

# **Syntactic Constraints on Paraphrases Extracted from Parallel Corpora**

**Chris Callison-Burch  
Johns Hopkins University  
October 25, 2008**

# Overview

- Extracting paraphrases from parallel corpora
- Syntactic constraints
- Results of manual evaluation

# Extracting paraphrases from Bilingual parallel corpora

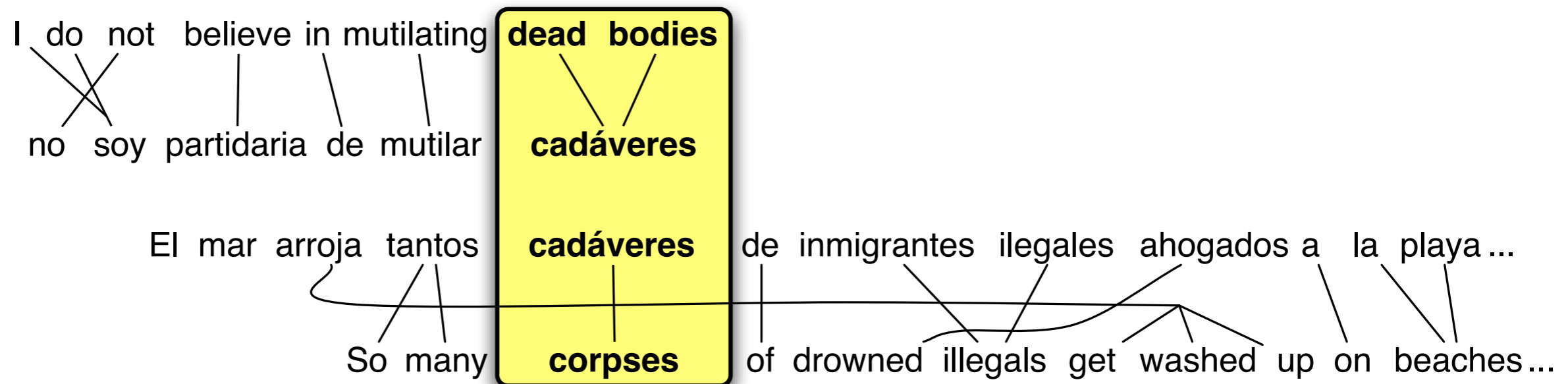
I do not believe in mutilating **dead bodies**  
no soy partidaria de mutilar **cadáveres**

# Extracting paraphrases from Bilingual parallel corpora

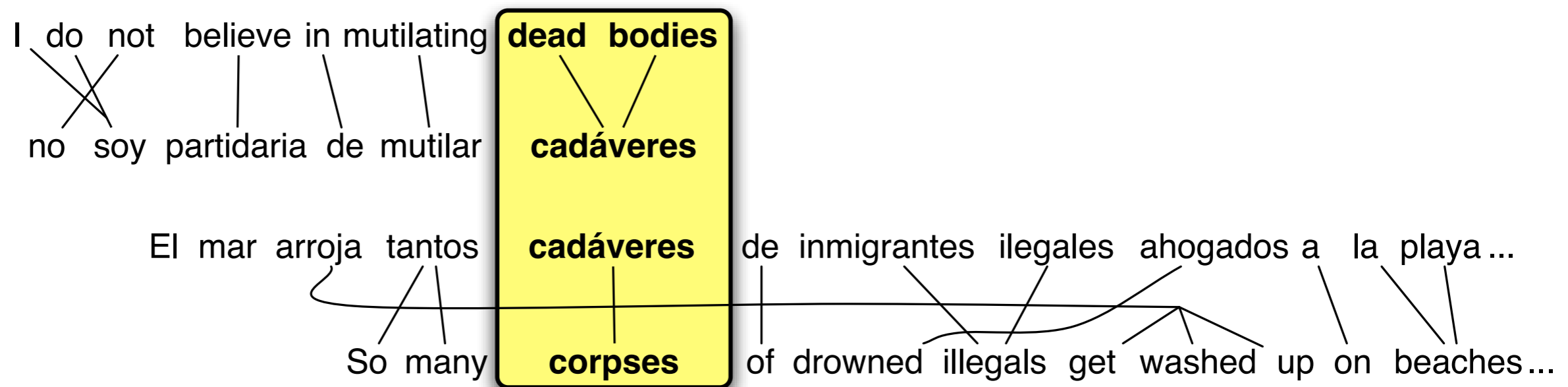
I do not believe in mutilating **dead bodies**  
no soy partidaria de mutilar **cadáveres**

El mar arroja tantos **cadáveres** de inmigrantes ilegales ahogados a la playa ...  
So many **corpses** of drowned illegals get washed up on beaches...

# Extracting paraphrases from Bilingual parallel corpora



# Extracting paraphrases from Bilingual parallel corpora



**dead bodies** → corpses, carcasses, bodies, skeletons, people

# Good examples

- **military force** → force, forces, peace-keeping personnel, armed forces, military forces, defense
- **sooner or later** → eventually, at some point
- **wish to clarify** → want to make perfectly clear, would like to ask, would like to comment on, would like to mention, would like to deal with, would comment on
- **every other** → any other, all, other, every, all other, everyone else, others, all the others

# Bad examples

- **are perfectly entitled** → perfectly entitled, have every right, right, are, has a legitimate, call for, has, legitimate right, have the right
- **for small-scale projects** → small-scale projects, small, of, only trifling amounts are at stake, for projects, for smaller-scale projects, to, for smaller projects
- **groundwork for** → for, groundwork, to, basis for, the, basis, preparation, foundations for, that
- **create equal** → equal, to create a, create, to create equality, same, created, conditions

# Paraphrase probability

$$\hat{e}_2 = \arg \max_{e_2: e_2 \neq e_1} p(e_2 | e_1)$$

# Paraphrase probability

$$\hat{e}_2 = \arg \max_{e_2: e_2 \neq e_1} p(e_2|e_1)$$

$$\begin{aligned} p(e_2|e_1) &= \sum_f p(f|e_1)p(e_2|f, e_1) \\ &\approx \sum_f p(f|e_1)p(e_2|f) \end{aligned}$$

