

international
international and
international and regional
and
and regional
and regional issues
regional
regional issues
issues
exchanged views on
.

the two sides also exchanged views on international and regional issues of common concern .

the two sides	international
also	international and
also	international and regional
common	and
common concern	and regional
common	and regional issues
common concern	regional
of common concern	regional issues
concern	issues
of	exchanged views on
	.

the two sides
also
also
common
common concern
common
common concern
of common concern
concern
of

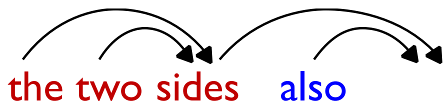
international
international and
international and regional
and
and regional
and regional issues
regional
regional issues
issues
exchanged views on
.

Carnegie Mellon

$ the two sides also exchanged views on international and regional issues of common concern .

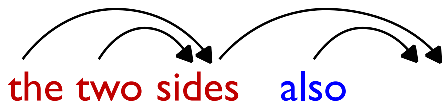
the two sides
also
exchanged views on
international
international and
international and regional
and
and regional
and regional issues
regional
regional issues
issues
of
of common concern
common
common concern
concern
.

the two sides
also
exchanged views on
international
international and
international and regional
and
and regional
and regional issues
regional
regional issues
issues
of
of common concern
common
common concern
concern
.



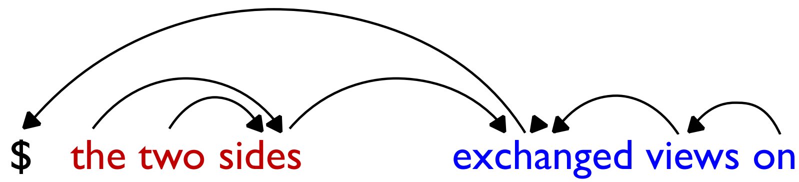
Carnegie Mellon

the two sides
also
exchanged views on
international
international and
international and regional
and
and regional
and regional issues
regional
regional issues
issues
of
of common concern
common
common concern
concern
.

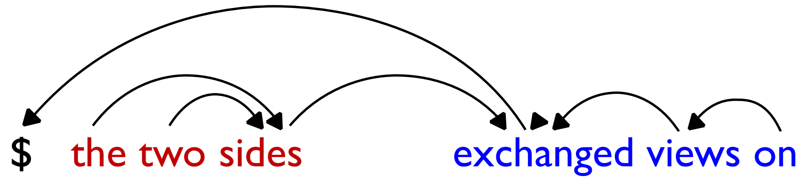


✘ No dependency edge between phrases

the two sides
also
exchanged views on
international
international and
international and regional
and
and regional
and regional issues
regional
regional issues
issues
of
of common concern
common
common concern
concern
.

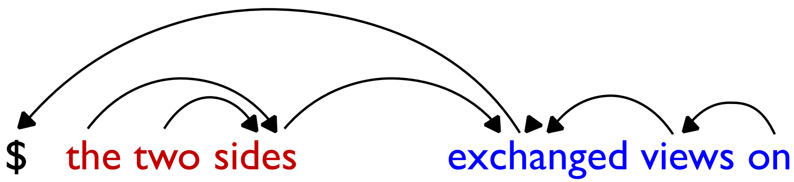


the two sides
also
exchanged views on
international
international and
international and regional
and
and regional
and regional issues
regional
regional issues
issues
of
of common concern
common
common concern
concern
.

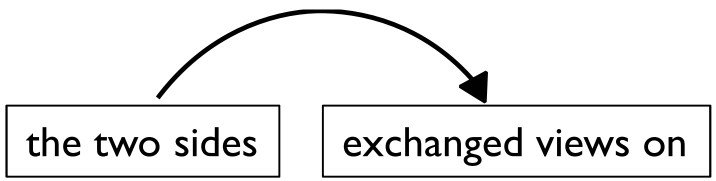


✓ Dependency edge between phrases

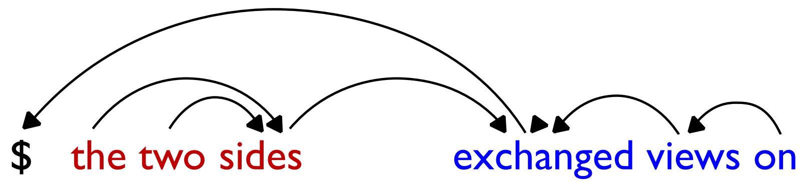
the two sides
also
exchanged views on
international
international and
international and regional
and
and regional
and regional issues
regional
regional issues
issues
of
of common concern
common
common concern
concern
.



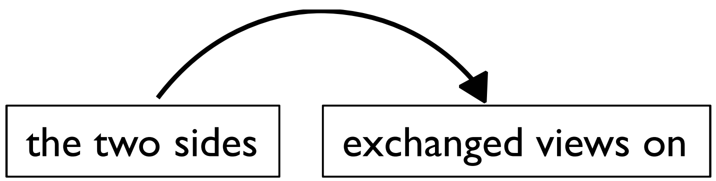
✓ Dependency edge between phrases

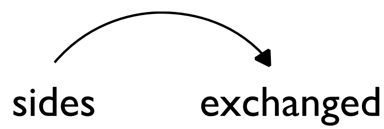


the two sides
also
exchanged views on
international
international and
international and regional
and
and regional
and regional issues
regional
regional issues
issues
of
of common concern
common
common concern
concern
.



Retain longest lexical dependency edge for lexical weighting features





Carnegie Mellon

New Features

- P(child | parent, direction)
- lex(child | parent, direction)
- Same two features using Brown clusters in place of words (Brown et al., 1992)
- Several additional local configuration features

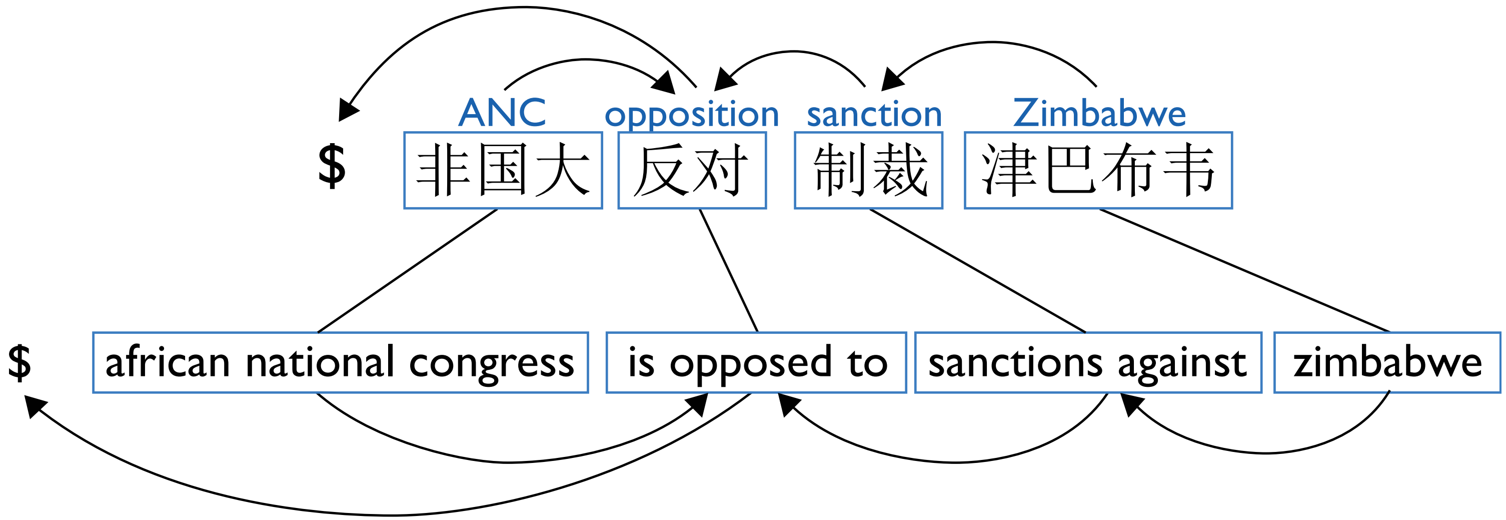
- We also include all Moses features

Left Child	Parent	Right Child	P(child | parent, direction)
,	there are		0.075
, and	there are		0.010
and	there are		0.009
but	there are		0.009
in	there are		0.008
at present	there are		0.006
...			
in addition to	there are		0.003
	there are	.	0.063
	there are	also	0.007
	there are	still	0.006
	there are	problems	0.004
	there are	people	0.004
		...	
	there are	differences between	0.002

