

# Interpreting Anaphoric Shell Nouns using Antecedents of Cataphoric Shell Nouns

**Varada Kolhatkar and Graeme Hirst**  
*University of Toronto*



**Heike Zinsmeister**  
*Universität Hamburg*



# Anaphoric Shell Nouns (ASNs)

(Schmid, 2000)

Our goal: Automatically interpreting anaphoric phrases such as *this issue* and *this fact*.

The municipal council will have to decide whether to balance the budget by raising revenue or cutting spending. The council will have to come to a resolution by the end of the month. **This issue** is dividing communities across the country.

# Anaphoric Shell Nouns (ASNs)

(Schmid, 2000)

Our goal: Automatically interpreting anaphoric phrases such as *this issue* and *this fact*.

antecedent

The municipal council will have to decide whether to balance the budget by raising revenue or cutting spending. The council will have to come to a resolution by the end of the month. This issue is dividing communities across the country.

anaphoric shell noun (ASN)

# Why ASN interpretation?

# Ubiquity of Shell Nouns

- Occur frequently in all kinds of texts
- *fact, idea, problem*: among 100 most frequently occurring nouns in the BNC  
(Schmid, 2000)
- ~25 million occurrences in the NYT corpus  
(~1.3 billion tokens)

# Organizing Discourse



The municipal council will have to decide whether to balance the budget by raising revenue or cutting spending. The council will have to come to a resolution by the end of the month. This issue is dividing communities across the country.

- Characterize and label complex information
- Cohesive devices
- Topic boundary markers

# Gap in the current research

# Limited Work in Resolving ASNs

- Most work: nominal anaphora
- Approaches for non-nominal antecedents  
(Eckert and Strube, 2000; Byron, 2004; Müller, 2008;  
Kolhatkar and Hirst, 2012)
- Domain-specific
- Limited syntactic types



# Lack of Annotated Data

Corpus	ASN instances
Poesio and Artstein, 2008 ARRAU	455 abstract anaphor instances, very few ASN instances
Kolhatkar and Hirst, 2012	188 <i>this issue</i> instances from Medline abstracts
Botley, 2006	462 ASN instances (not available)

No large-scale annotated corpus available

Need an approach that does not rely on  
manual annotation

# An observation

# Cataphoric Shell Nouns

cataphoric shell noun (CSN)

Of course, the central, and probably insoluble, **issue** is **whether animal testing is cruel.**

antecedent

# Cataphoric Shell Nouns

cataphoric shell noun (CSN)

Of course, the central, and probably insoluble, **issue** is **whether animal testing is cruel.**

antecedent

- Resemblance to ASNs: antecedents encode similar kind of complex abstract objects
- Contrast with ASNs: **easy to interpret** using syntactic structure alone

Can we use properties of  
CSN antecedents to  
interpret ASNs?

# Hypothesis of our Work

CSN antecedents and ASN antecedents share some linguistic properties, and hence linguistic knowledge encoded in CSN antecedents will help in interpreting ASNs.

- Antecedents of both represent the general notion of the shell noun (e.g., the notion of an *issue* is something that is unresolved)
- There are patterns, for example, of syntactic shape or words when people state *issues* or *facts*

# Hypothesis of our Work

## ASN example

The municipal council had to decide whether to balance the budget by raising revenue or cutting spending. The council had to come to a resolution by the end of the month. This issue was dividing communities across the country.

*whether* clause in both cases

## CSN example

Of course, the central, and probably insoluble, issue is whether animal testing is cruel.

# Behaviour of Shell Nouns



# CSN Patterns

(Schmid, 2000)

Pattern	Example
<i>N-to</i>	Several people at the group said the <b>decision to</b> <i>write the letters</i> was not controversial internally.
<i>N-be-to</i>	The principal <b>reason is to</b> <i>create a representative government rather than to select the most talented person.</i>
<i>N-that</i>	Mr. Shoval left open the <b>possibility that</b> <i>Israel would move into other West Bank cities.</i>
<i>N-be-that</i>	The simple and reassuring <b>fact is that</b> a future generation of leaders is seeking new challenges during challenging times .
<i>N-wh</i>	There is now some <b>question whether</b> <i>the country was ever really in a recession.</i>
<i>N-be-wh</i>	Of course, the central, and probably insoluble, <b>issue is whether</b> <i>animal testing is cruel.</i>

# Shell Nouns: Selection

Schmid's  
Category

Examples

Factual

***fact, problem, reason***

Linguistic

***news, proposal, question***

Mental

***idea, belief, decision, issue***

Modal

***possibility, need, trend***

High-frequency nouns from each category

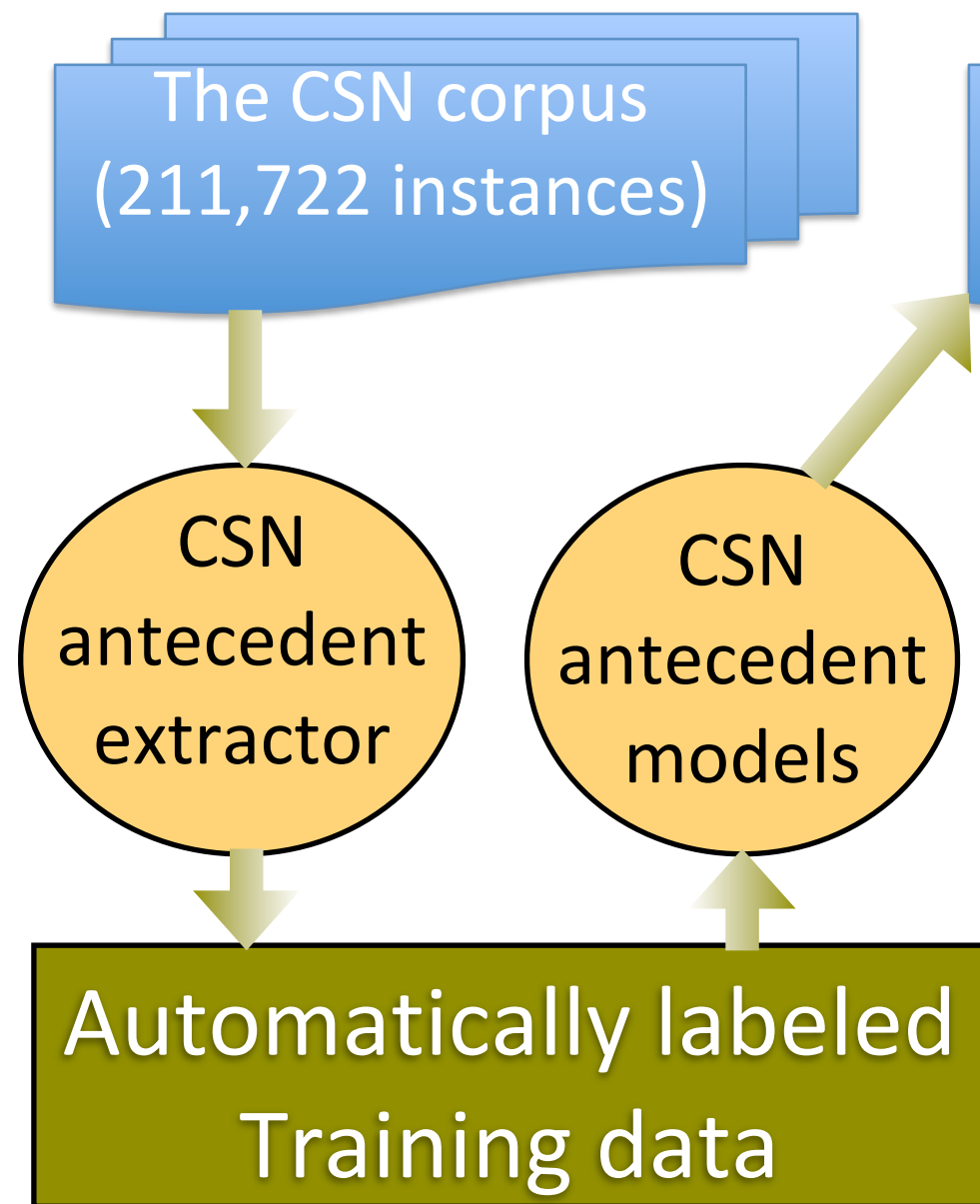
# Extraction Rules

Pattern	<i>fact</i>	<i>reason</i>	<i>issue</i>	<i>decision</i>	<i>question</i>	<i>possibility</i>
<i>N-to</i>	–	–	–	<i>inf</i> clause	predicate	<i>inf</i> clause
<i>N-be-to</i>	–	<i>inf</i> clause	<i>inf</i> clause	<i>inf</i> clause	<i>inf/wh</i> clause	<i>inf</i> clause
<i>N-that</i>	<i>that</i> clause	predicate	predicate	–	predicate	<i>that</i> clause
<i>N-be-that</i>	<i>that</i> clause	<i>that</i> clause	<i>that</i> clause	<i>that</i> clause	<i>that</i> clause	<i>that</i> clause
<i>N-wh</i>	–	predicate	<i>wh</i> clause	<i>wh</i> clause	<i>wh</i> clause	–
<i>N-be-wh</i>	<i>wh</i> clause	<i>wh</i> clause	<i>wh</i> clause	<i>wh</i> clause	<i>wh</i> clause	<i>wh</i> clause