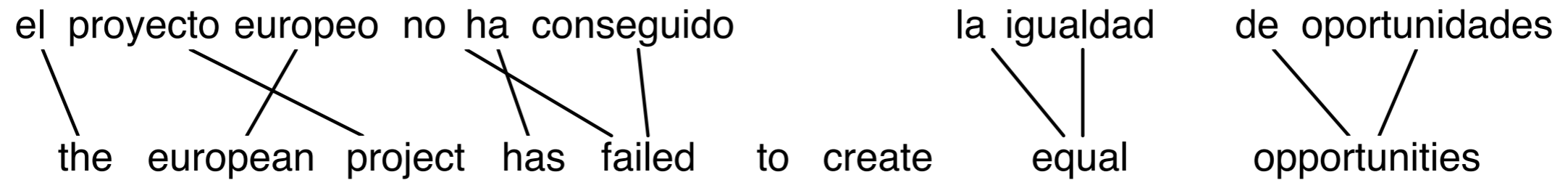
# Paraphrase probability

$$\hat{e}_2 = \arg \max_{e_2: e_2 \neq e_1} p(e_2|e_1)$$

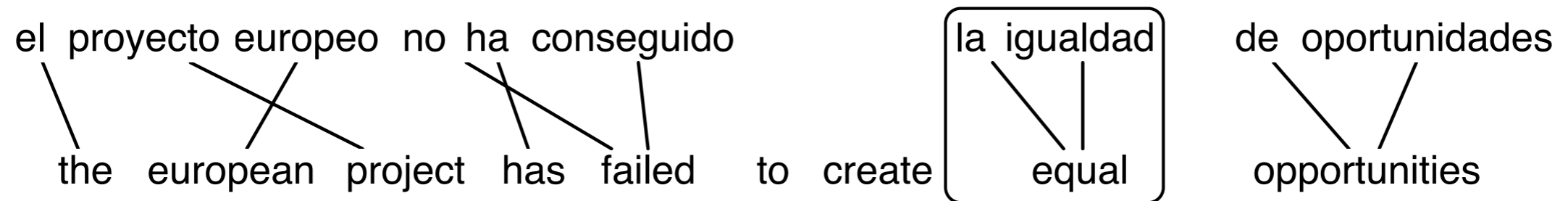
$$\begin{aligned} p(e_2|e_1) &= \sum_f p(f|e_1)p(e_2|f, e_1) \\ &\approx \sum_f p(f|e_1)p(e_2|f) \end{aligned}$$

$$p(f|e) = \frac{\text{count}(e, f)}{\sum_f \text{count}(e, f)}$$

# Phrase extraction with Unaligned words

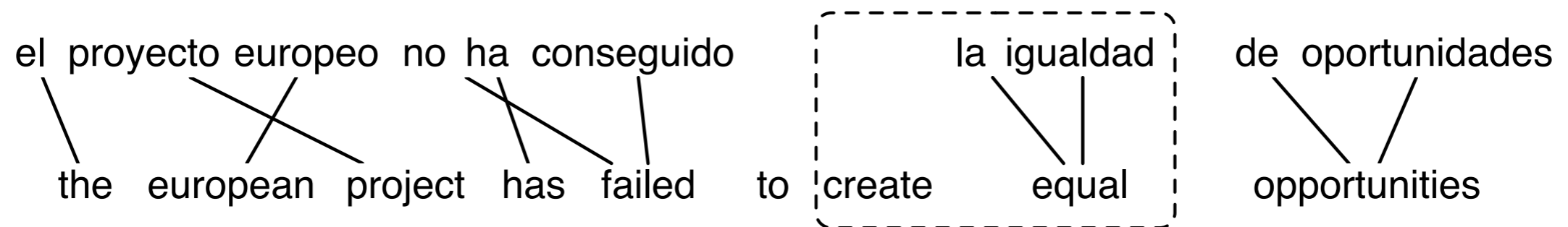


# Phrase extraction with Unaligned words



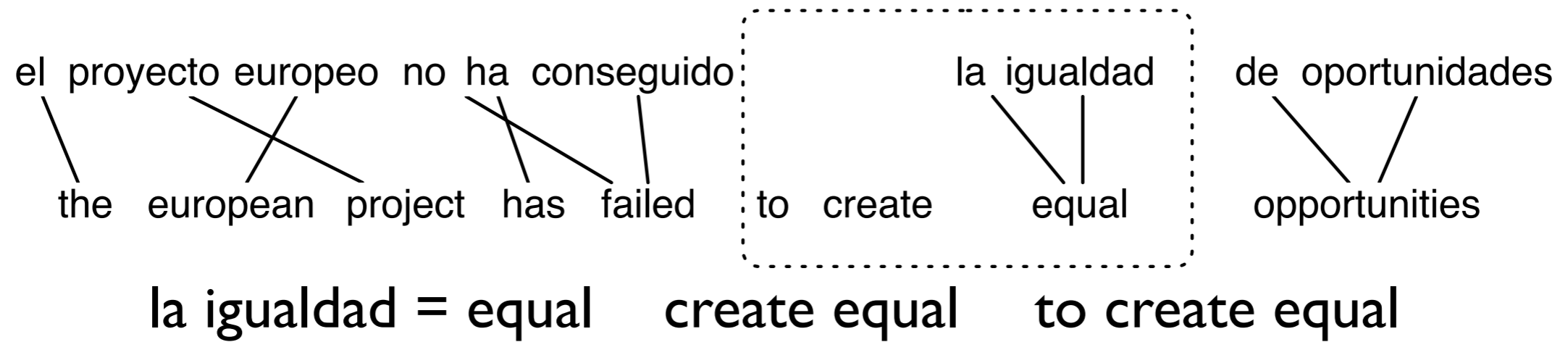
la igualdad = equal

# Phrase extraction with Unaligned words

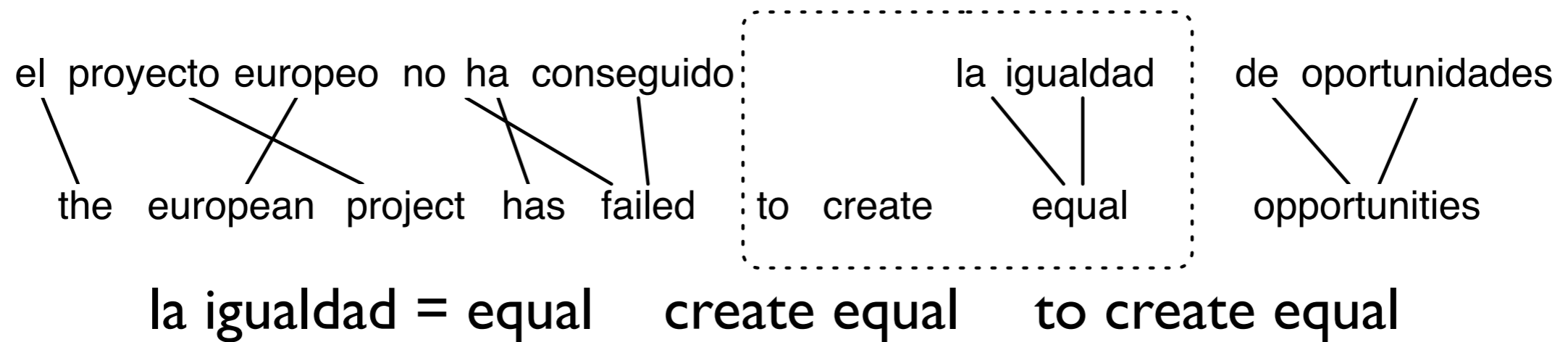


la igualdad = equal      create equal

# Phrase extraction with Unaligned words



# Phrase extraction with Unaligned words



- For 3.7m paraphrases of 400k phrases
  - 34% were sub- or super-strings
  - 73% of the paraphrases that were ranked highest by the paraphrase probability

# Potential solutions

- Use multiple parallel corpora to eliminate systematic misalignments in one language
- Re-rank results with a language model when paraphrases are substituted into a sentence
- Impose requirement that paraphrases cannot be substrings and superstrings