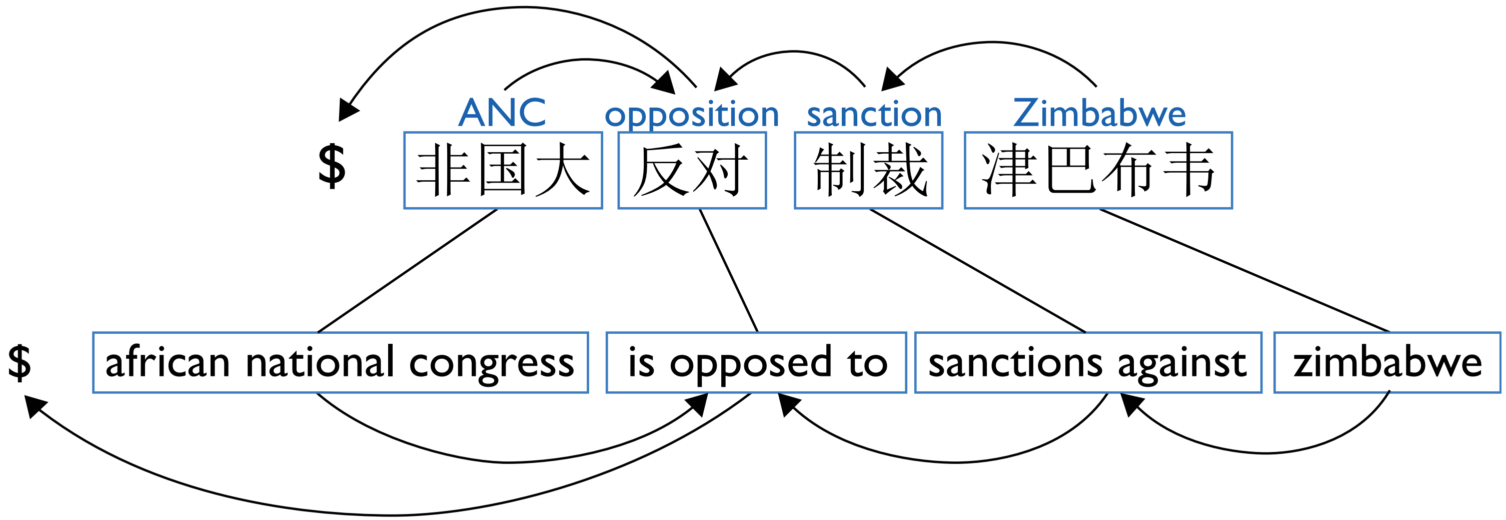
Root Phrases	P(root)
is	0.012
will	0.004
are	0.004
has	0.004
said	0.003
was	0.003
said :	0.003
said that	0.002
…	
has been	0.0007
he said	0.0007
pointed out that	0.0006

Root Phrases for Brown Clusters	P(root)
{is, shows, remains, provides}	0.014
{made, held, set, called}	0.012
{will, would, can, should}	0.009
{said, says, added, stressed}	0.006
{are, were, 're}	0.005
{said, says, added, stressed} that	0.004
{has, had}	0.004
{we, they, i, you} {will, would, can, should}	0.003
…	
{he, she} {said, says, added, stressed} that	0.001
{will, would, can, should} {also, still, already, always}	0.001
{also, still, already, always} {made, held, set, called}	0.001

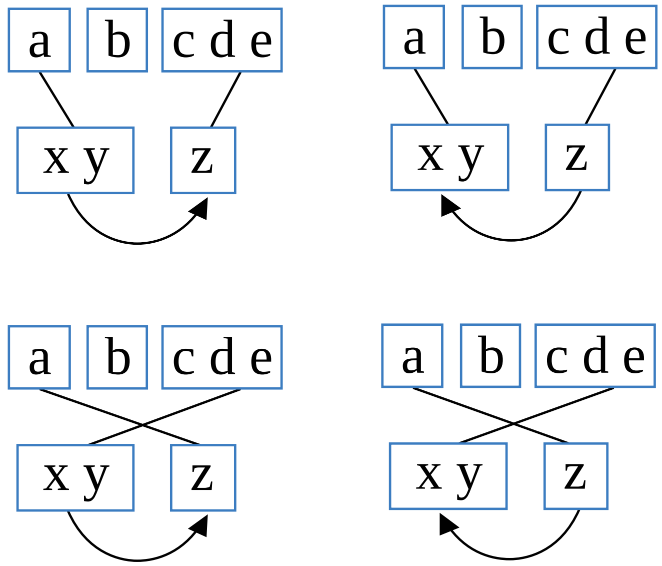
Configuration Features



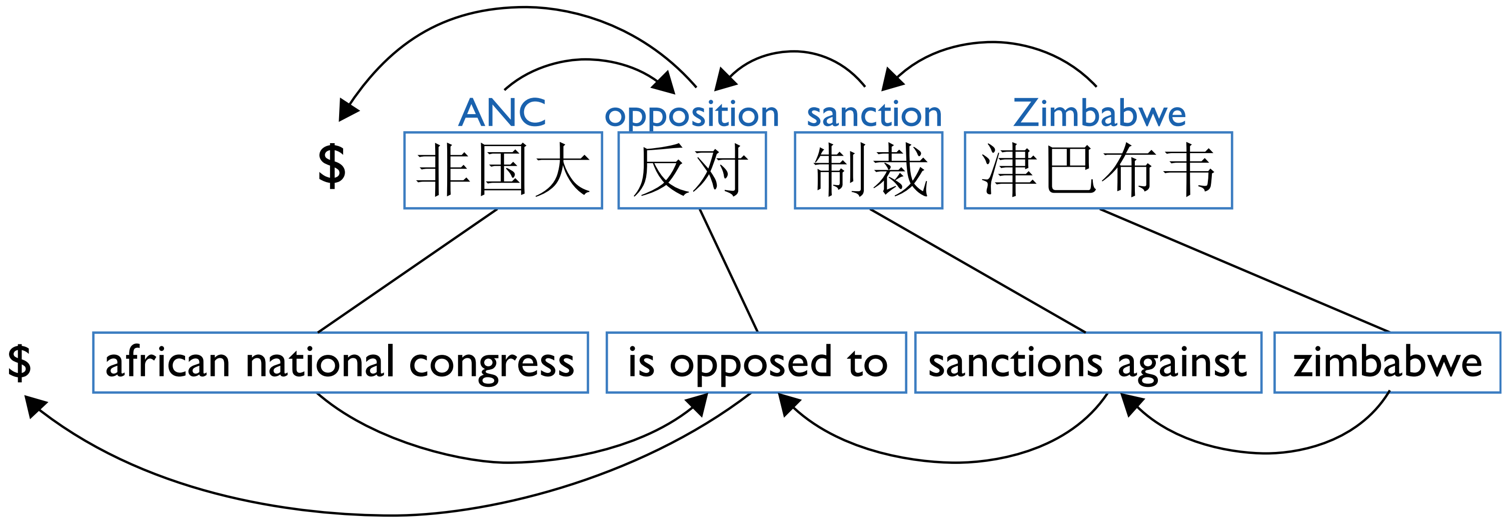
Configuration Features



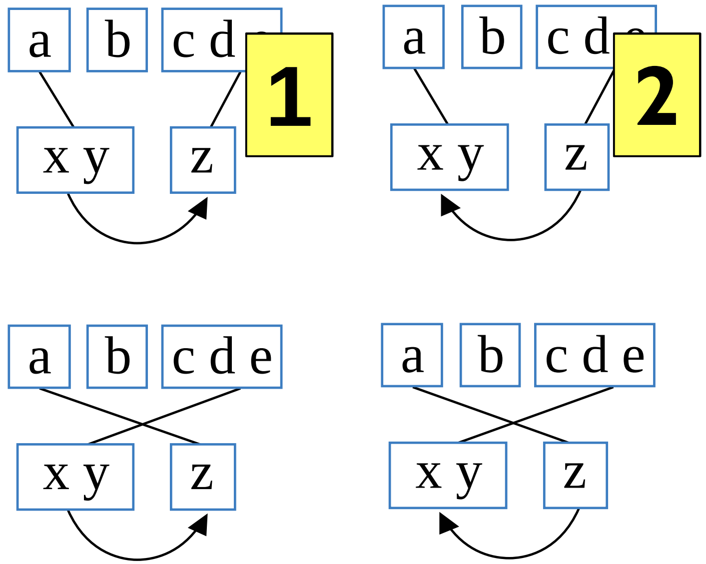
- String-to-tree configurations:
  - A feature to count each type of configuration