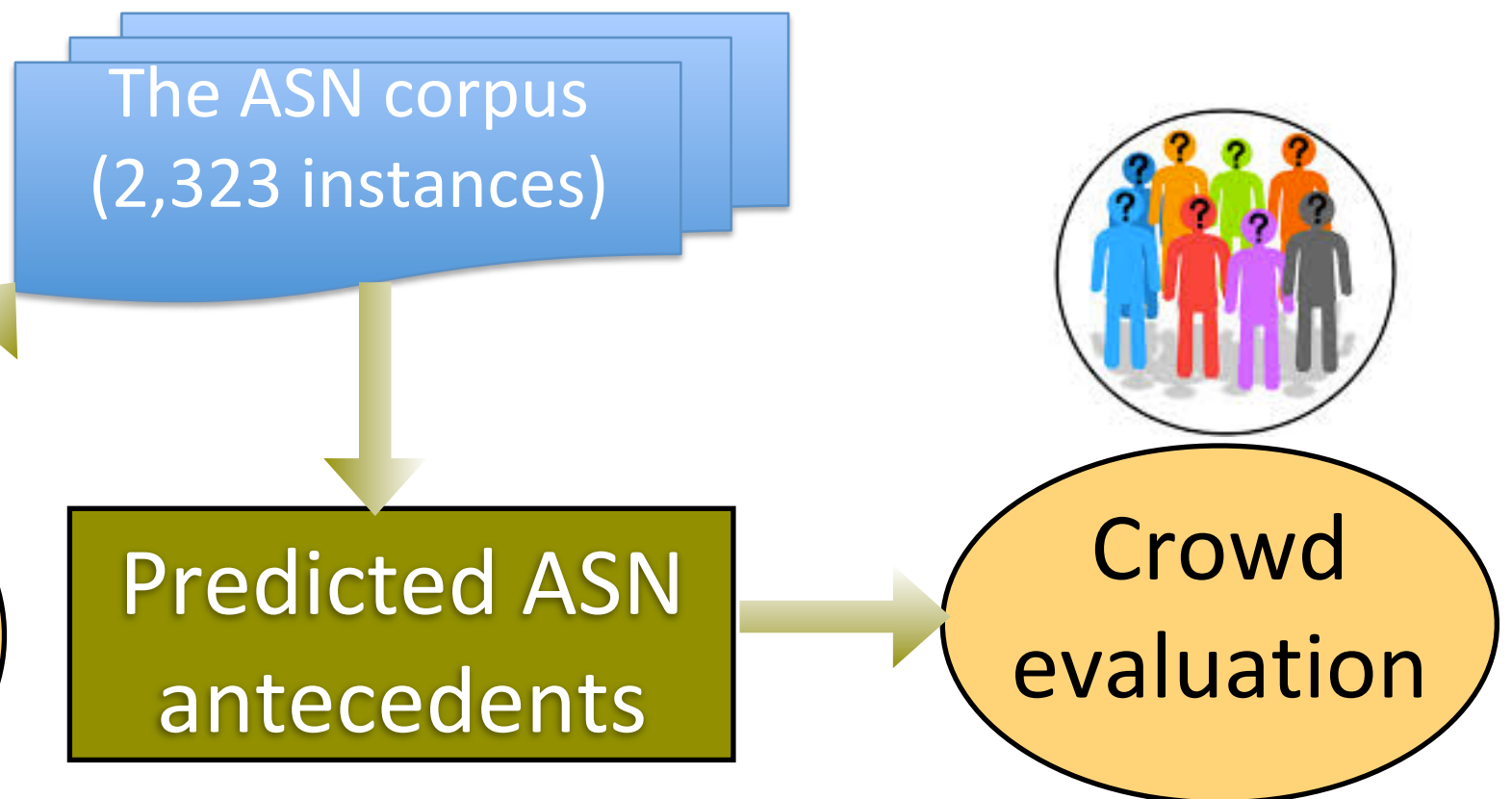
# Our approach

# Overview

## Training



## Testing



# Training

# The CSN Corpus

- Base corpus: The NYT corpus  
(Sandhaus, 2008)
- 211,722 sentences following CSN patterns

Example:

*Of course, the central, and probably insoluble, **issue is whether** animal testing is cruel.*

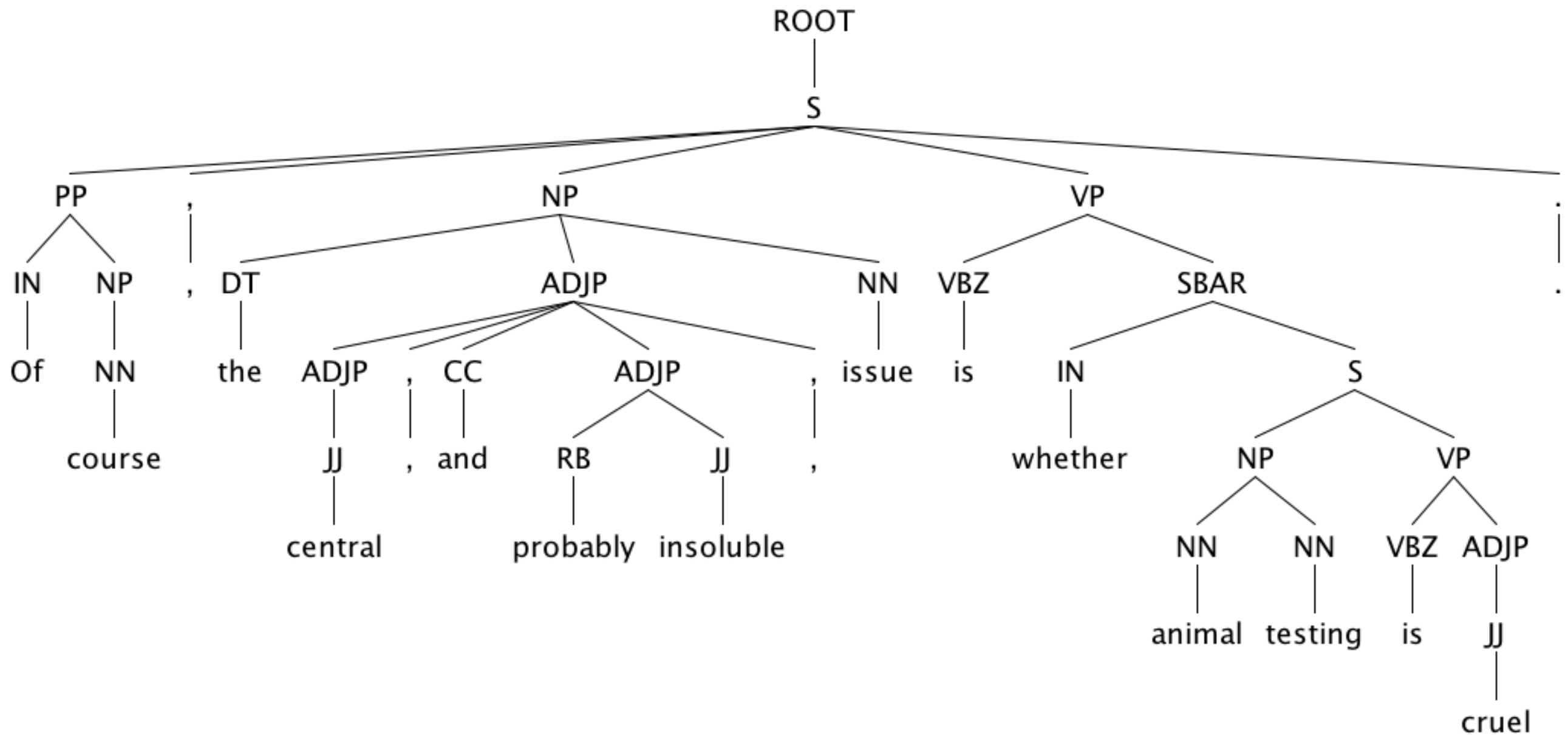
# CSN Antecedent Extractor

Goal: To create automatically labeled CSN antecedent data

- Parse each sentence using the Stanford parser (Klein and Manning, 2003)
- Apply manually derived extraction rules
- Extract appropriate syntactic constituent as the antecedent