# Syntactic Constraints

- Require phrases and their paraphrase to be the same syntactic type
- Redefine the paraphrase probability to condition on syntactic labels
- Change the phrase extraction algorithm so that it enumerates phrase pairs and syntactic labels

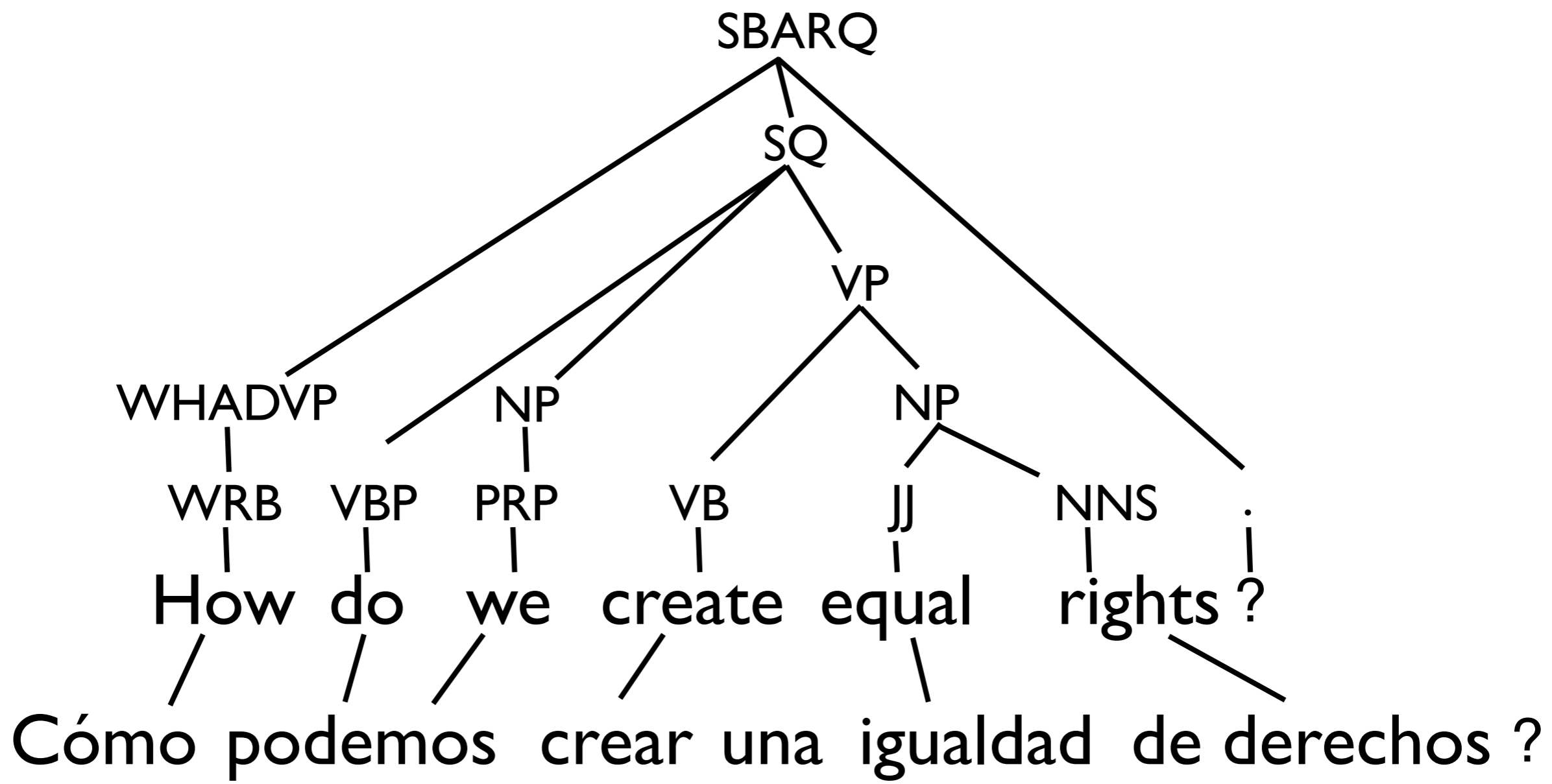
# Redefined paraphrase prob

$$\begin{aligned}\hat{e}_2 &= \arg \max_{e_2: e_2 \neq e_1 \wedge s(e_2) = s(e_1)} p(e_2 | e_1, s(e_1)) \\ &\approx \arg \max_{e_2: e_2 \neq e_1 \wedge s(e_2) = s(e_1)} \sum_f p(f | e_1, s(e_1)) p(e_2 | f, s(e_1))\end{aligned}$$

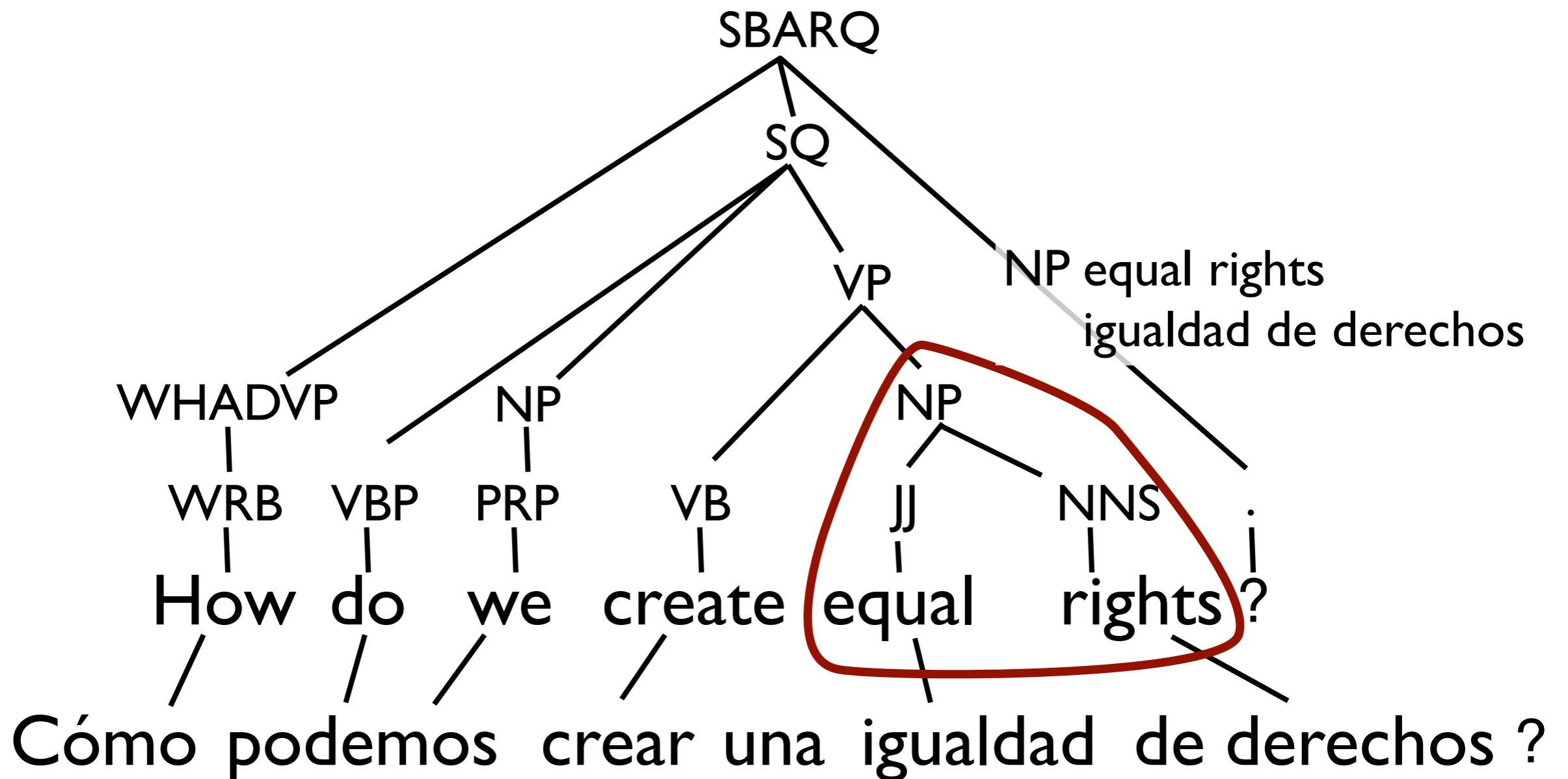
$$p(f | e_1, s(e_1)) = \frac{\text{count}(f, e_1, s(e_1))}{\sum_f \text{count}(f, e_1, s(e_1))}$$