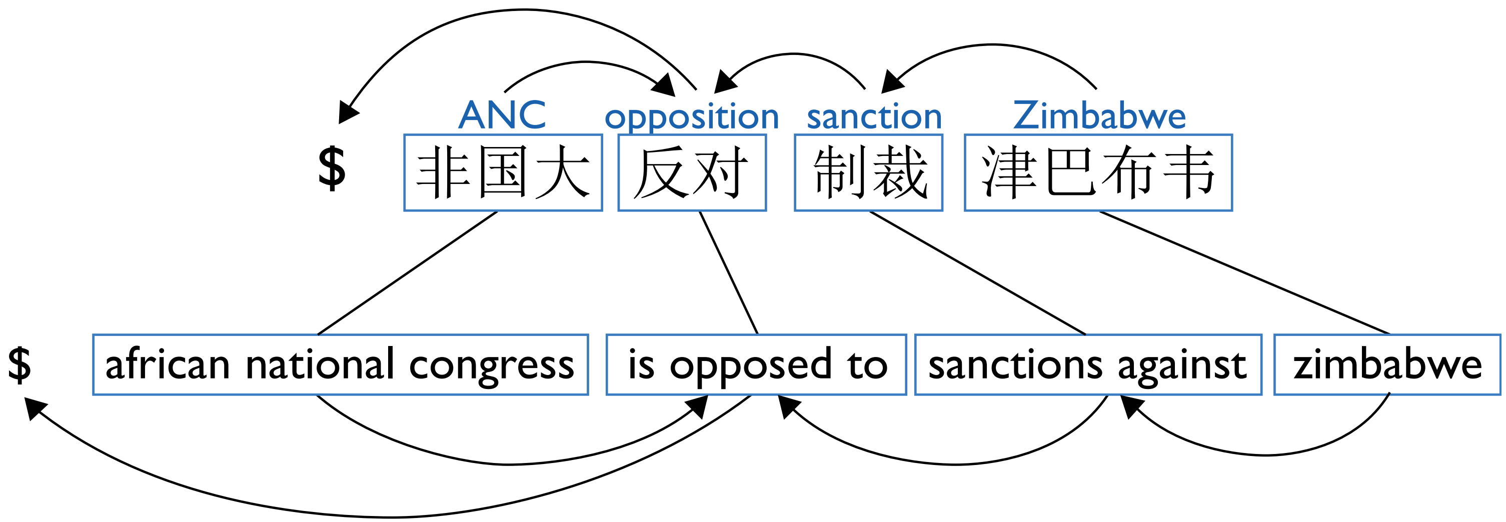


Configuration Features

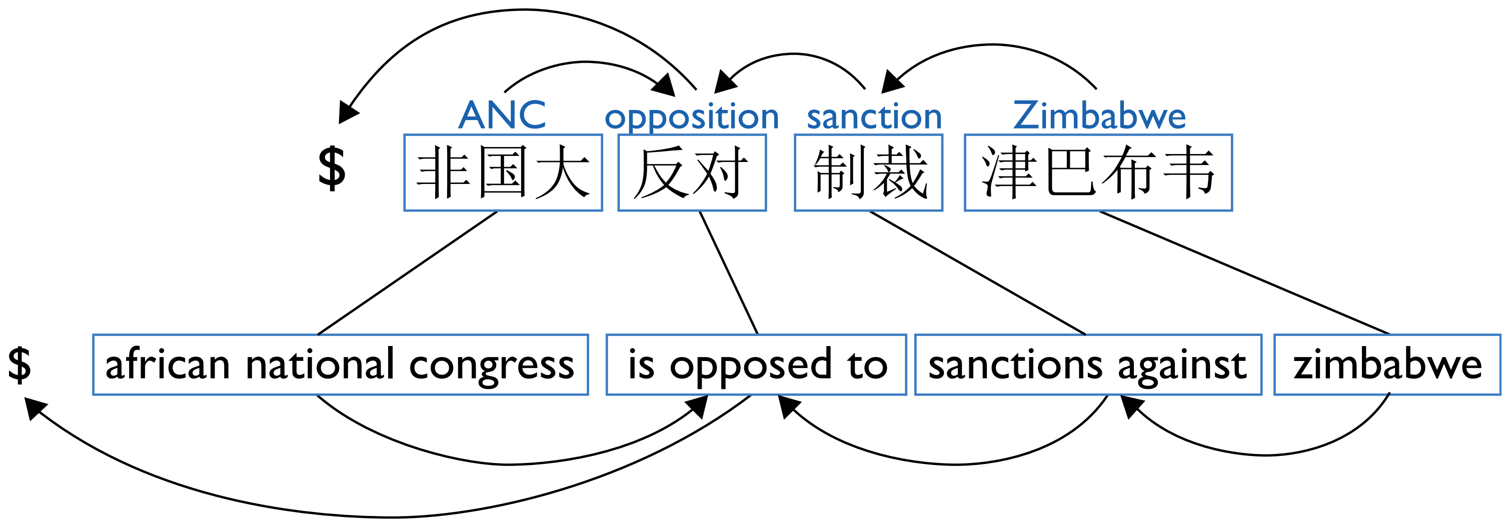


- String-to-tree configurations:
  - A feature to count each type of configuration



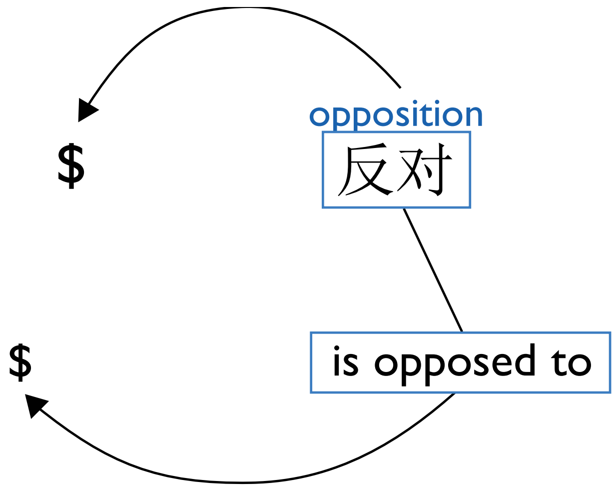


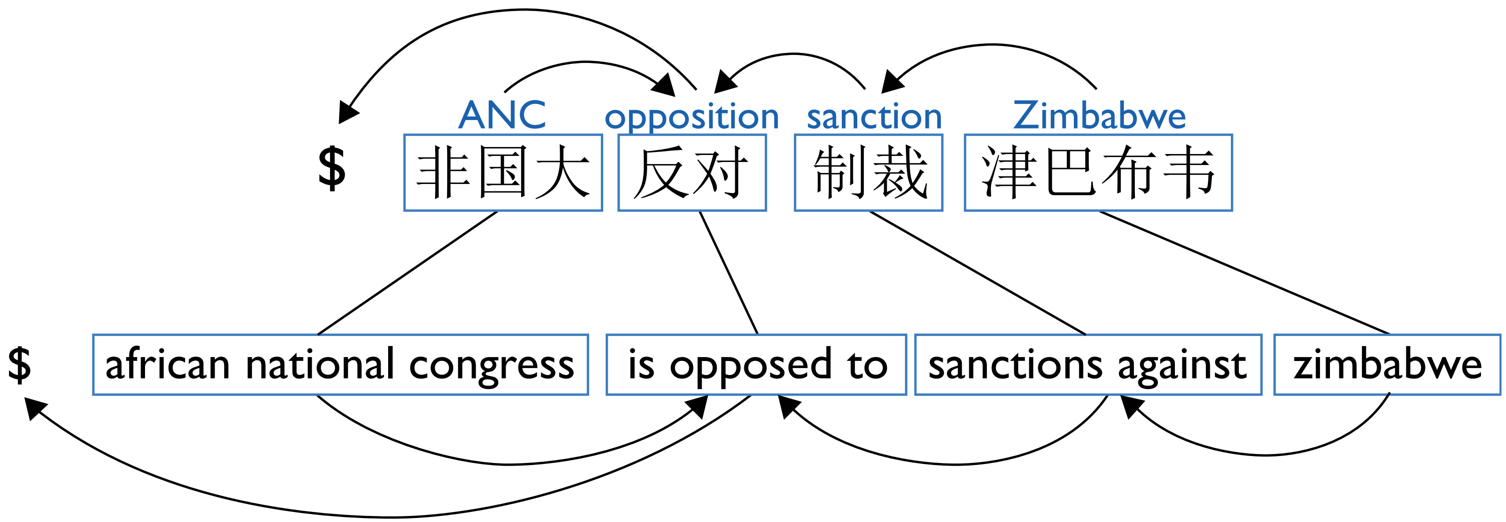
- Tree-to-tree configurations:
  - Quasi-synchronous features (Smith and Eisner, 2006)
  - These require a source-language parser



- Tree-to-tree configurations:
  - Quasi-synchronous features (Smith and Eisner, 2006)

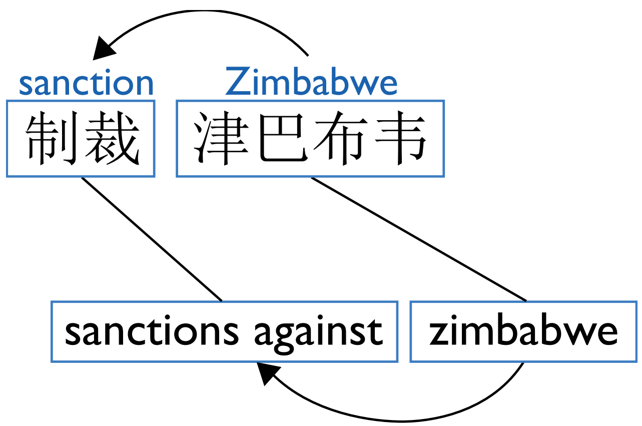
"Root-Root"





- Tree-to-tree configurations:
  - Quasi-synchronous features (Smith and Eisner, 2006)

"Parent-Child"



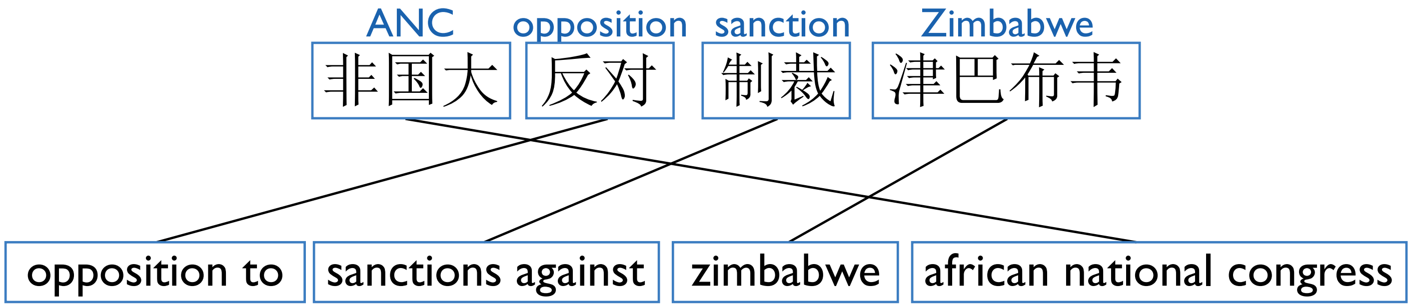
Two Questions

- How do we score dependency trees on phrases?
- How do we decode?