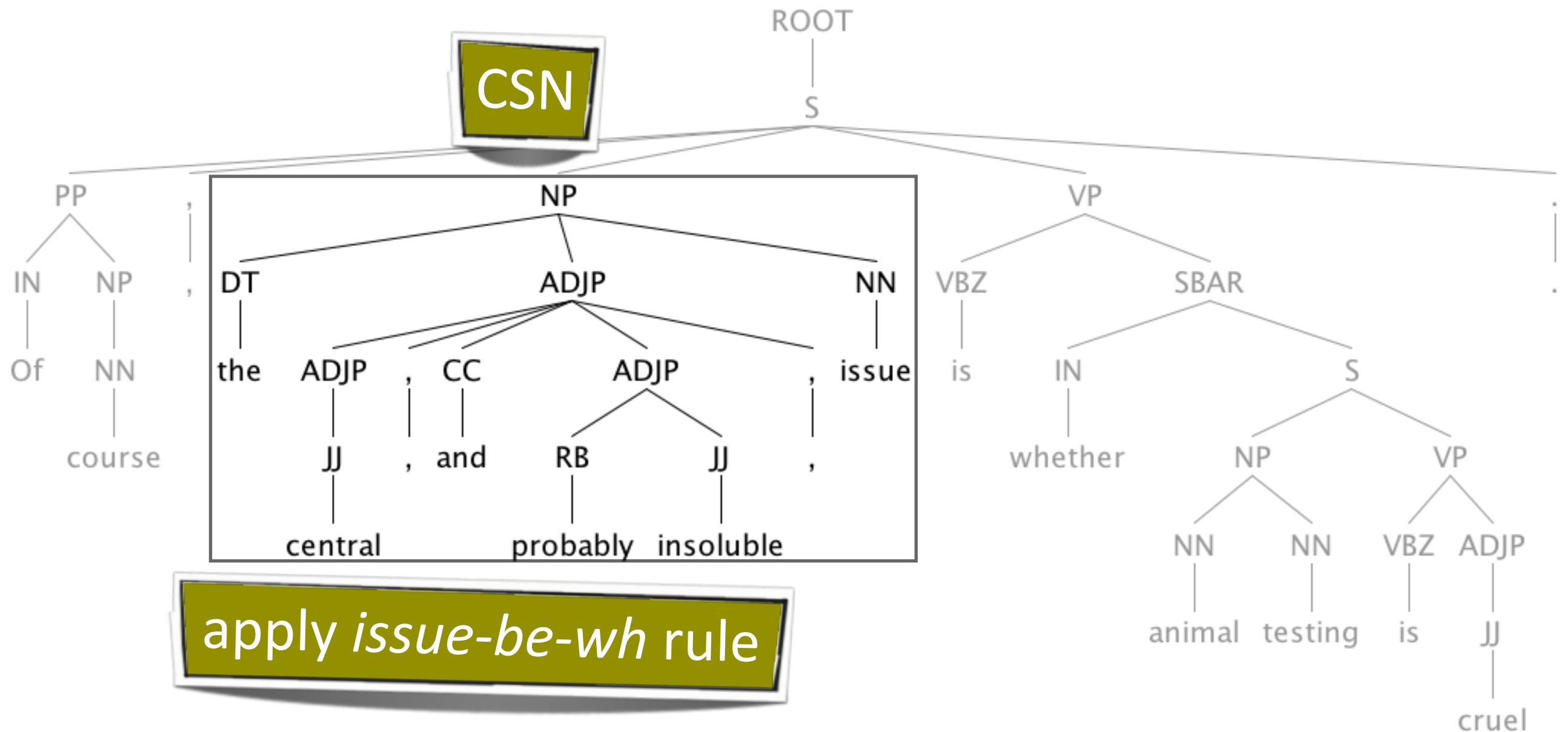


# wh Clause



Of course, the central, and probably insoluble, **issue** is **whether** animal testing is cruel.

# wh Clause

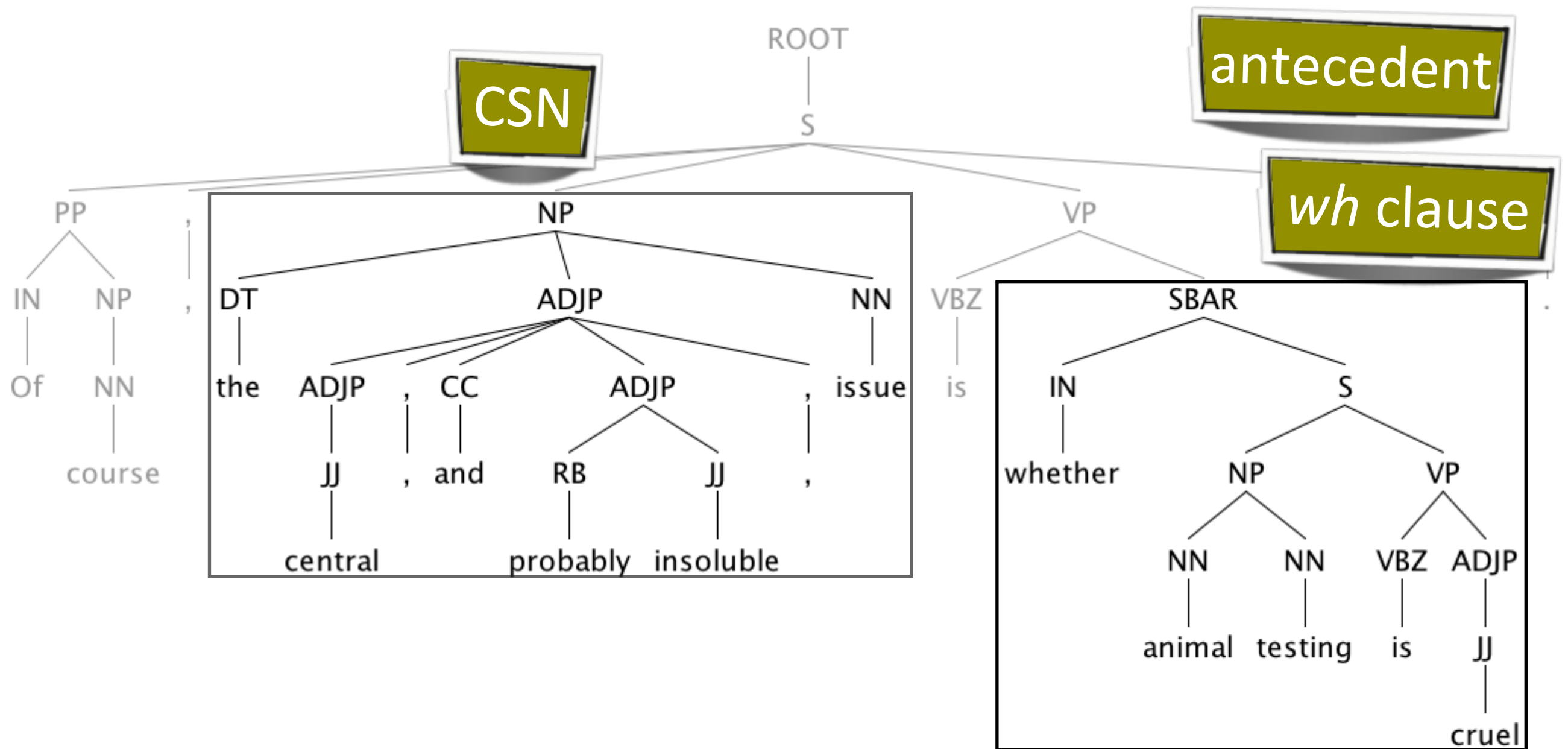


Of course, the central, and probably insoluble, **issue** is **whether** animal testing is cruel.

# Extraction Rules

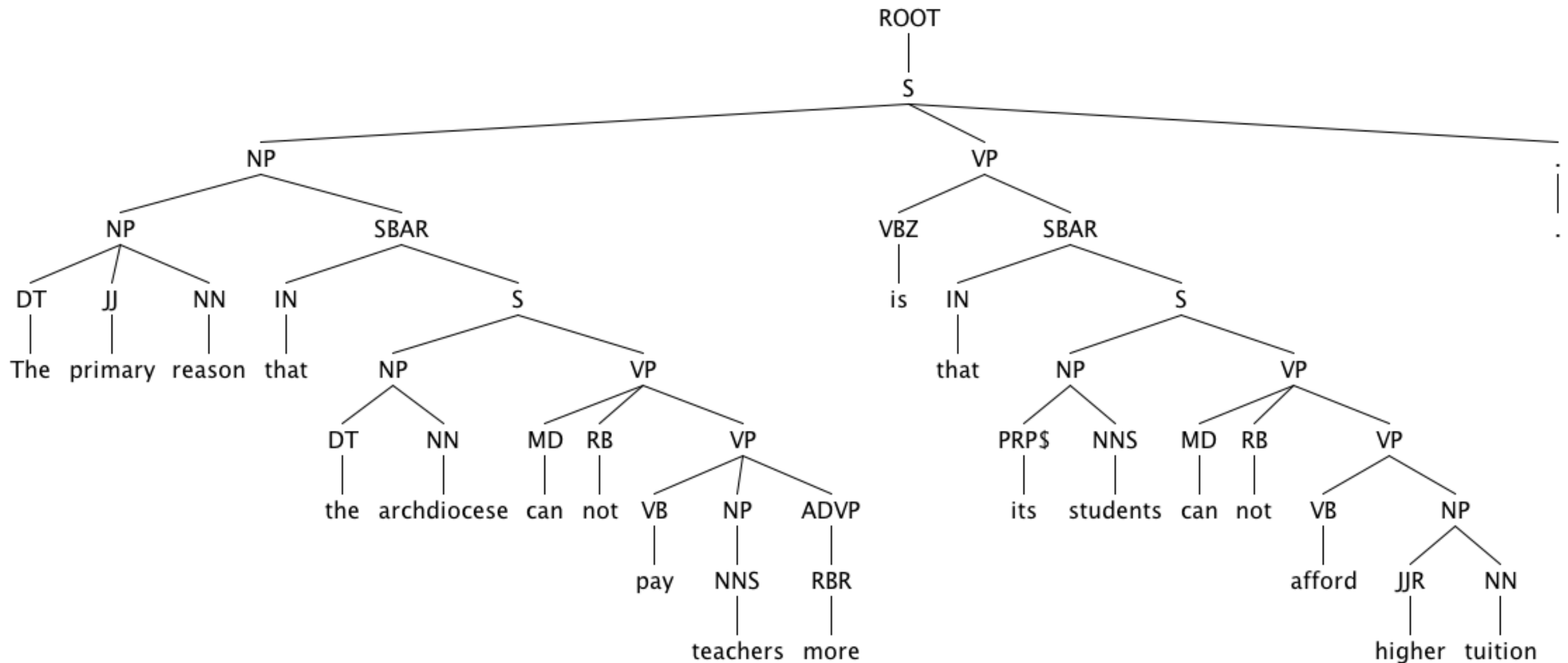
Pattern	<i>fact</i>	<i>reason</i>	<i>issue</i>	<i>decision</i>	<i>question</i>	<i>possibility</i>
<i>N-to</i>	–	–	–	<i>inf</i> clause	predicate	<i>inf</i> clause
<i>N-be-to</i>	–	<i>inf</i> clause	<i>inf</i> clause	<i>inf</i> clause	<i>inf/wh</i> clause	<i>inf</i> clause
<i>N-that</i>	<i>that</i> clause	predicate	predicate	–	predicate	<i>that</i> clause
<i>N-be-that</i>	<i>that</i> clause	<i>that</i> clause	<i>that</i> clause	<i>that</i> clause	<i>that</i> clause	<i>that</i> clause
<i>N-wh</i>	–	predicate	<i>wh</i> clause	<i>wh</i> clause	<i>wh</i> clause	–
<i>N-be-wh</i>	<i>wh</i> clause	<i>wh</i> clause	<b><i>wh</i> clause</b>	<i>wh</i> clause	<i>wh</i> clause	<i>wh</i> clause