$$p(e_2 | f, s(e_1)) = \frac{\text{count}(f, e_2, s(e_1))}{\sum_{e_2} \text{count}(f, e_2, s(e_1))}$$

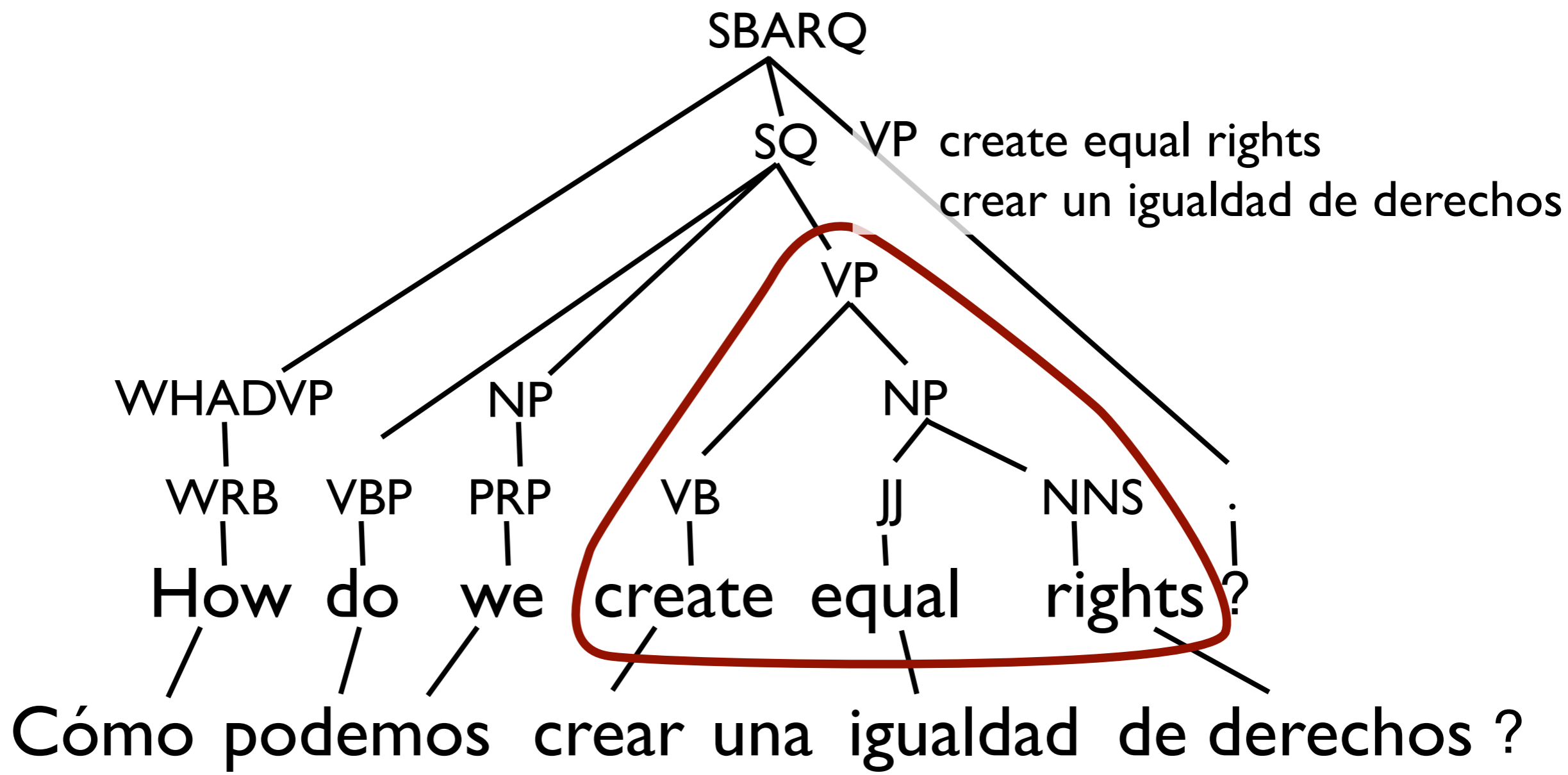
# Phrase Extraction + Syntactic Labels



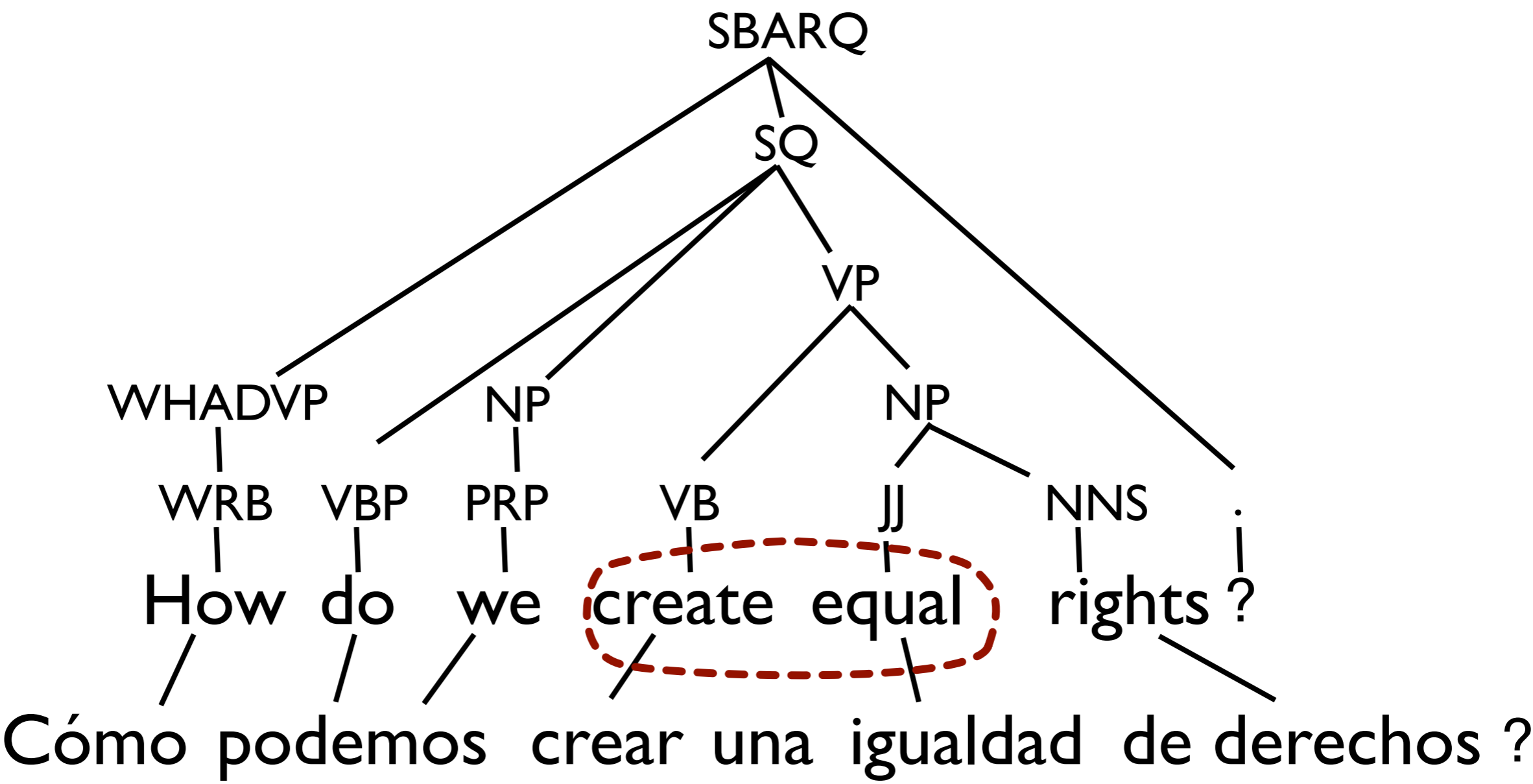