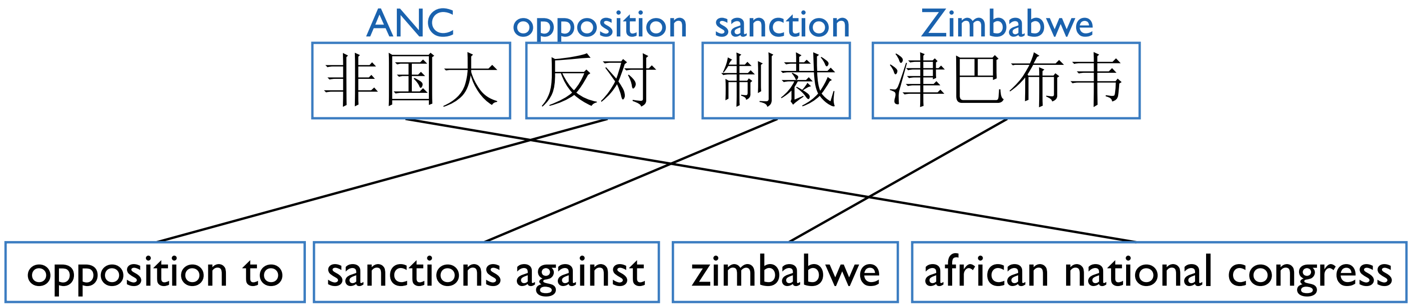
Two Questions

- How do we score dependency trees on phrases?

- How do we decode?
  - A coarse-to-fine algorithm based on lattice parsing of phrase lattices

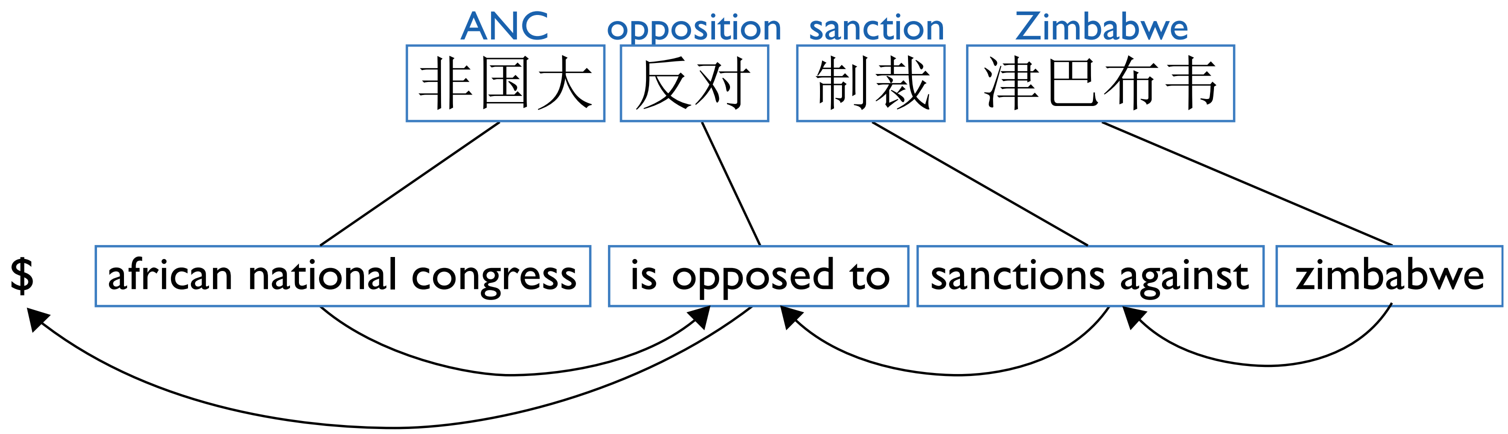


Reference: african national congress opposes sanctions against zimbabwe



Reference: african national congress opposes sanctions against zimbabwe

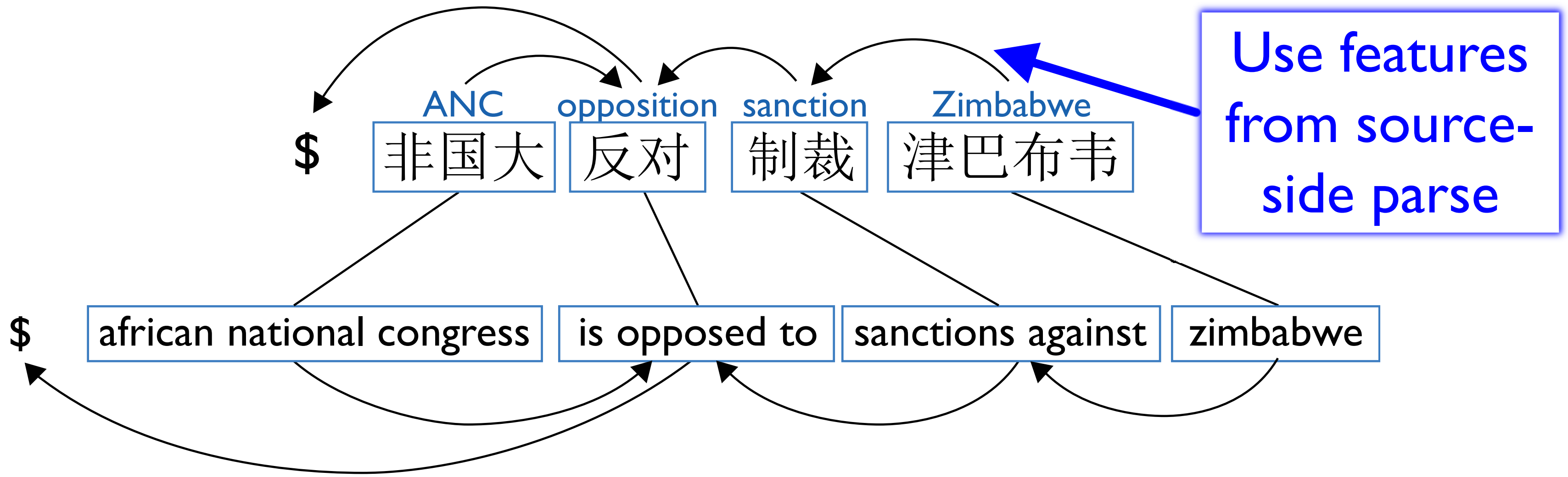
- For phrase-based decoding, search over:
  - Phrase segmentations
  - Translations for each phrase
  - Orderings of the translated phrases



Reference: african national congress opposes sanctions against zimbabwe

- For our model, search over:
  - Phrase segmentations
  - Translations for each phrase
  - Orderings of the translated phrases
  - Projective dependency trees on the phrases

ARK lti Carnegie Mellon

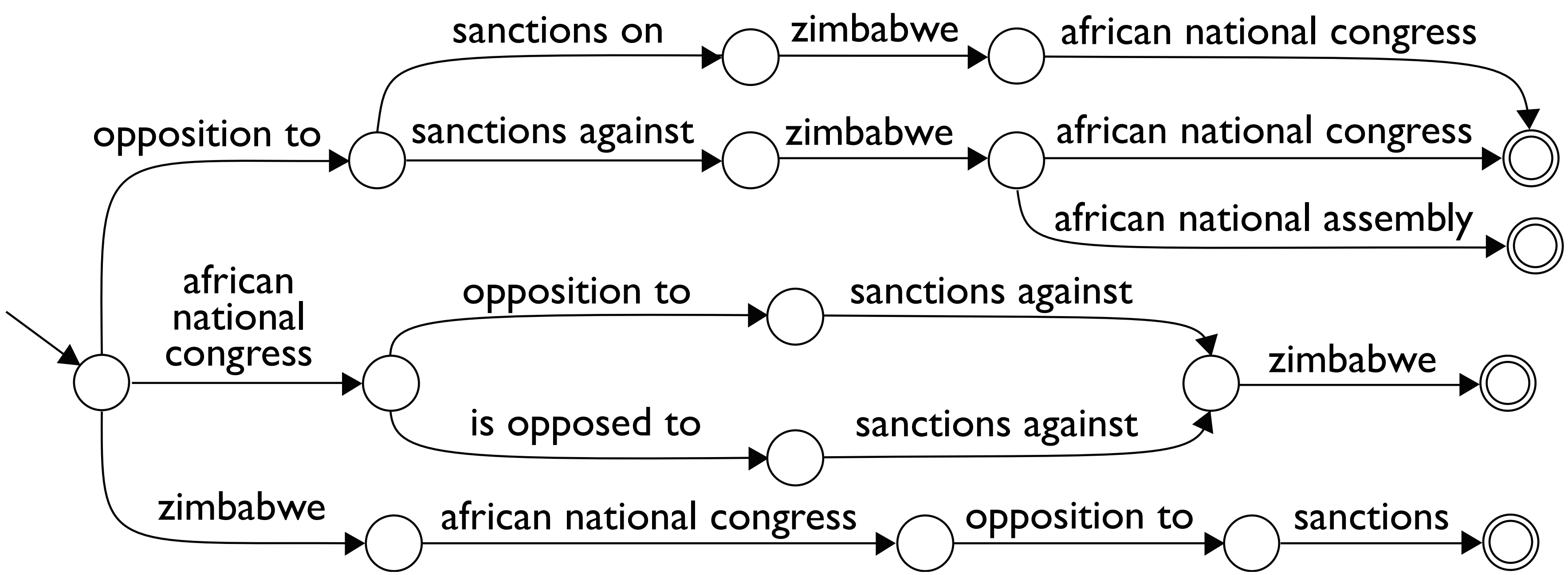


Reference: african national congress opposes sanctions against zimbabwe

- For our model, search over:
  - Phrase segmentations
  - Translations for each phrase
  - Orderings of the translated phrases
  - Projective dependency trees on the phrases