# wh Clause



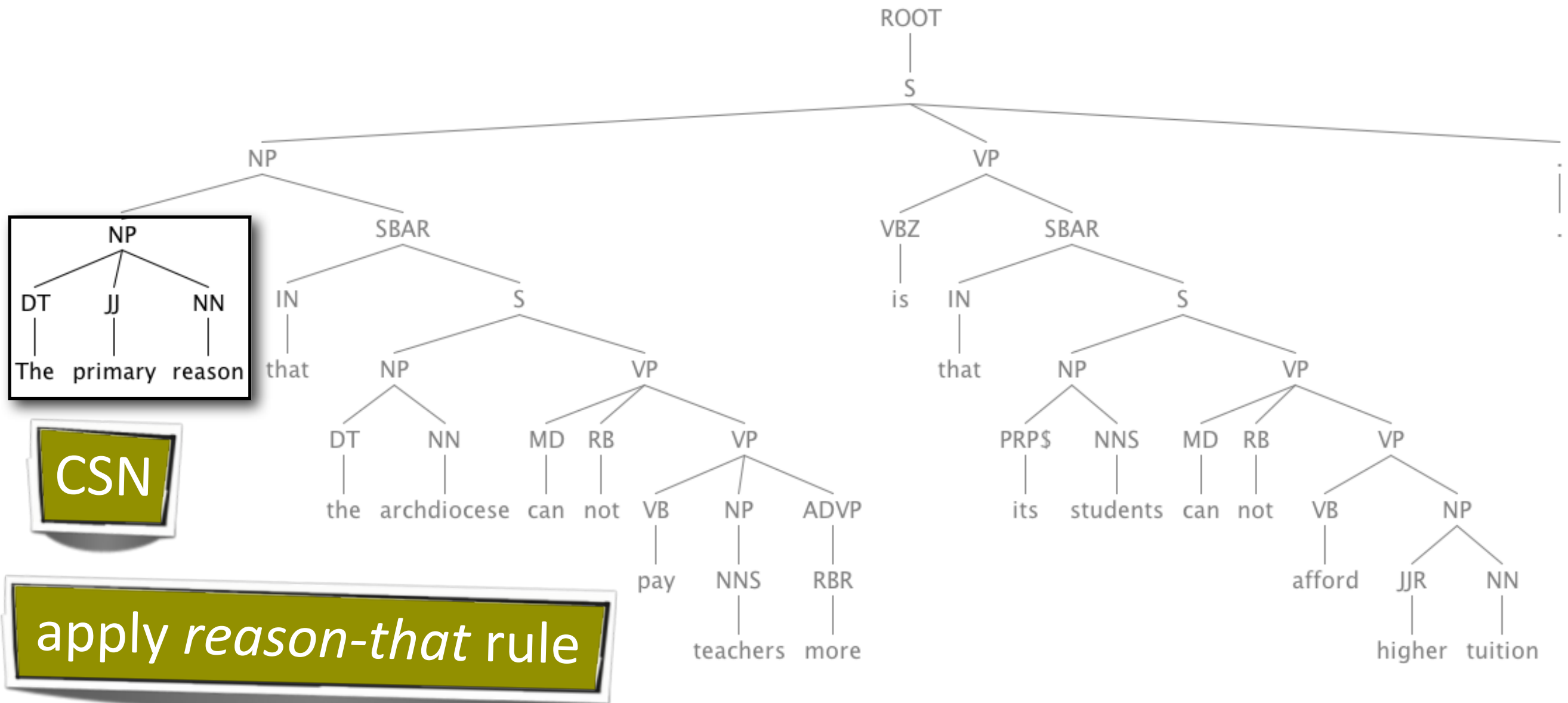
Of course, the central, and probably insoluble, issue is whether animal testing is cruel.

# Predicate Clause



The primary **reason** that the archdiocese cannot pay teachers more is that its students cannot afford higher tuition.

# Predicate Clause

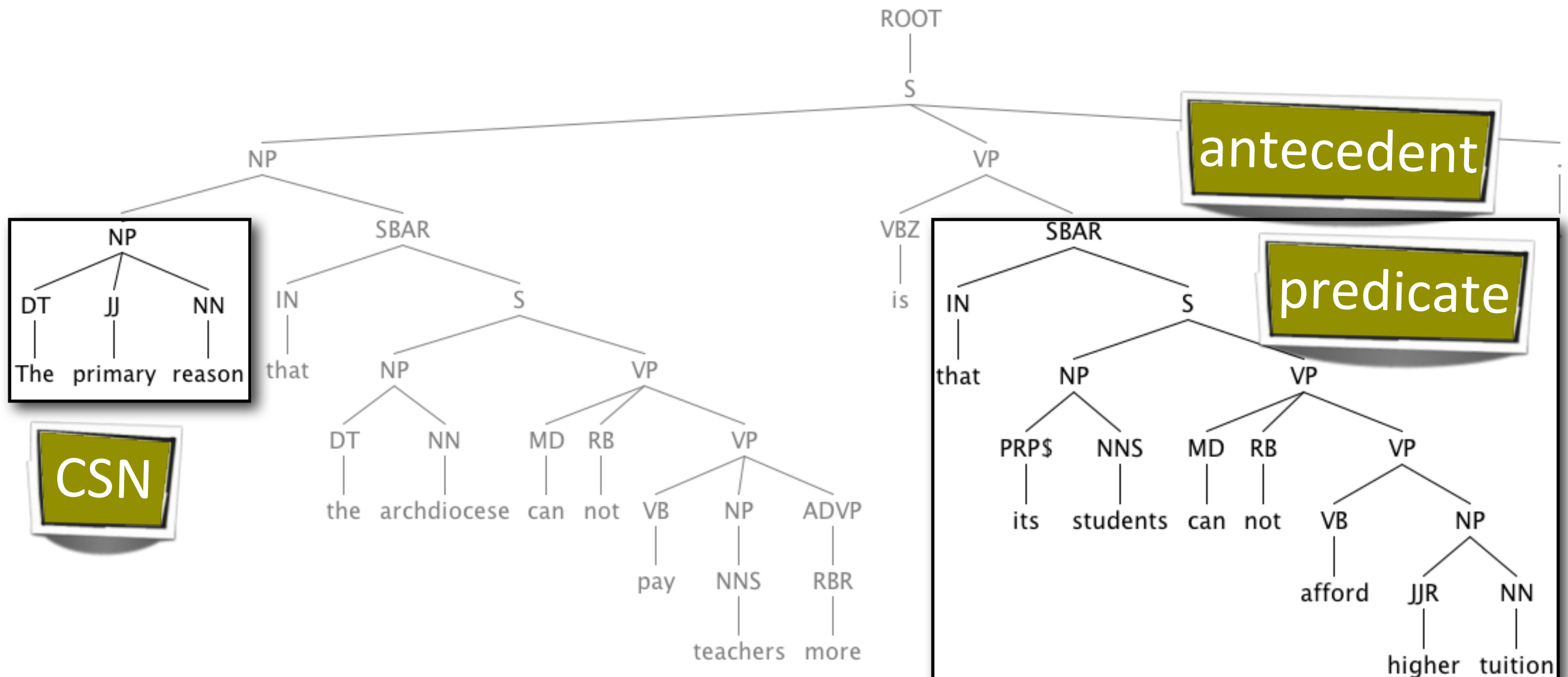


The primary **reason** that the archdiocese cannot pay teachers more is that its students cannot afford higher tuition.

# Extraction Rules

Pattern	<i>fact</i>	<i>reason</i>	<i>issue</i>	<i>decision</i>	<i>question</i>	<i>possibility</i>
<i>N-to</i>	–	–	–	<i>inf</i> clause	predicate	<i>inf</i> clause
<i>N-be-to</i>	–	<i>inf</i> clause	<i>inf</i> clause	<i>inf</i> clause	<i>inf/wh</i> clause	<i>inf</i> clause
<i>N-that</i>	<i>that</i> clause	<b>predicate</b>	predicate	–	predicate	<i>that</i> clause
<i>N-be-that</i>	<i>that</i> clause	<i>that</i> clause	<i>that</i> clause	<i>that</i> clause	<i>that</i> clause	<i>that</i> clause
<i>N-wh</i>	–	predicate	<i>wh</i> clause	<i>wh</i> clause	<i>wh</i> clause	–
<i>N-be-wh</i>	<i>wh</i> clause	<i>wh</i> clause	<i>wh</i> clause	<i>wh</i> clause	<i>wh</i> clause	<i>wh</i> clause

# Predicate Clause



The primary **reason** that the archdiocese cannot pay teachers more is **that its students cannot afford higher tuition.**




# Supervised ML Models

Have:

Automatically labeled CSN antecedent data

- Candidate extraction
- Feature extraction
- Candidate ranking



Extract all syntactic  
constituents

# Features

Feature	Description
syntactic type	coarse-grained and fine-grained syntactic types of the candidate (e.g., VP, S)
<b>embedding level</b>	top and immediate embedding levels (Müller, 2008)
context features	syntactic type and POS tag of left and right siblings
<b>subordinating conjunction</b>	clause preferences for shell nouns (Vendler, 1968) e.g., <i>facts</i> prefer <i>that</i> -clause, <i>issues</i> prefer <i>wh</i> -clause
verb features	presence of verbs or modals, finite/non-finite
length features	length of the candidate in words
<b>lexical features</b>	characteristic words in antecedent part based on information gain (Yang and Pedersen, 1997)

# Candidate Ranking Models

- Gather labeled data for each shell noun
- Train SVM candidate ranking models  
(Joachims, 2002)
- Consider the automatically labeled antecedents as the *true* antecedents

# Testing

# Testing Phase

Have:

Six SVM ranking models for six shell nouns that capture characteristic properties of antecedents for that shell noun

Want:

Prediction for ASN antecedents

# The ASN Corpus

- Base corpus: The NYT corpus  
(Sandhaus, 2008)
- ~475 instances per 6 selected shell nouns  
*fact, reason, issue, decision, question,*  
*possibility*
- Total: 2,323 ASN instances

# Antecedent Prediction

- Candidate extraction
- Feature extraction
- Candidate ranking

# Candidate Extraction

- Large search space
  - On average 49.5 distinct constituents per sentence
  - With  $n$  preceding and  $n$  following sentences  
 $49.5 \times (n+2)$  candidates
- Identify sentence containing the antecedent  
(Kolhatkar et al., 2013)
- Extract all syntactic constituents given by the Stanford parser



# Feature Extraction and Candidate Ranking

- Similar to the training phase
- Invoke the appropriate trained model to rank the antecedent candidates of the given ASN instance



# Evaluation

# CrowdFlower Task

The municipal council will have to decide whether to balance the budget by raising revenue or cutting spending. The council will have to come to a resolution by the end of the month. This issue is dividing communities across the country.