# Phrase Extraction + Syntactic Labels



# Phrase Extraction + Syntactic Labels

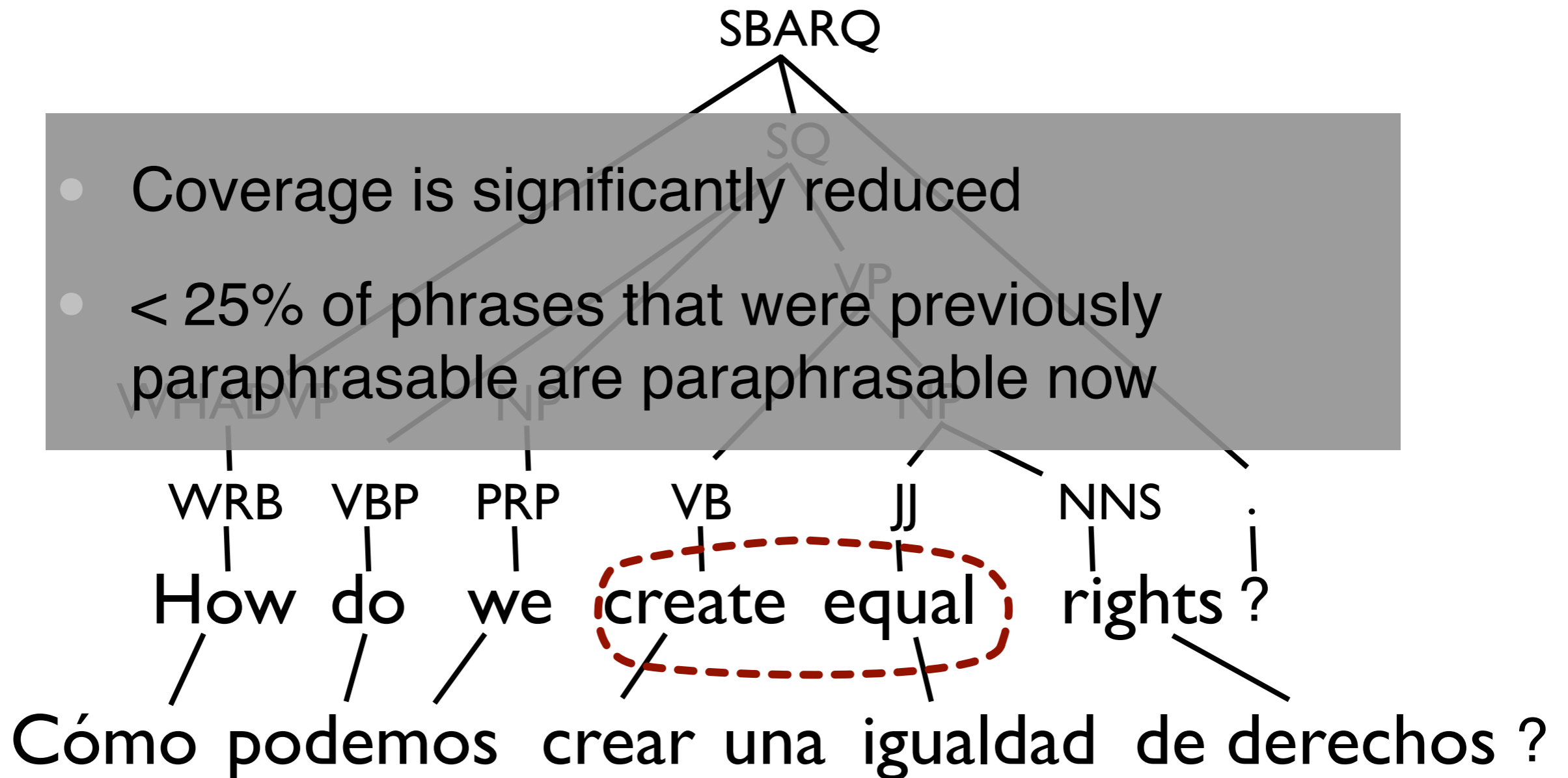


# Phrase Extraction + Syntactic Labels

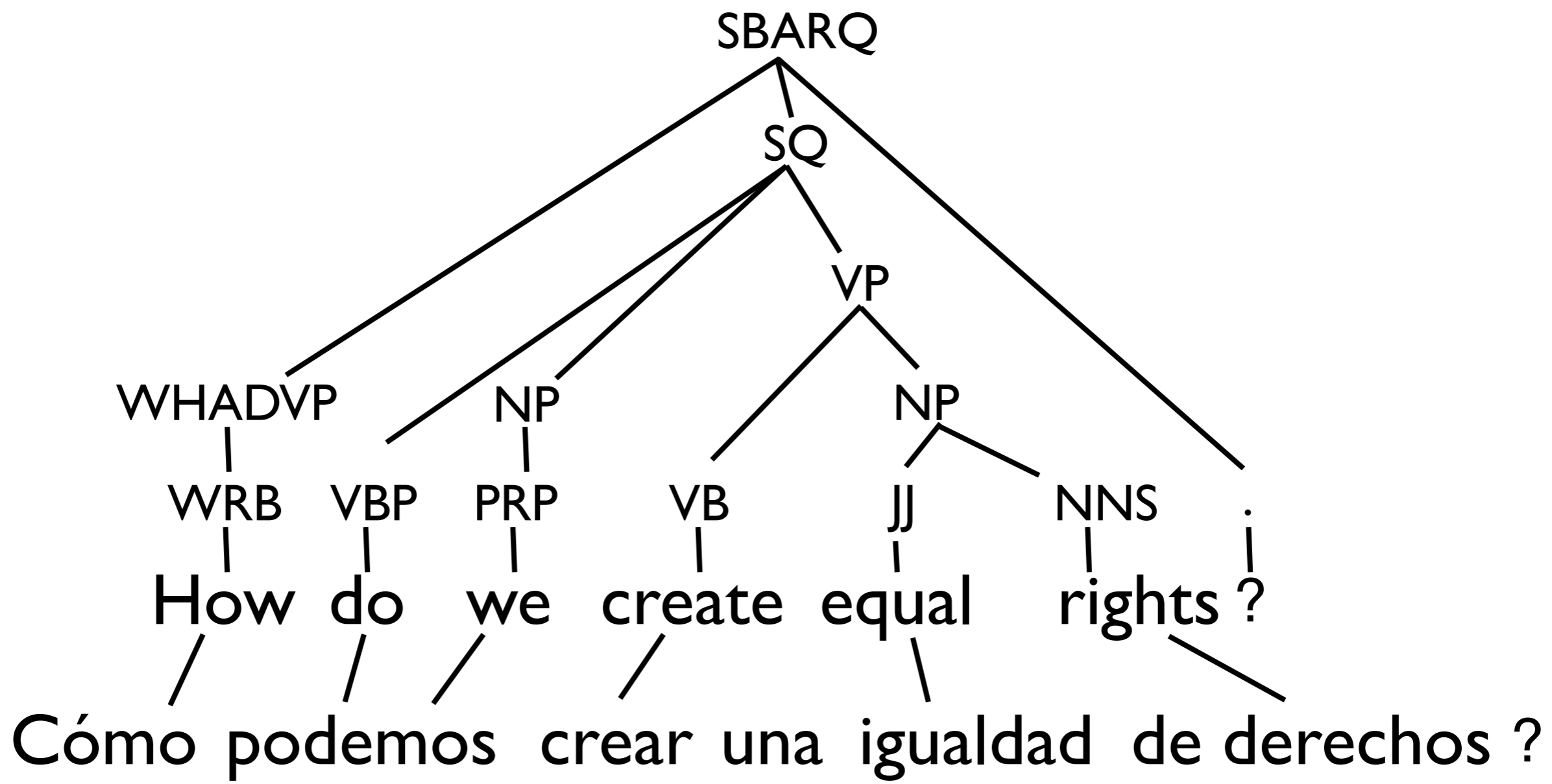