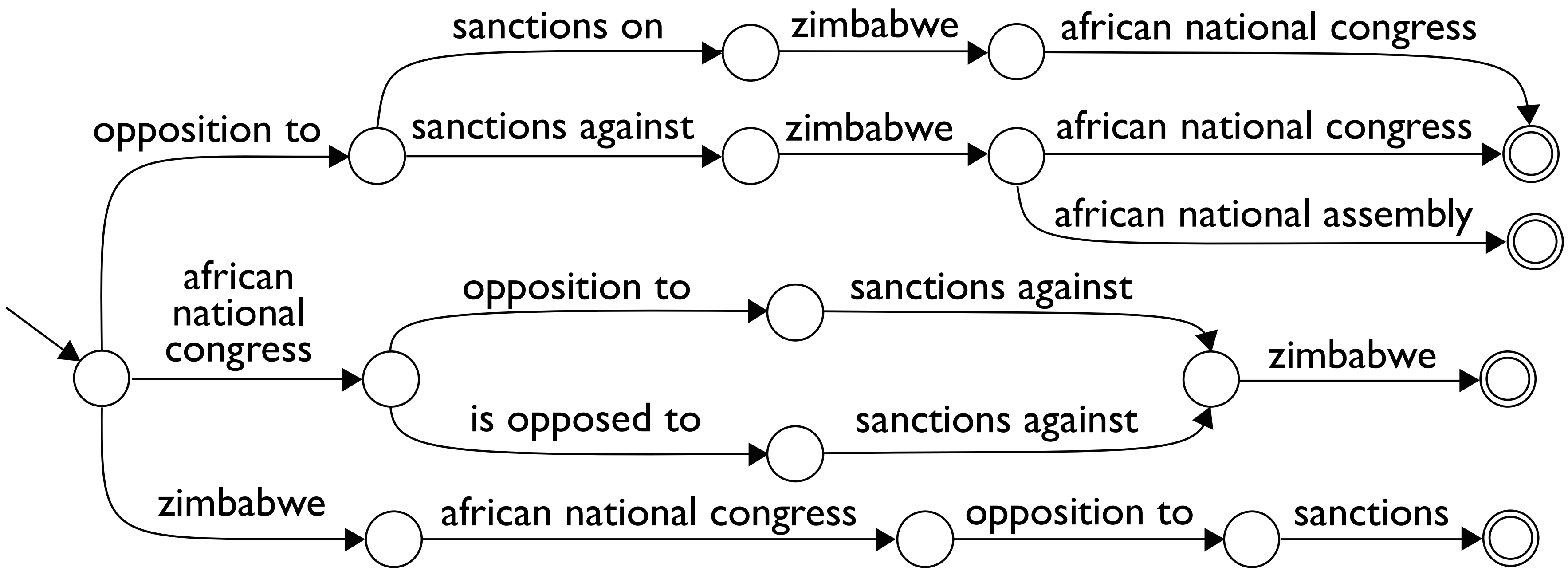
Coarse-to-Fine Decoding

- Pass 1: Generate phrase lattice



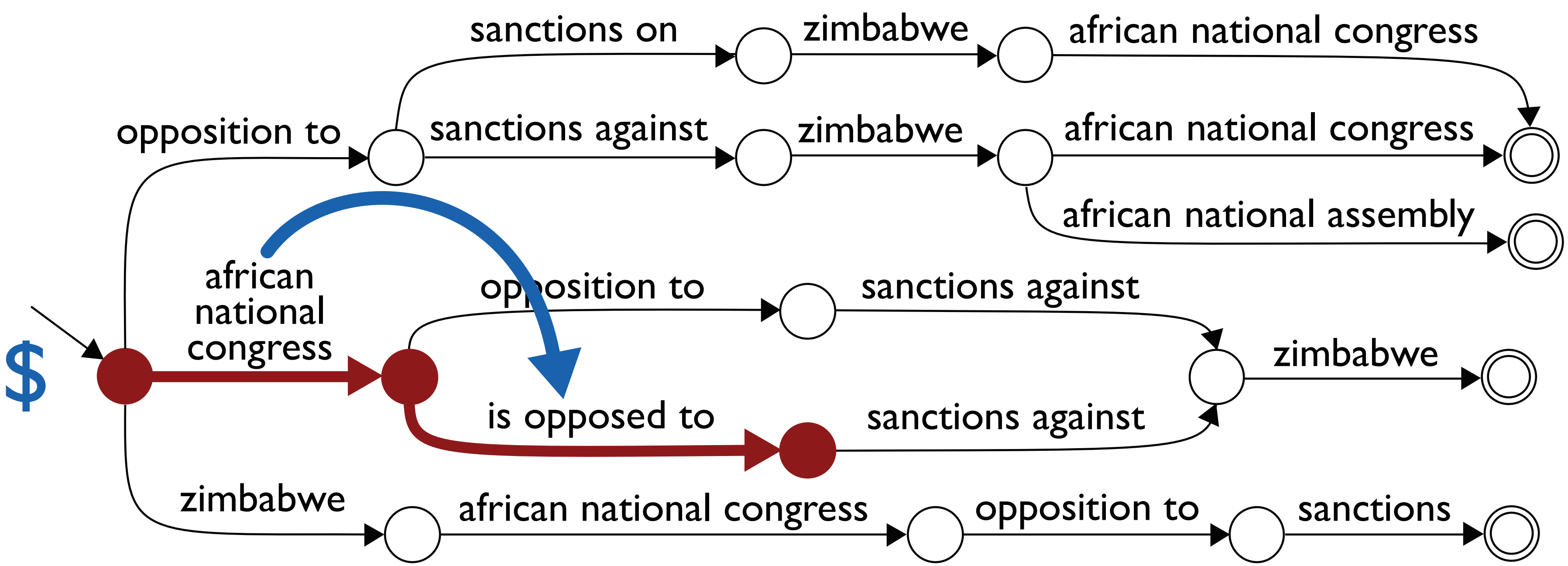
Reference: african national congress opposes sanctions against zimbabwe

- Pass 1: Generate phrase lattice
- Pass 2: Perform lattice parsing on phrase lattice
  - **Jointly** find best path + dependency tree



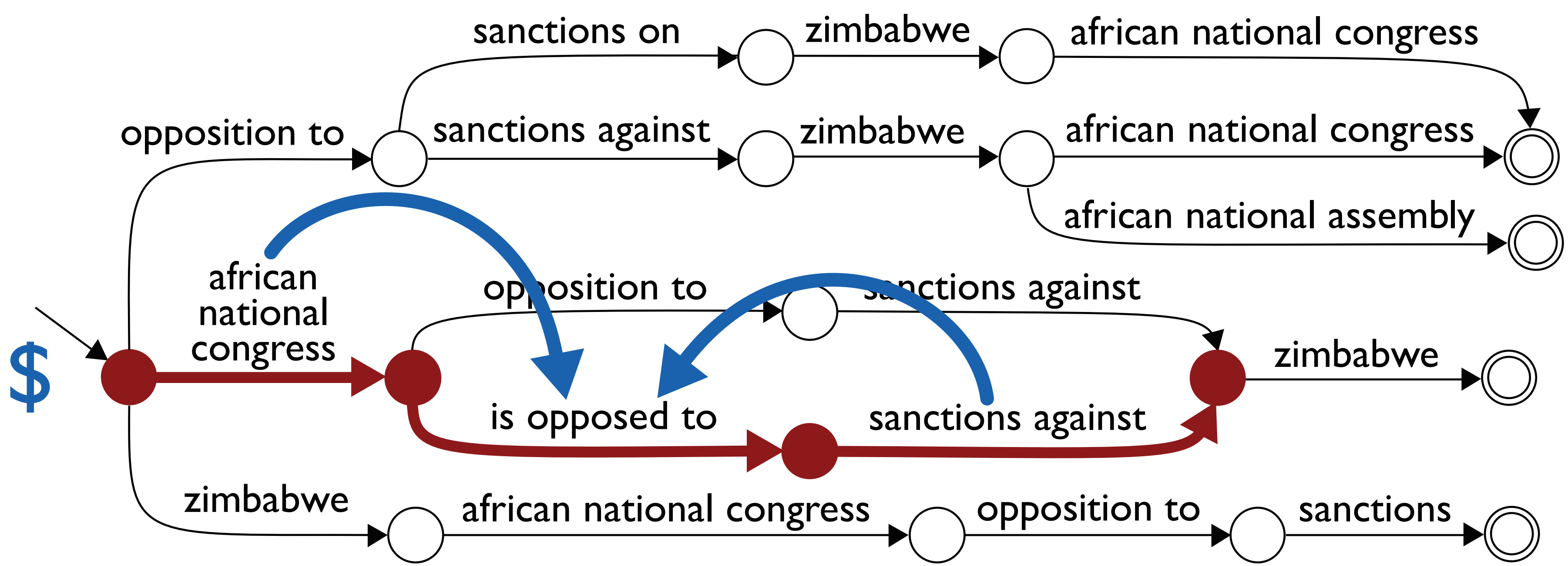
Reference: african national congress opposes sanctions against zimbabwe

- Pass 1: Generate phrase lattice
- Pass 2: Perform lattice parsing on phrase lattice
  - **Jointly** find best path + dependency tree