# CrowdFlower Task

The municipal council will have to decide whether to balance the budget by raising revenue or cutting spending.

**Select one of the options**

- ☐ None
- ☐ whether to balance the budget
- ☐ have to decide whether to balance the budget by raising revenue or cutting spending
- ☐ decide whether to balance the budget by raising revenue or cutting spending
- ☐ whether to balance the budget by raising revenue or cutting spending
- ☐ to decide whether to balance the budget by raising revenue or cutting spending
- ☐ to balance the budget by raising revenue or cutting spending
- ☐ The municipal council will have to decide whether to
- ☐ balance the budget by raising revenue or cutting spending
- ☐ the budget by raising revenue or cutting spending
- ☐ will have to decide whether to balance the budget by raising revenue or cutting spending

10 top-ranked candidates  
(randomly ordered)

**i** Select one of the above options that provides meaning to the underlined phrase in blue.

**Are you satisfied with the above options?**

- ☐ Satisfied
- ☐ Partially satisfied
- ☐ Unsatisfied

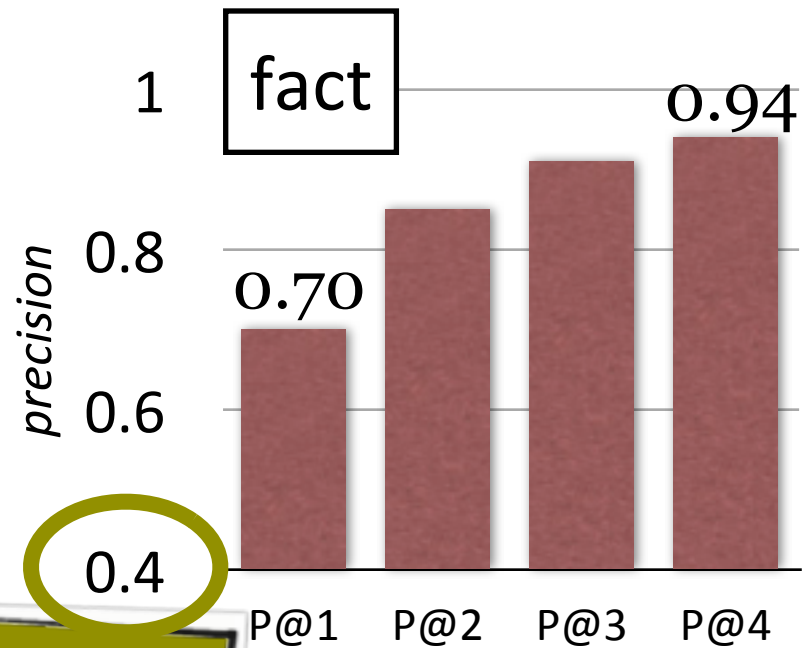
**i** Your satisfaction level of the above options depending upon whether your answer was present in the given options.

# Results

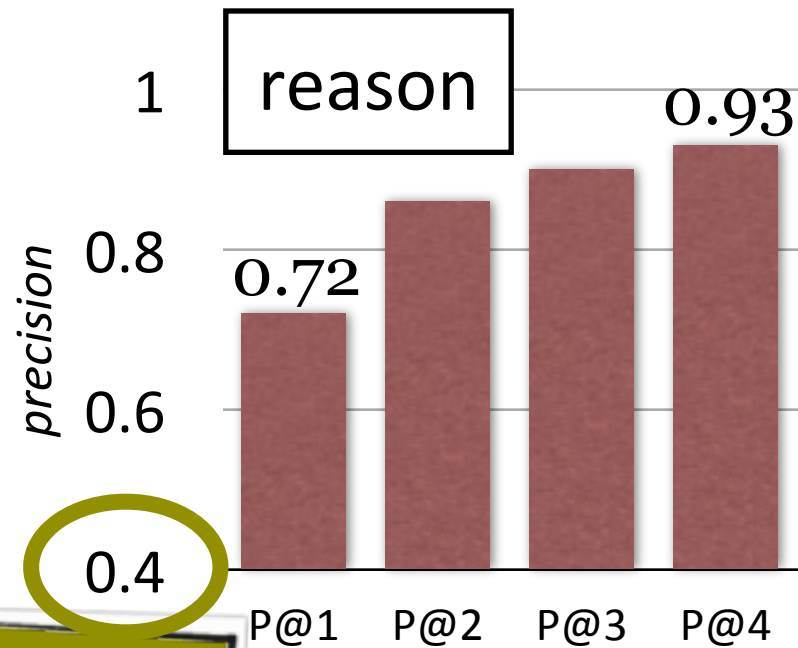
# Metric and Baseline

- Precision at  $n$  ( $P@n$ )
  - Proportion of instances where the crowd's answers occur within our ranker's first  $n$  choices
  - $P@1$  is standard precision
- Baseline
  - Consider the previous sentence as the correct antecedent