# Phrase Extraction + Syntactic Labels

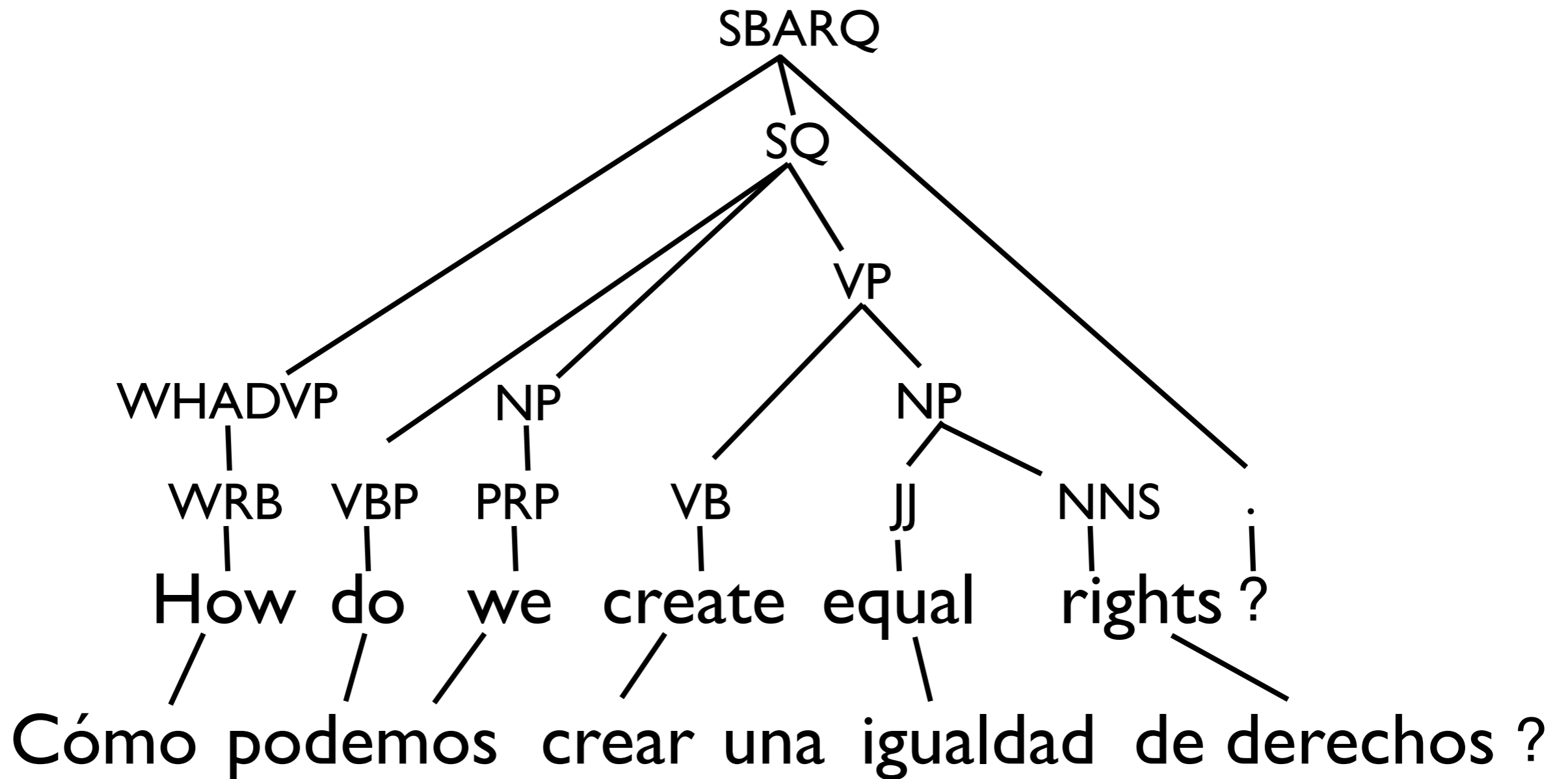
- Coverage is significantly reduced
- < 25% of phrases that were previously paraphrasable are paraphrasable now