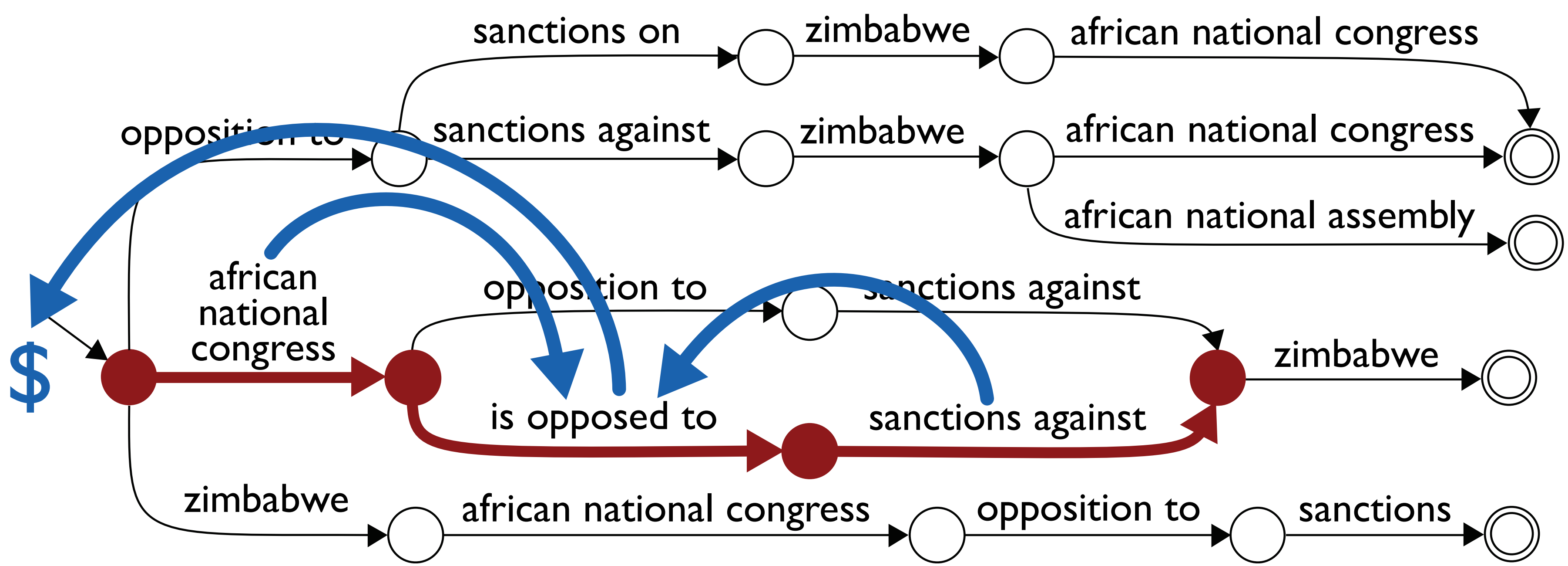
Reference: african national congress opposes sanctions against zimbabwe

- Pass 1: Generate phrase lattice
- Pass 2: Perform lattice parsing on phrase lattice
  - **Jointly** find best path + dependency tree



Reference: african national congress opposes sanctions against zimbabwe

- Pass 1: Generate phrase lattice
- Pass 2: Perform lattice parsing on phrase lattice
  - **Jointly** find best path + dependency tree

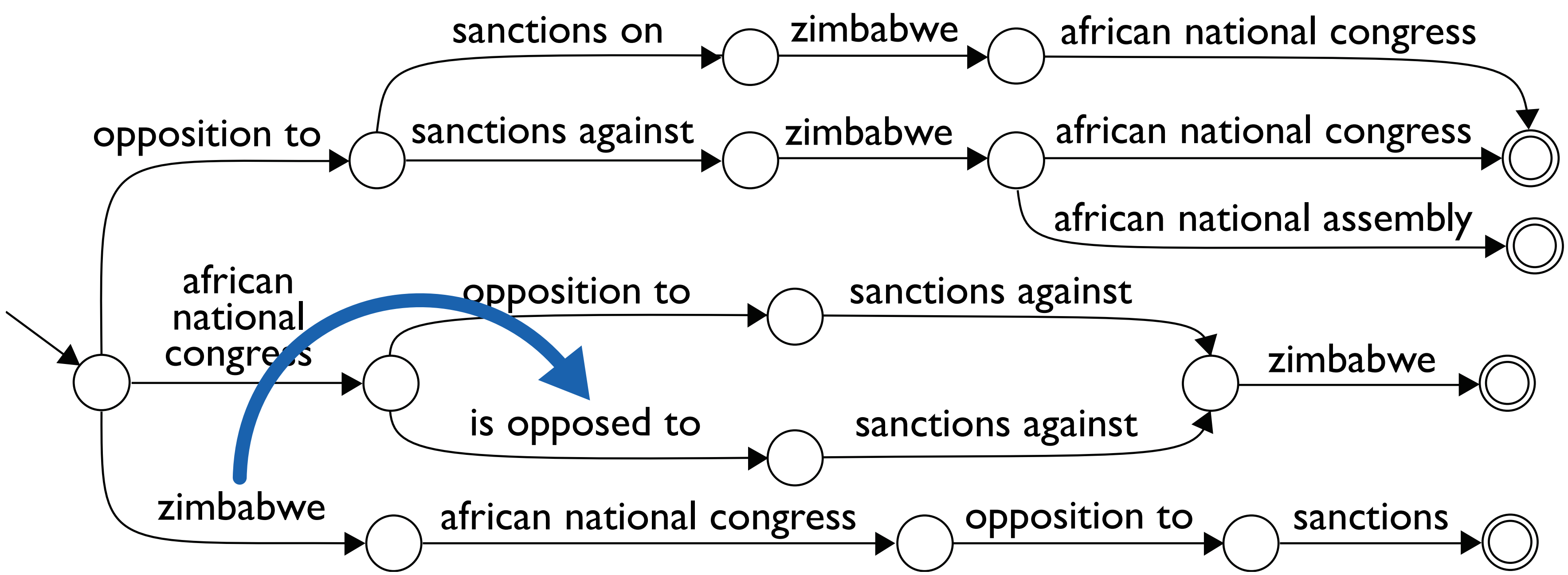


Reference: african national congress opposes sanctions against zimbabwe

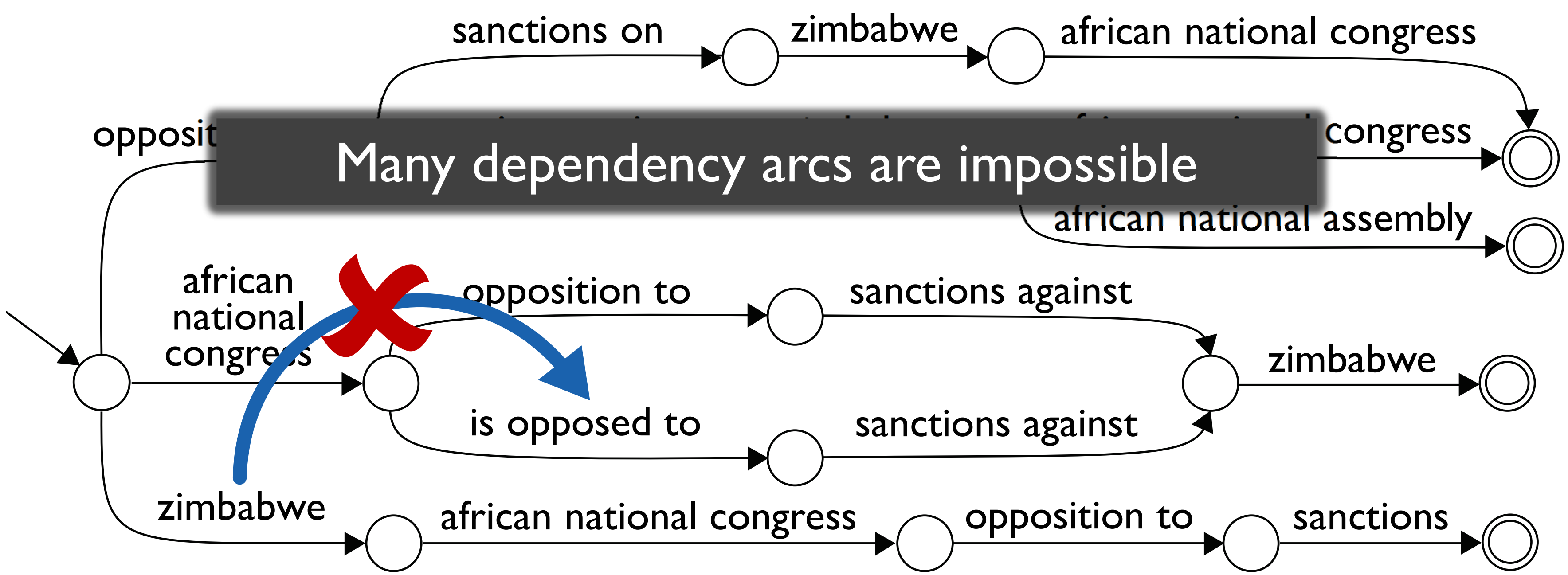
- Pass 1: Generate phrase lattice
- Pass 2: Perform lattice parsing on phrase lattice
  - Requires $O(E^2V)$ time (E = # edges, V = # vertices in lattice)
  - We prune the lattices with forward-backward pruning so that $E \approx 1000$ (lattices still have 10^{15} derivations on average)
  - But still prohibitively expensive to do exact search

- Pass 1: Generate phrase lattice
- Pass 2a: Filter candidate dependency arcs
- Pass 2b: Lattice parsing using arcs that remain