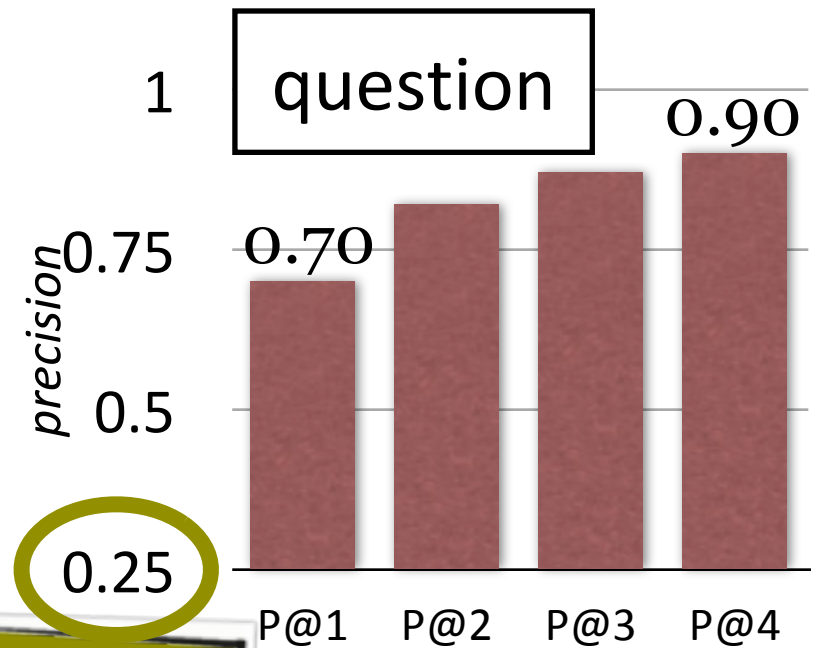
# Ranker Evaluation



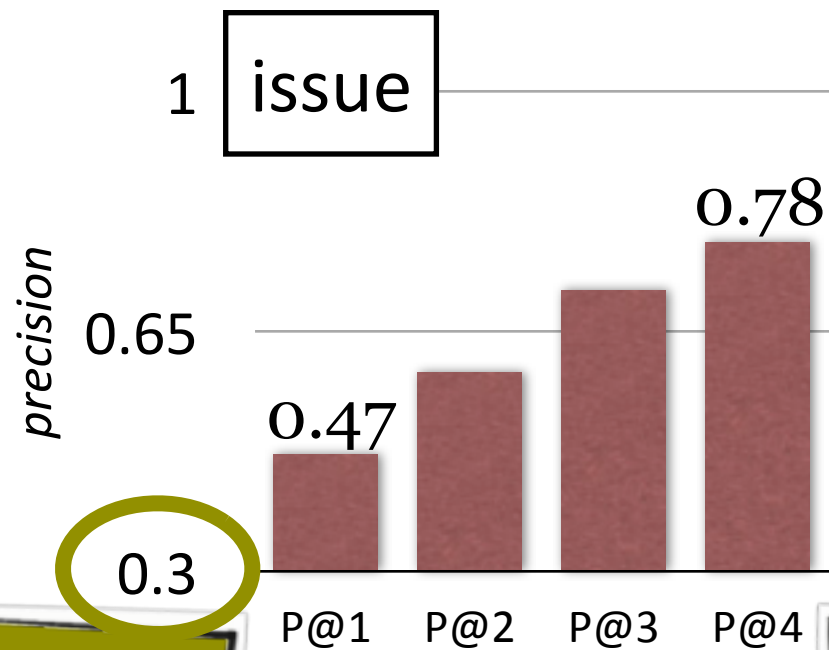
baseline



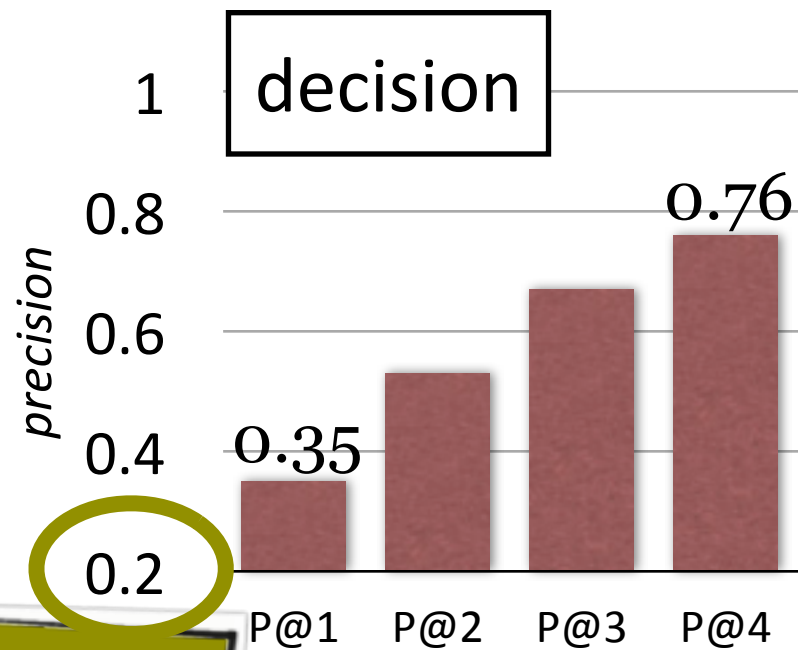
baseline



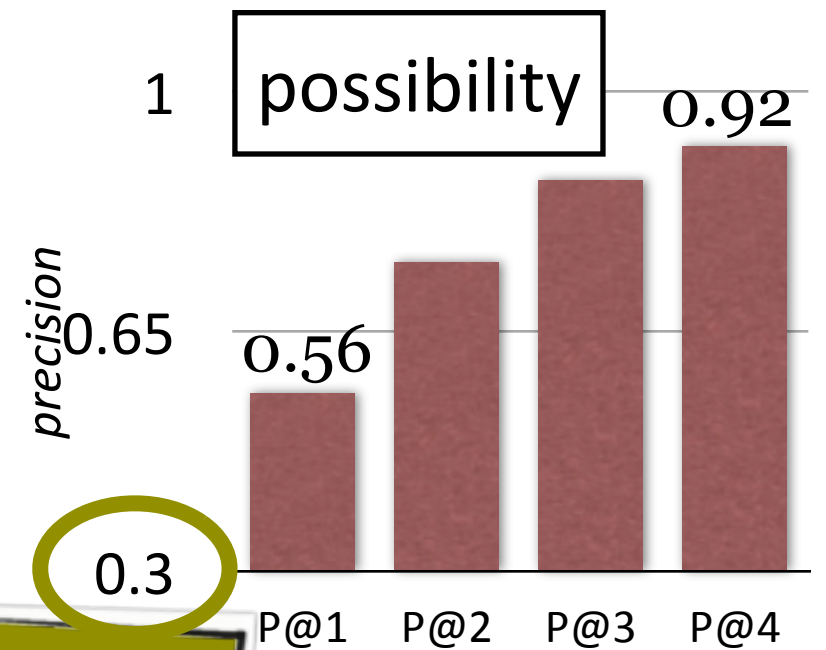
baseline



baseline

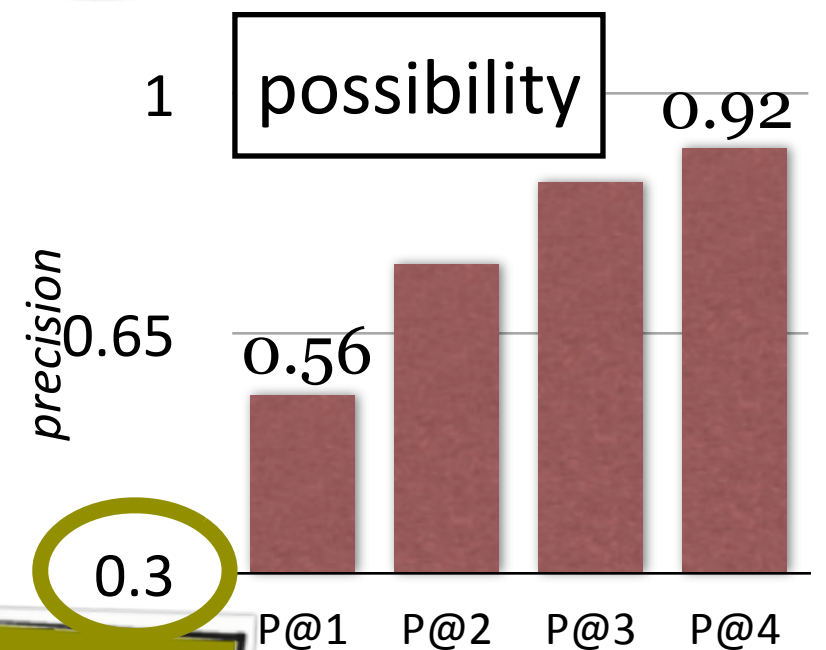
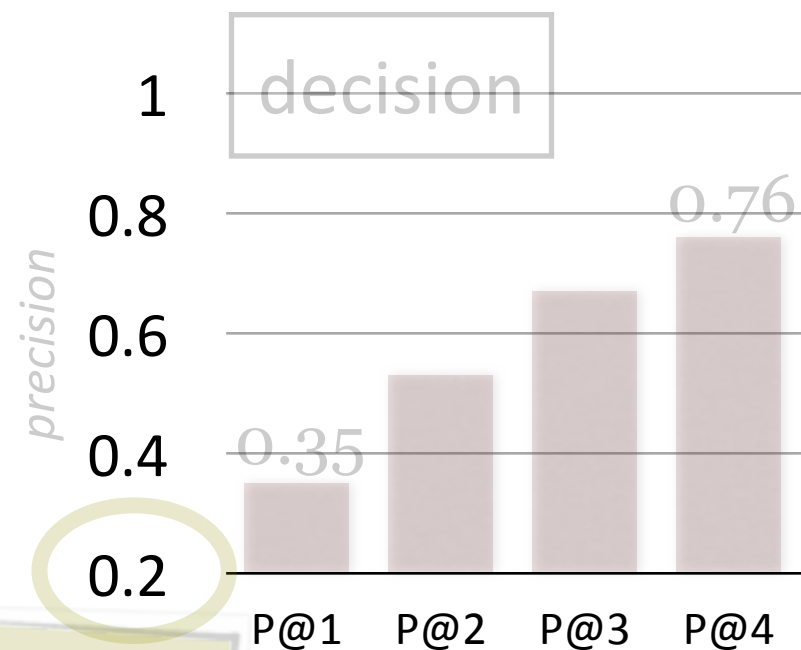
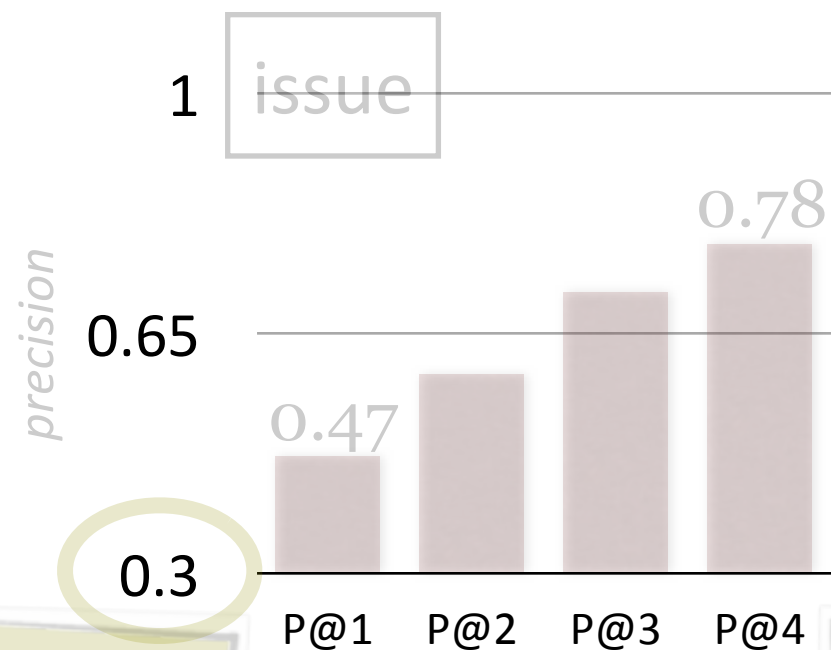
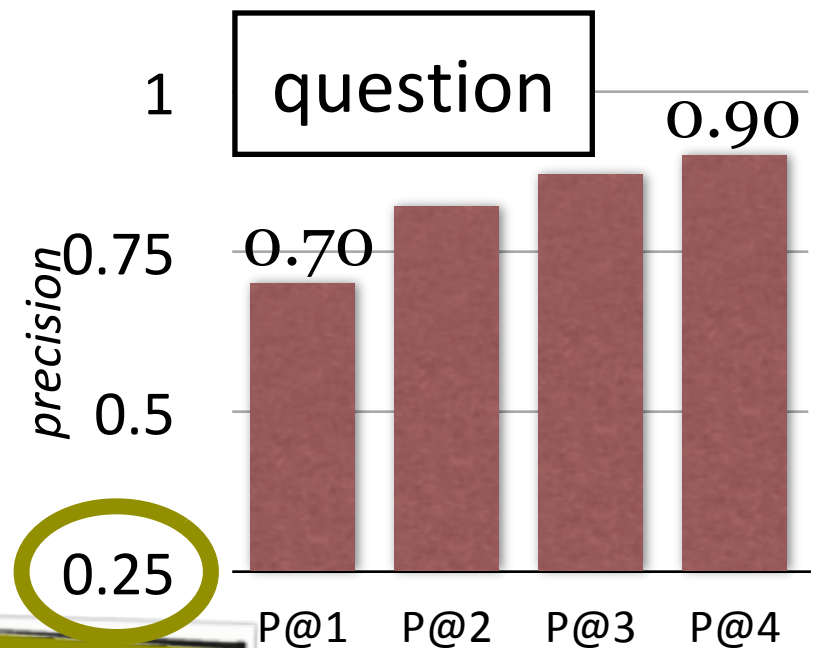
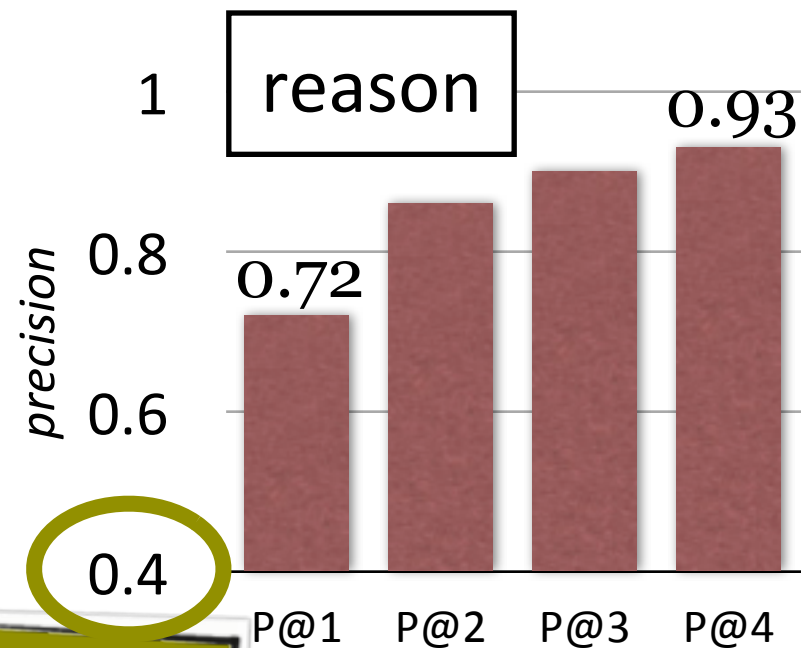
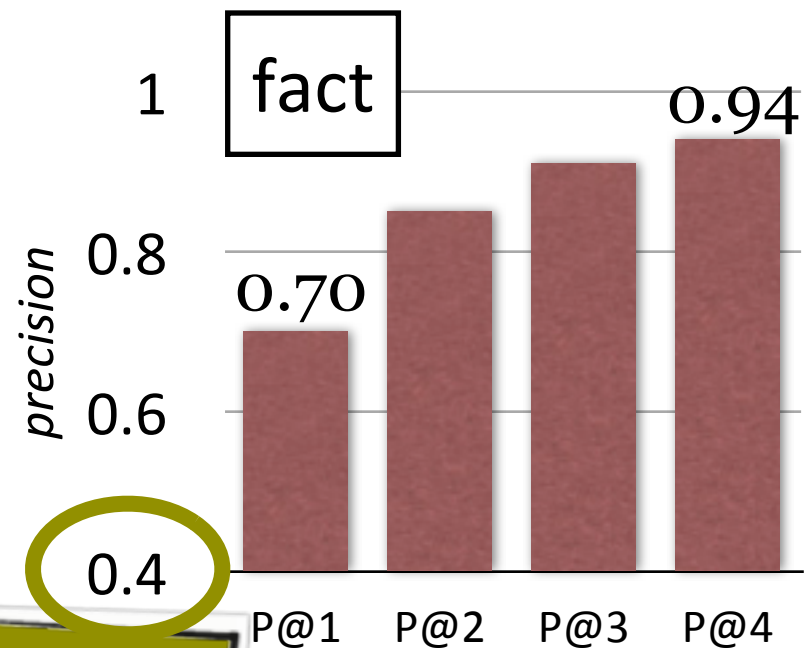


baseline



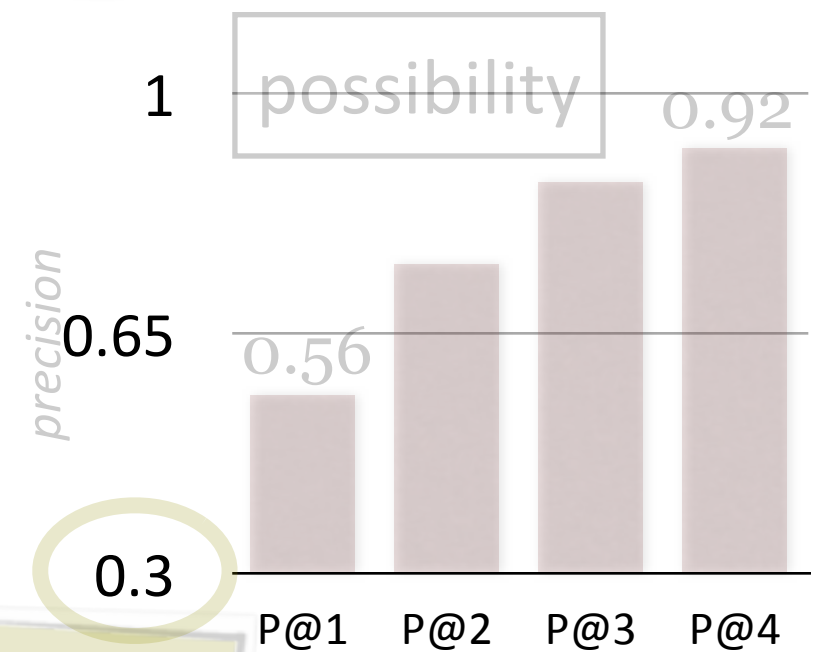
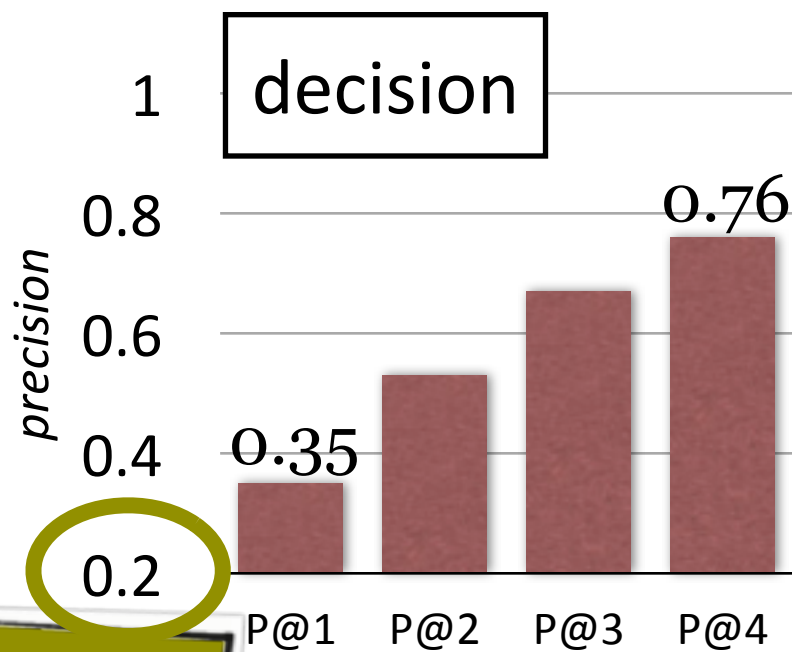
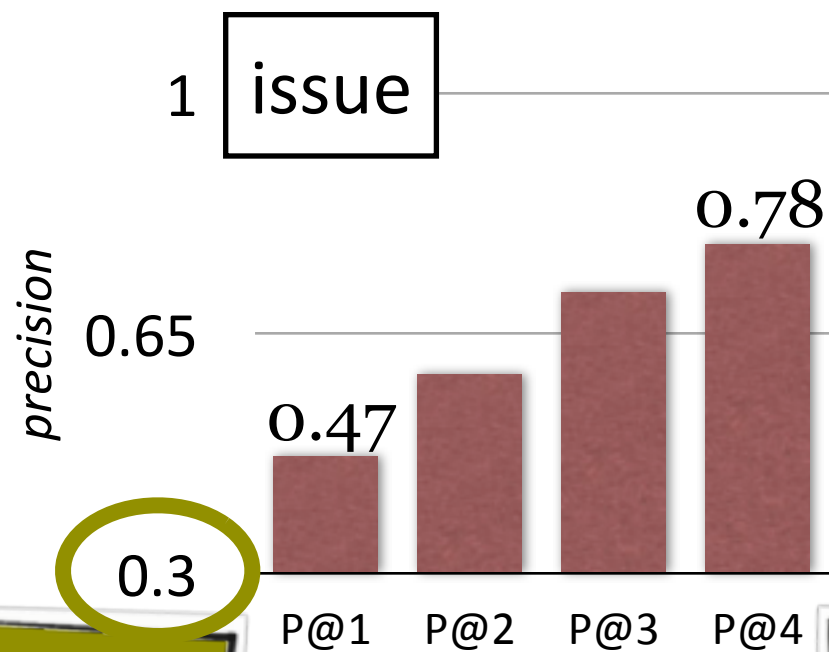
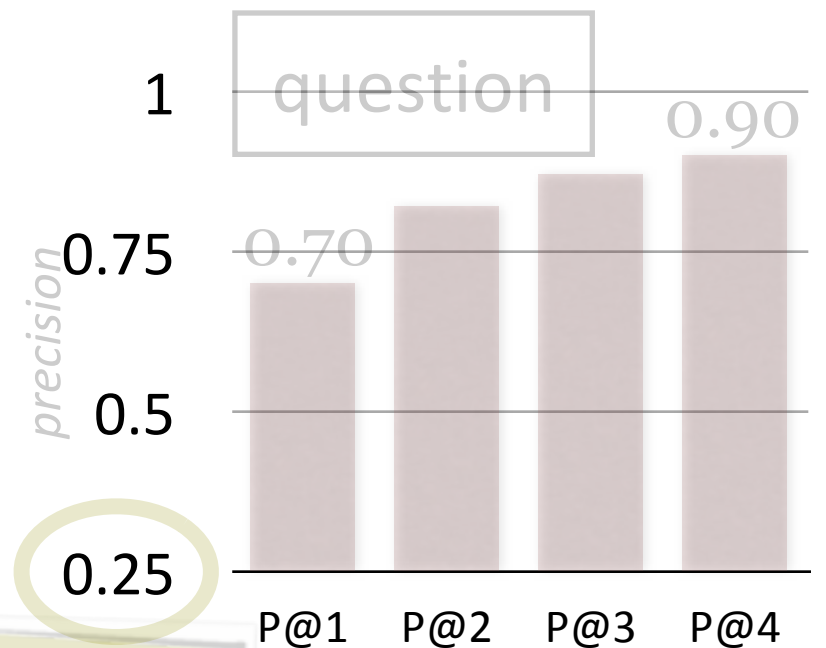
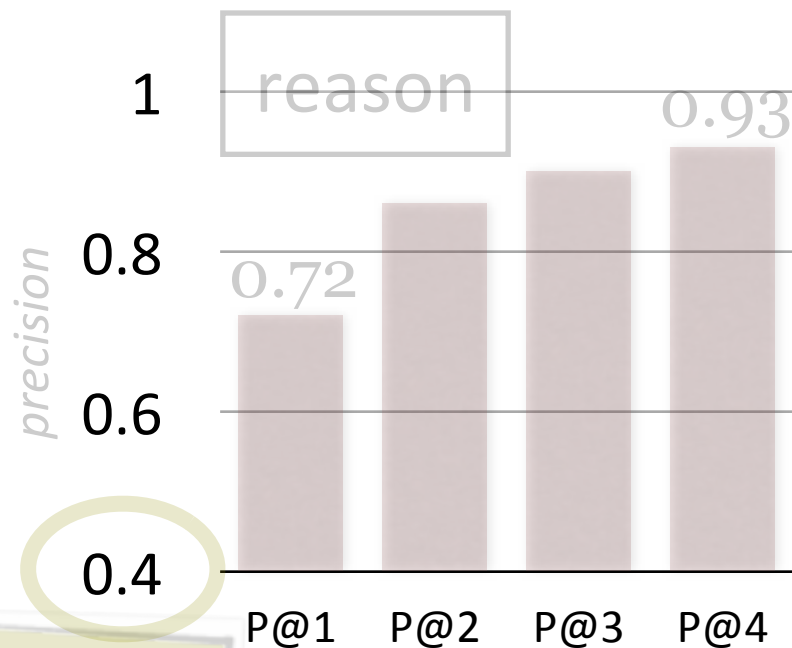
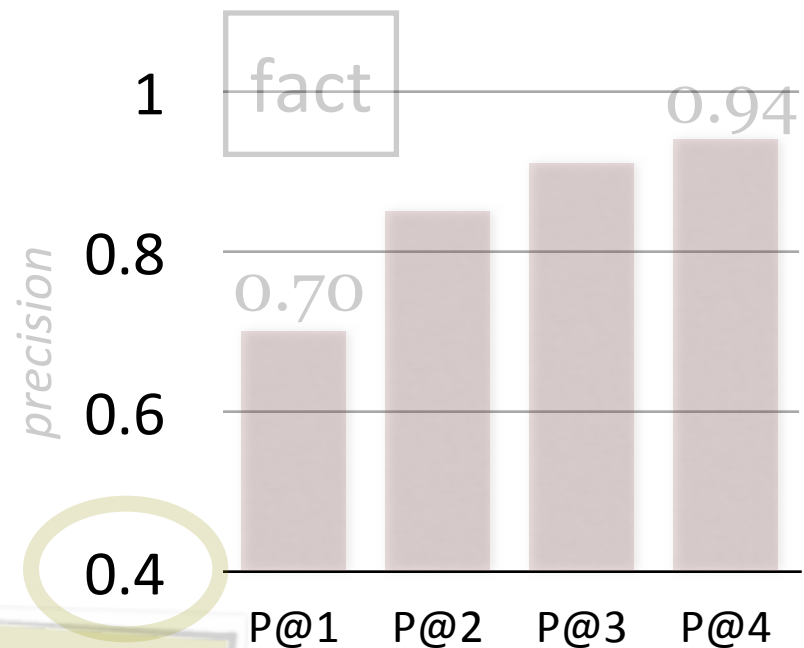
baseline