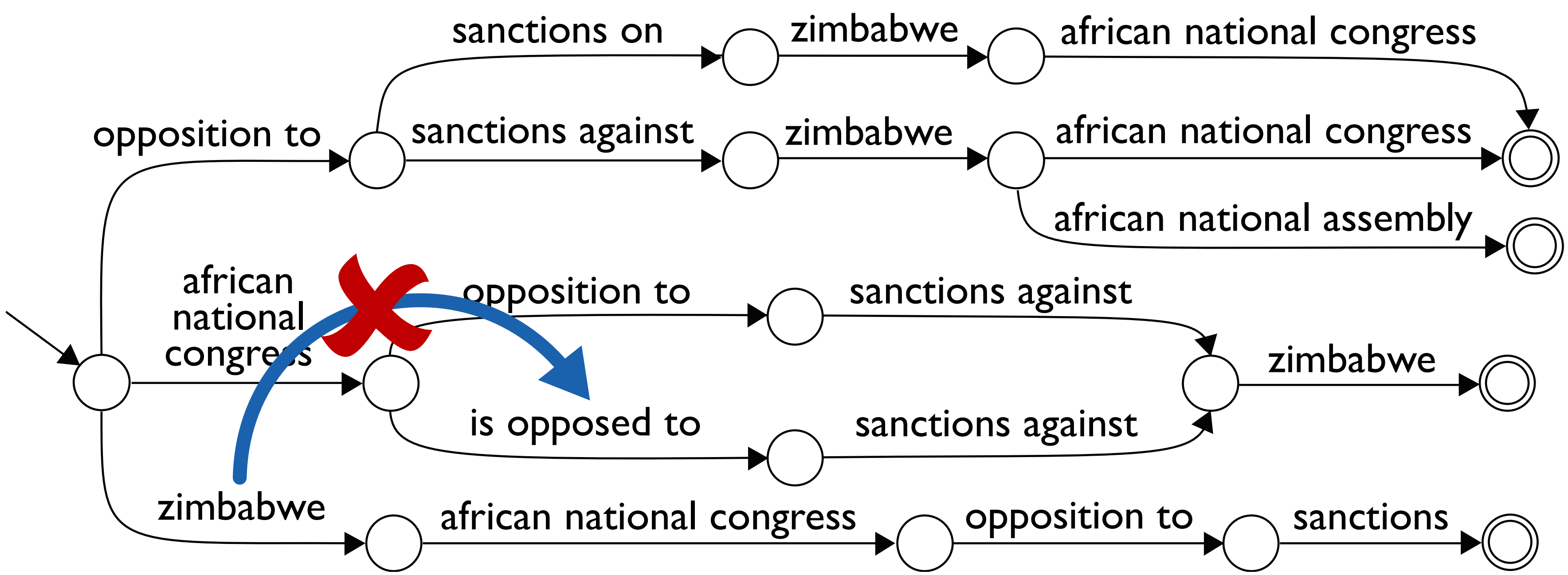
- Pass 1: Generate phrase lattice
- Pass 2a: Filter candidate dependency arcs



- Pass 1: Generate phrase lattice
- **Pass 2a: Filter candidate dependency arcs**



- Pass 2a: Filter candidate dependency arcs
  - Floyd-Warshall algorithm used to determine reachability, impossible arcs filtered (not lossy)



- Pass 2a: Filter candidate dependency arcs
  - Floyd-Warshall algorithm used to determine reachability, impossible arcs filtered (not lossy)
  - Remaining arcs ranked using full feature set, arcs for top 300 parents for each child are kept; all others filtered (lossy)

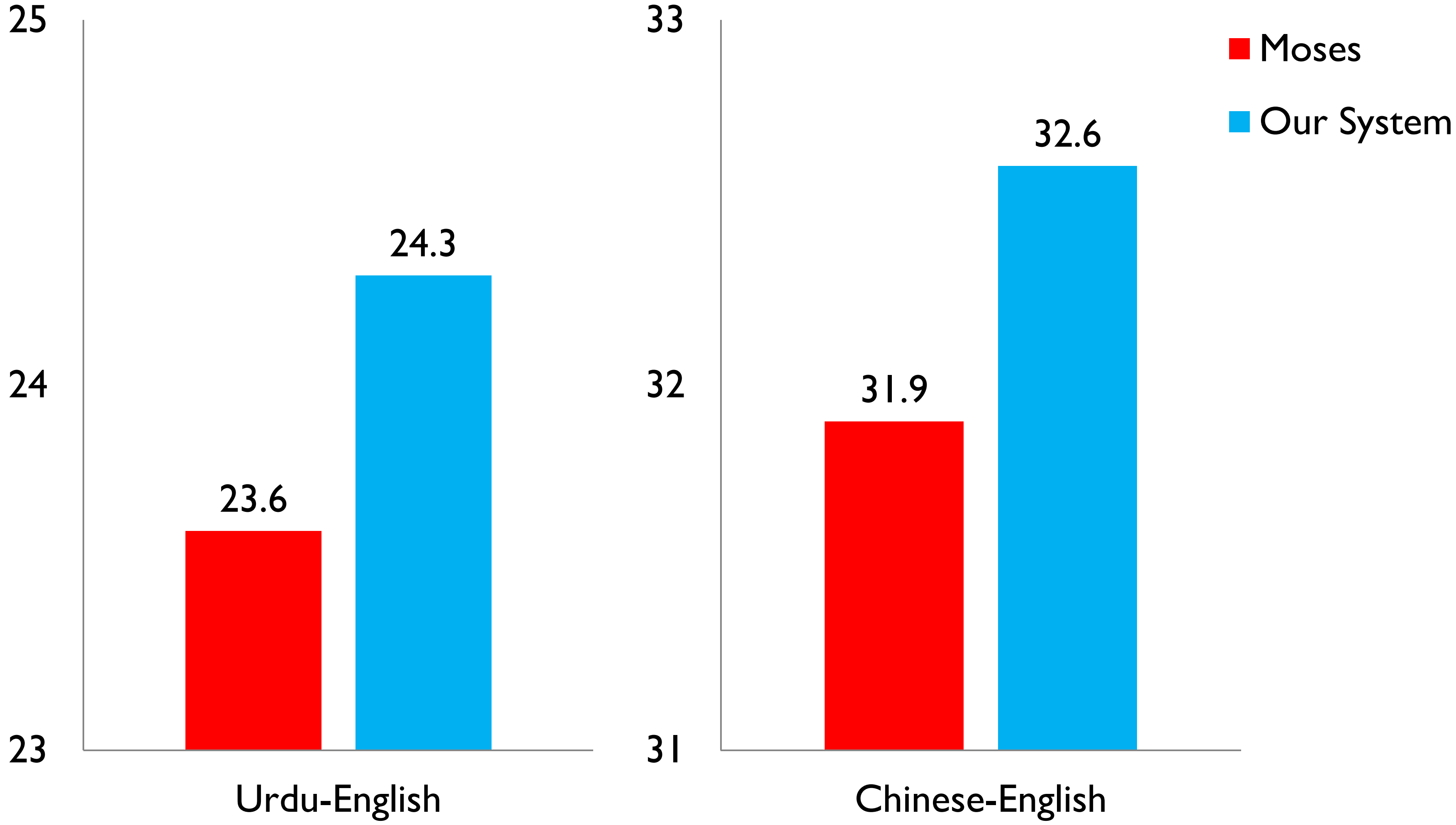
- Pass 2a: Filter candidate dependency arcs
  - Floyd-Warshall algorithm used to determine reachability, impossible arcs filtered (not lossy)
  - Remaining arcs ranked using full feature set, arcs for top 300 parents for each child are kept; all others filtered (lossy)
- Pass 2b: Lattice parsing using arcs that remain
  - Extension of edge-factored dependency parsing algorithm (Eisner, 1996)
  - Agenda algorithm + heuristic search: Floyd-Warshall scores used to weight agenda items

Experiments

Experimental Setup

- Chinese-English
  - 300k sentence pairs from FBIS corpus (~9M words)
  - Tuned on MT03, tested on MT02, MT05, and MT06
- Urdu-English
  - Data from NIST08 evaluation (~1M words)
  - Tuned on MT08, tested on MT09
- Trigram language models trained on 200M words
- MERT run 3 times
- Chinese parsed with Stanford Parser (Levy & Manning, 2003)
- English parsed with TurboParser (Martins et al., 2009)
- Baseline: Moses
  - Max phrase len 7, distortion limit 10, lexicalized reordering model
  - Used to generate phrase lattices for our system