# Ranker Evaluation





# Ranker Evaluation



# Feature Analysis

- Best performing features: embedding level, lexical, subordinating conjunction
- Syntactic type: not very useful
  - ASNs had a wide variety of syntactic type than what was available in CSN data
- No specific features associated with Schmid's semantic categories

# Conclusion

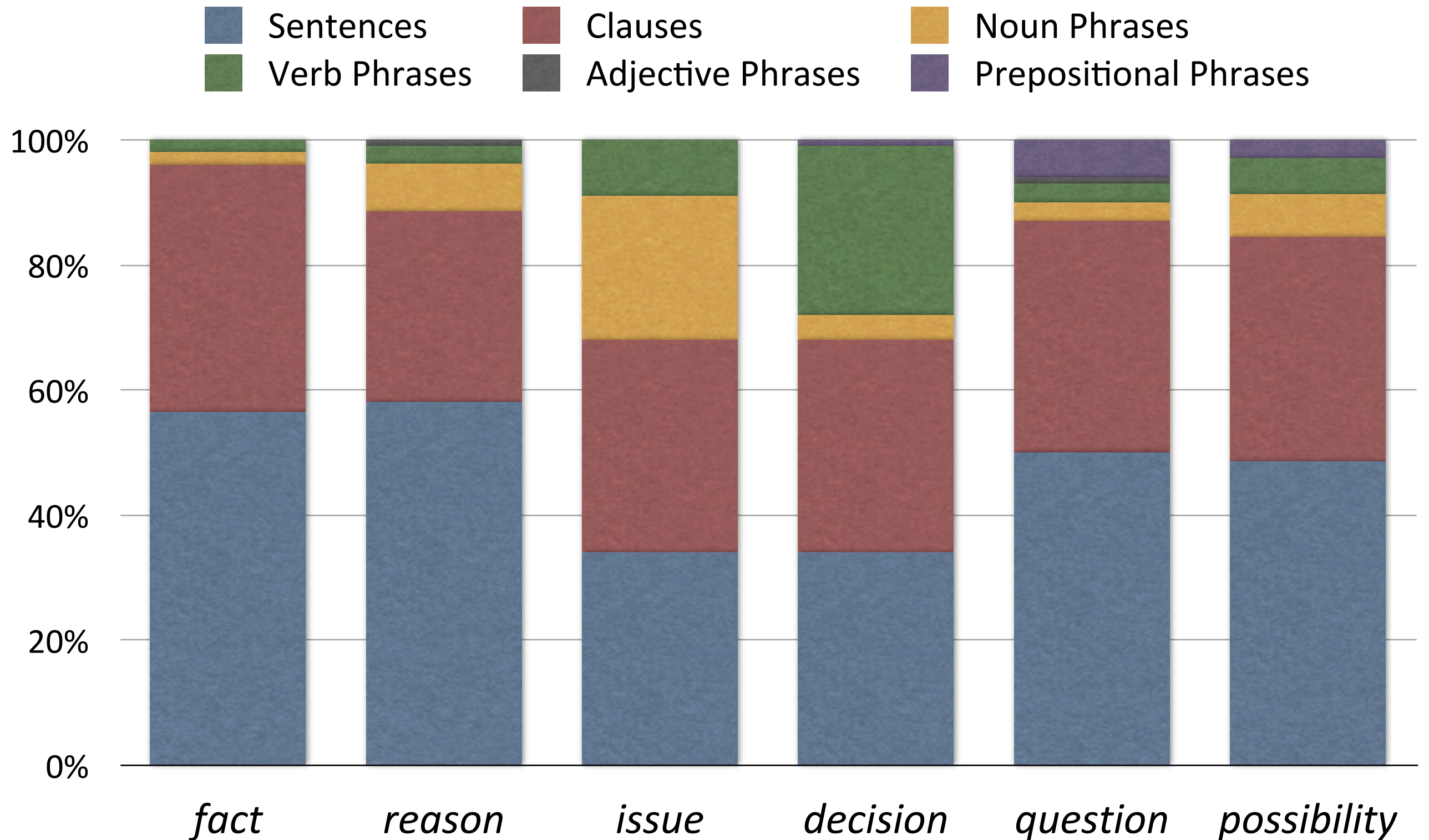
Revisit the hypothesis: CSN antecedents and ASN antecedents share some linguistic properties, and hence linguistic knowledge encoded in CSN antecedents will help in interpreting ASNs.

- Precision results as high as 0.72 for *reason* and 0.70 for *fact* support the hypothesis
- The mental nouns *issue* and *decision* were harder to interpret than other shell nouns

Our models can be used as base models to reduce the large search space of ASN antecedent candidates.



# Syntactic Type Distribution



# Hard Examples

The teacher erased the solutions before John had time to copy them out, as he had momentarily been distracted by a band playing outside.

- This fact infuriated him, as the teacher always erased the board quickly and John suspected it was just to punish anyone who was lost in thought, even for a moment.
- This fact infuriated the teacher, who had already told John several times to focus on class work.

# Hard Examples

Several Vatican officials said, however, that any such talk has little meaning because the church does not take sides in elections. But the statements by several American bishops that Catholics who vote for Mr. Kerry would have to go to confession have raised the question in many corners about whether this is an official church position.

The church has not addressed **this question** publicly and, in fact, seems reluctant to be dragged into the fight...”

# Hard Examples

Any biography of Thomas More has to answer one fundamental question. Why? Why, out of all the many ambitious politicians of early Tudor England, did only one refuse to acquiesce to a simple piece of religious and political opportunism? What was it about More that set him apart and doomed him to a spectacularly avoidable execution?

The innovation of Peter Ackroyd's new biography of More is that he places the answer to **this question** outside of More himself.