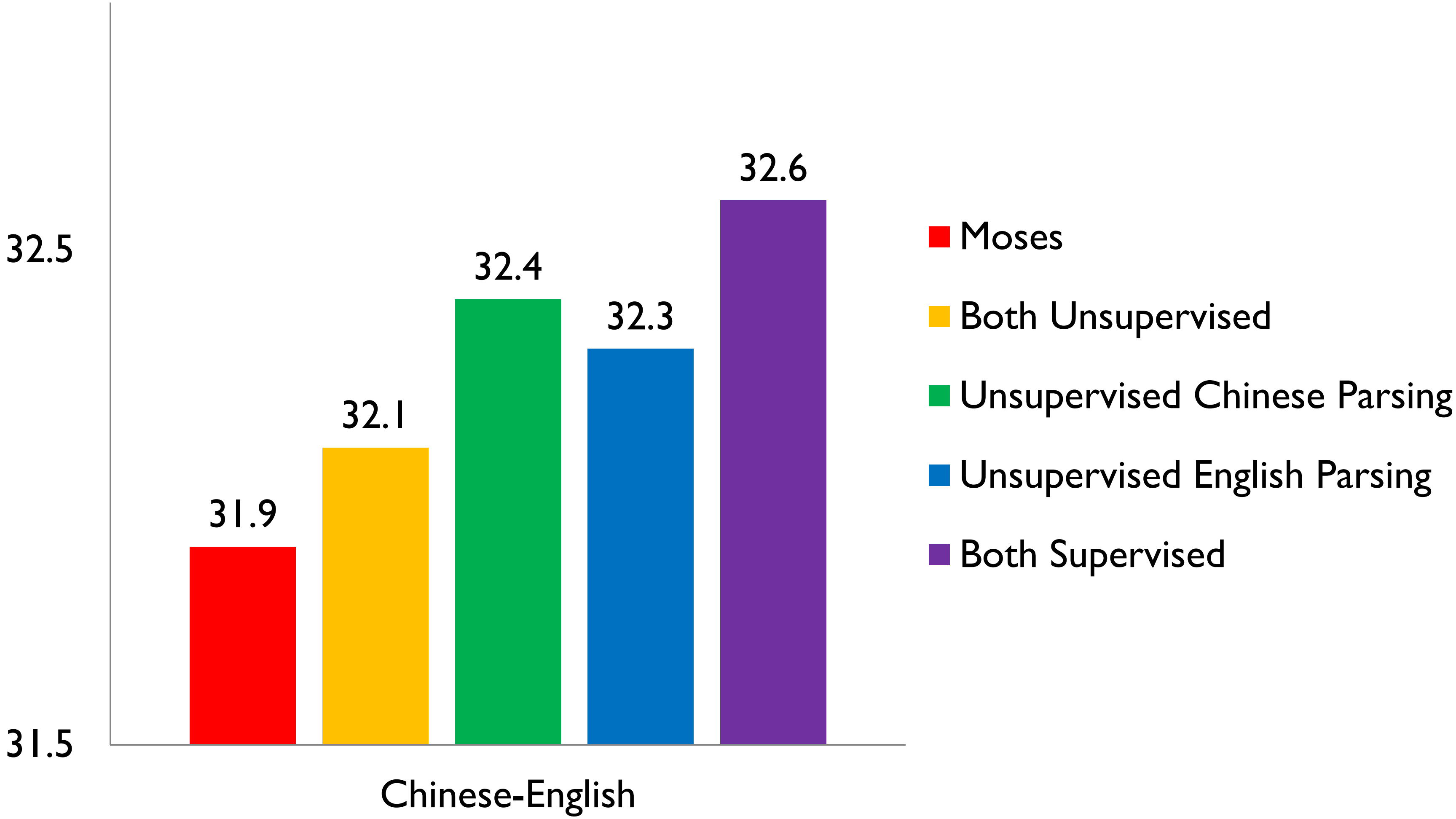
Results


25
24
23
24.3
23.6
Urdu-English
33
32
31
32.6
31.9
Chinese-English
Moses
Our System

- Our best results use supervised parsers for both source and target languages
- What about **unsupervised** parsing?

- Our best results use supervised parsers for both source and target languages
- What about **unsupervised** parsing?
  - We use the dependency model with valence (DMV; Klein & Manning, 2004)
  - When initialized with a model with a concave log-likelihood function, DMV approaches state-of-the-art performance (Gimpel & Smith, 2011)
    - 53.1% attachment accuracy on Penn Treebank
    - 44.4% on Chinese Treebank

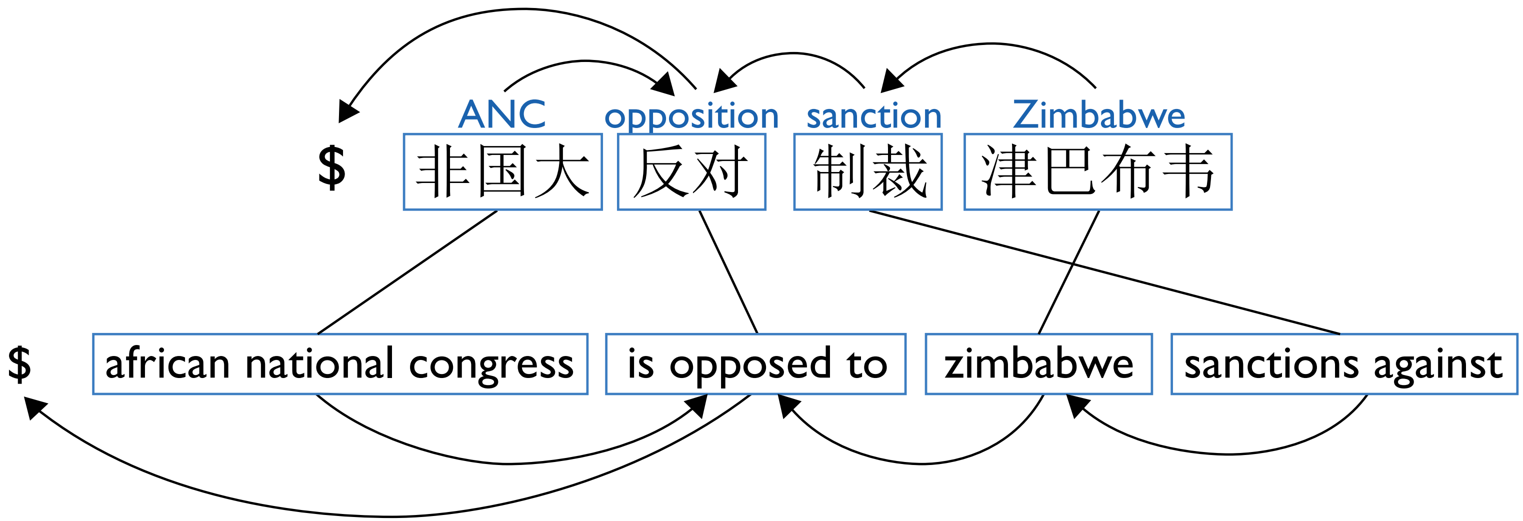
Using Unsupervised Parsers



Conclusions

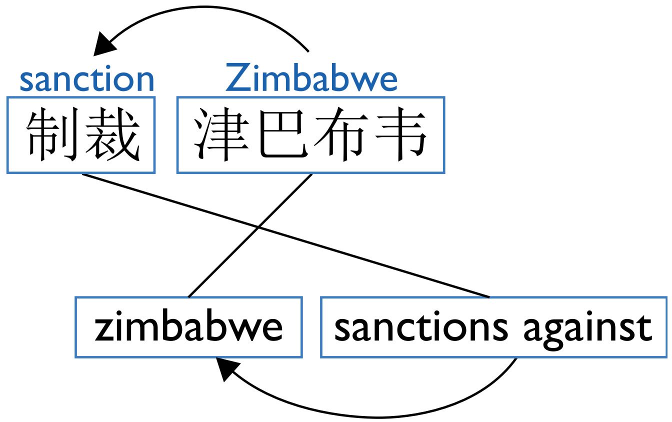
- A translation model built on a dependency grammar on phrases
- Features for scoring phrase dependency trees and algorithms for decoding
- Encouraging experimental results for Urdu-English and Chinese-English translation, including unsupervised dependency parsing

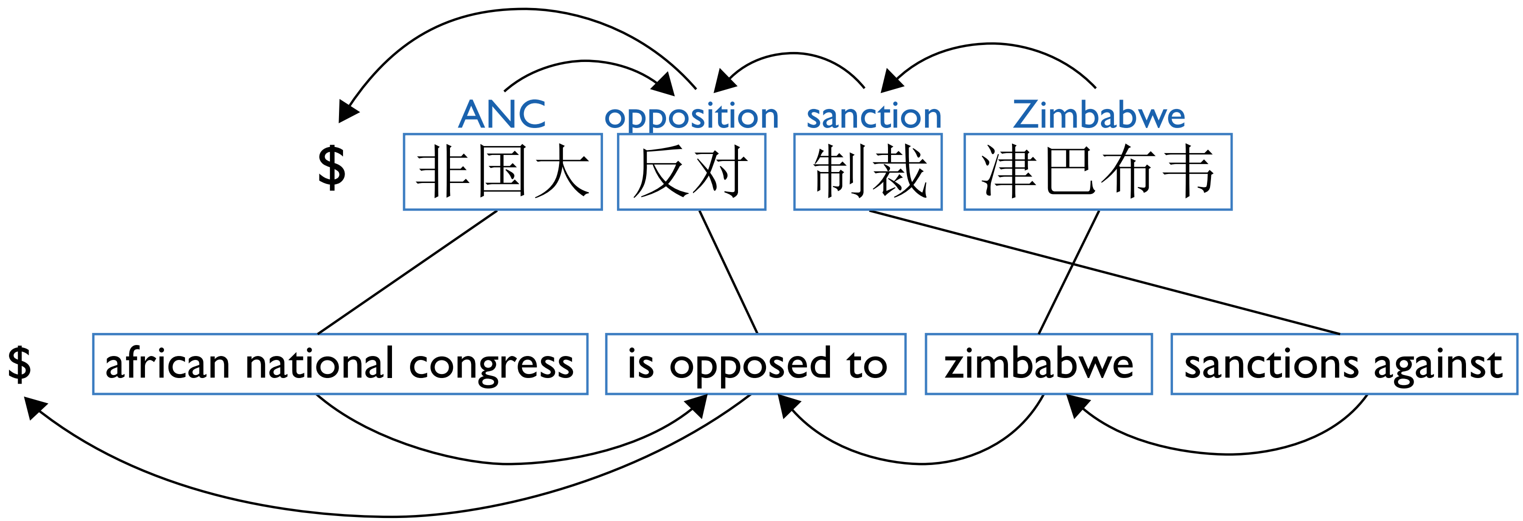
Thanks!



- Tree-to-tree configurations:
  - Quasi-synchronous features (Smith and Eisner, 2006)

"Child-Parent"





- Tree-to-tree configurations:
  - Quasi-synchronous features (Smith and Eisner, 2006)

“Grandparent-Grandchild”