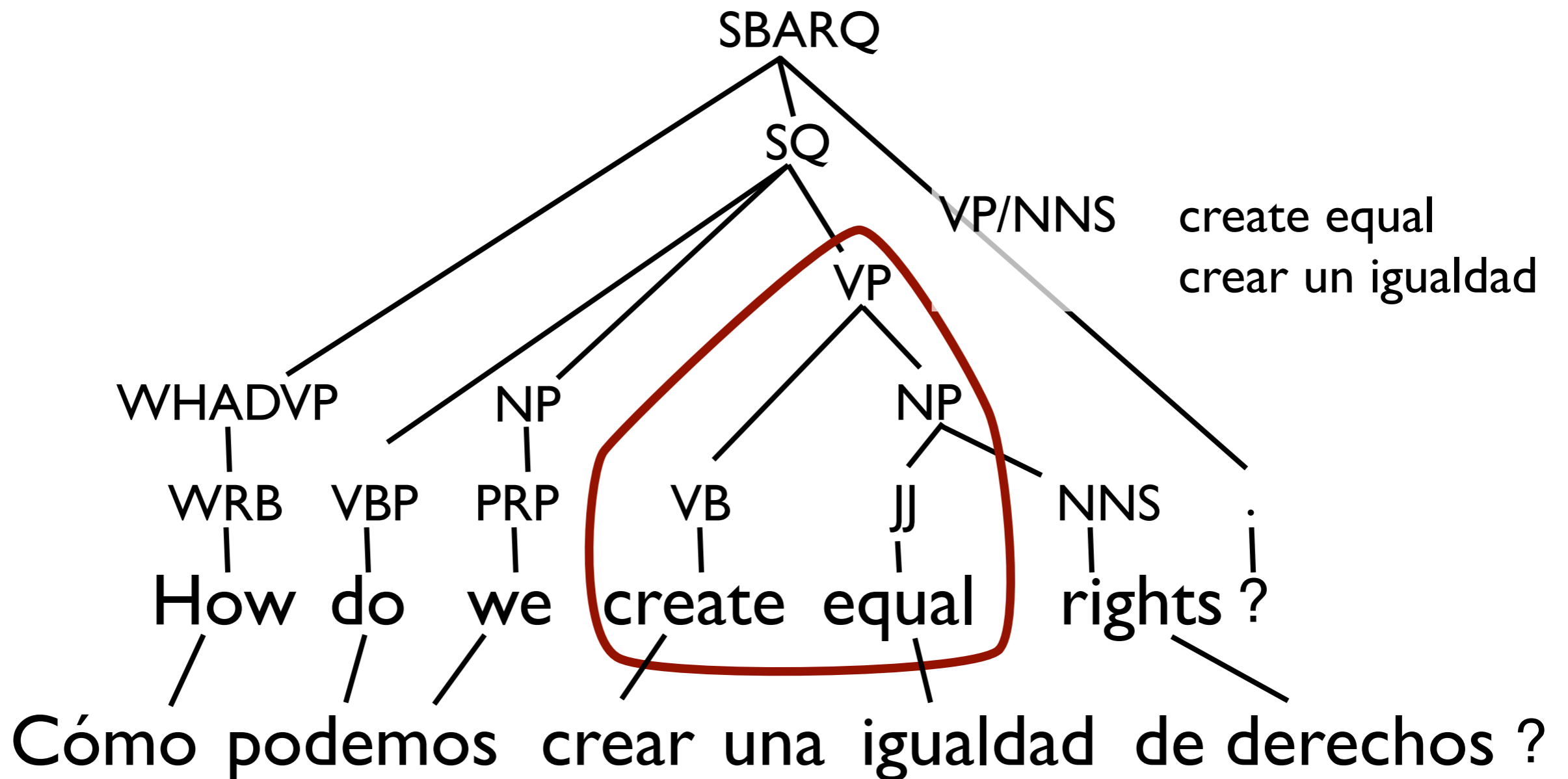
# Phrase Extraction + Syntactic Labels



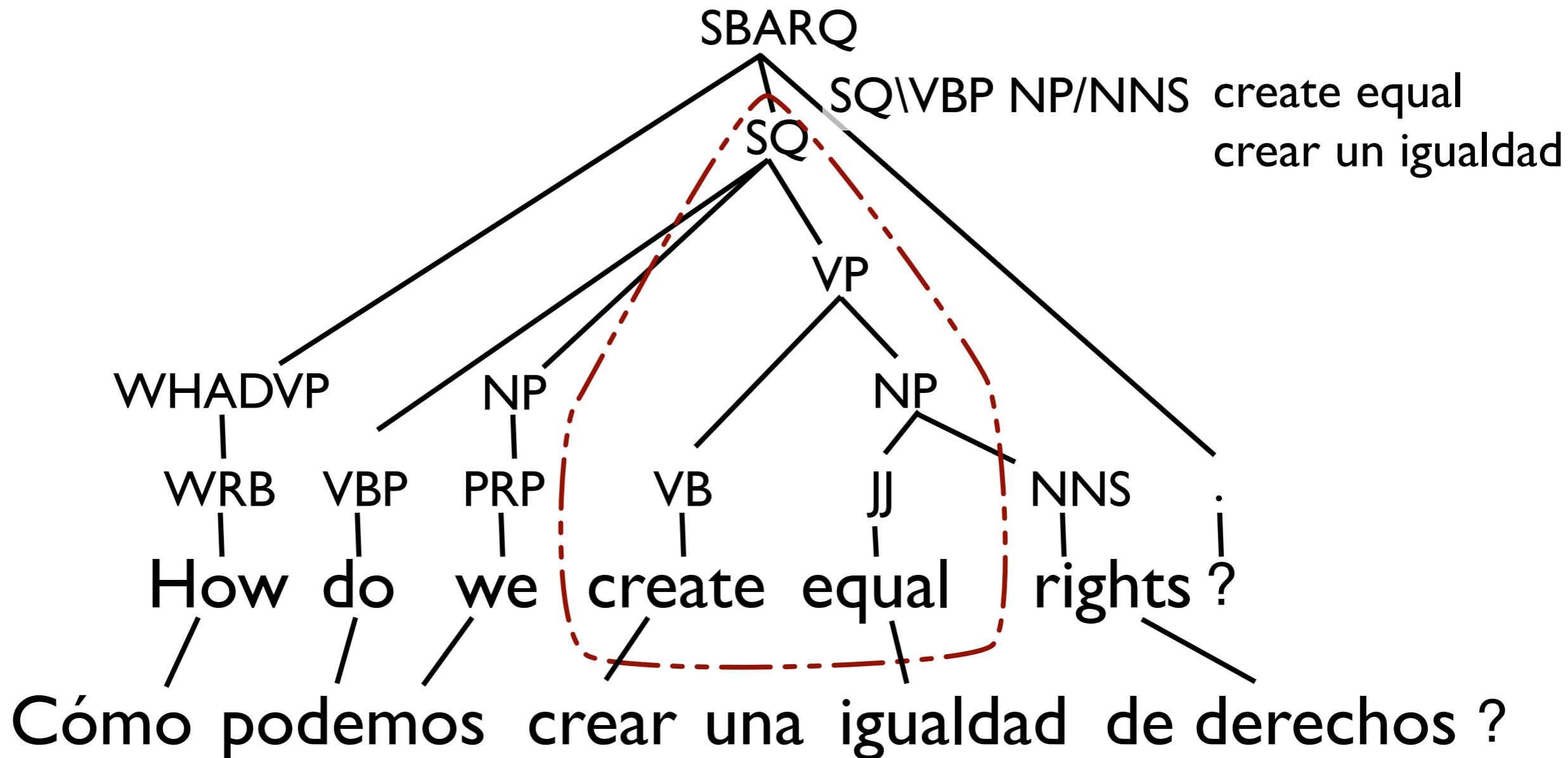
# Using Complex Syntactic Labels



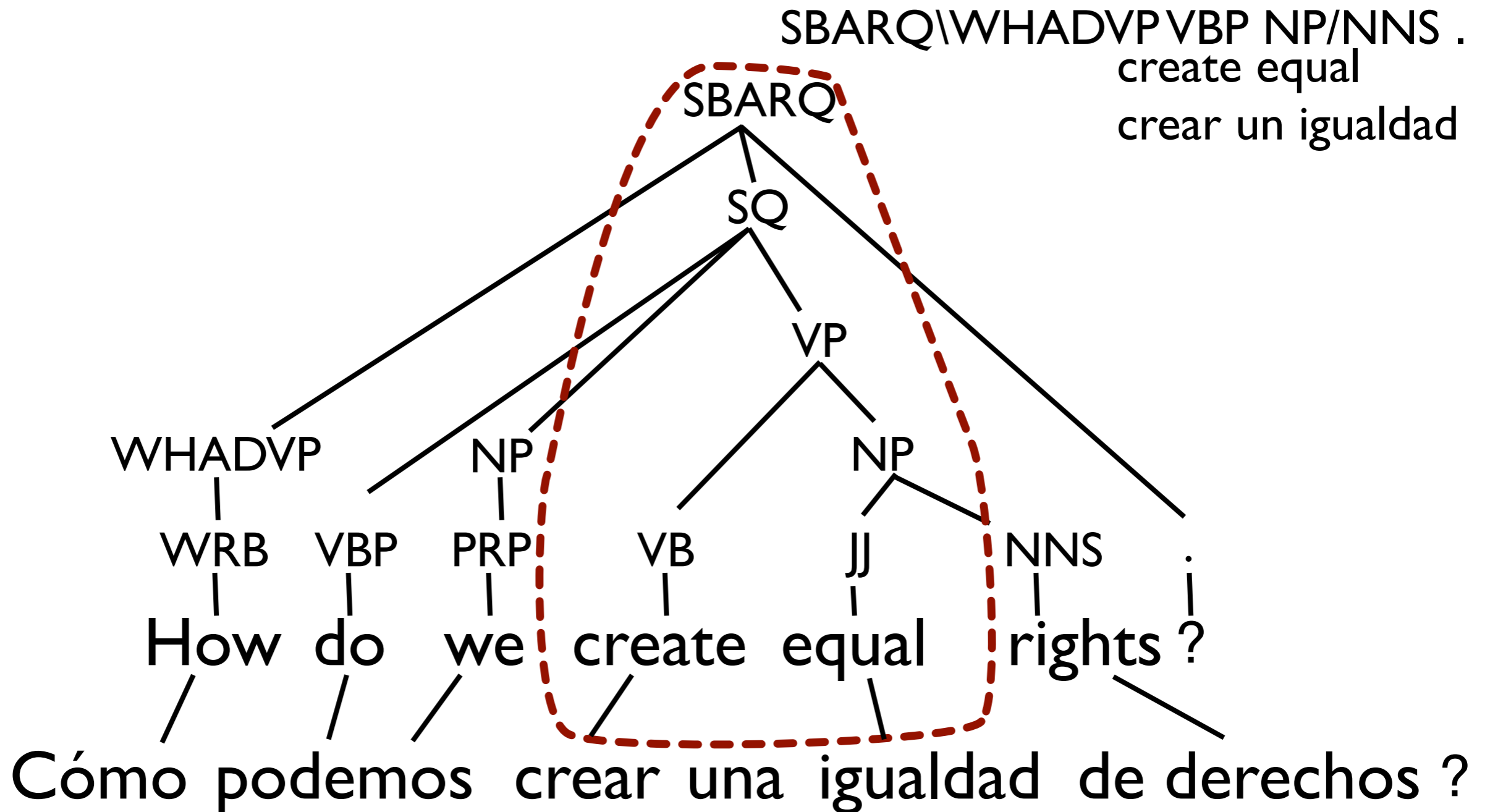
# Using Complex Syntactic Labels



# Using Complex Syntactic Labels



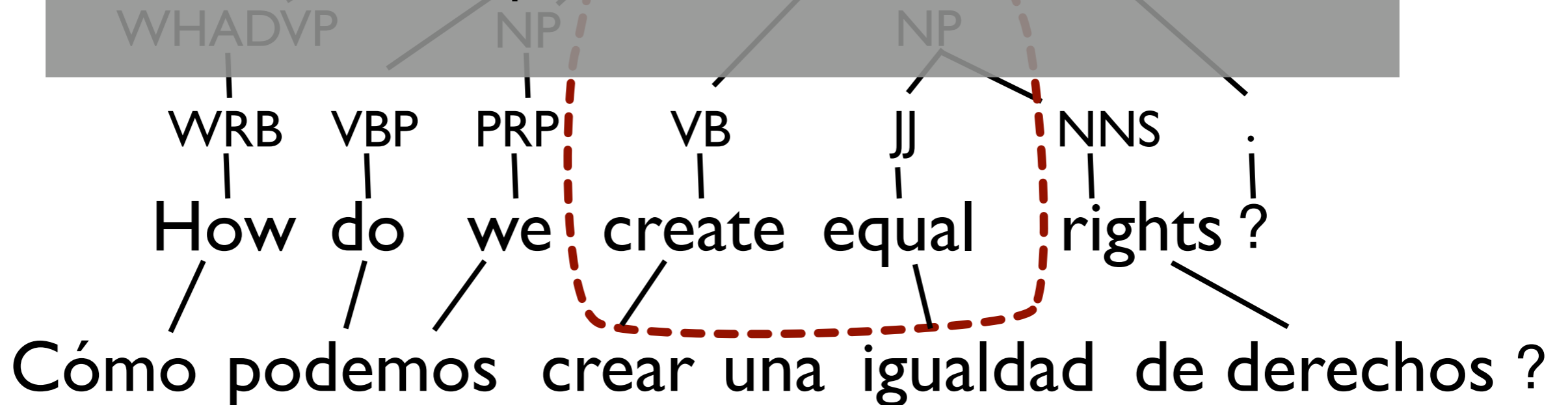
# Using Complex Syntactic Labels



# Using Complex Syntactic Labels

SBARQ\WHADVPVBP NP/NNS .  
create equal  
crear un igualdad

- Coverage improves 3x over simple labels
- Covers 2/3 of phrases that the baseline does



# Example improvements

- **create equal** → equal, to create a, create, to create equality, same, created, conditions
- **create equal (VP/NNS)** → creating equal
- **create equal (VP/NNS PP)** → promote equal, establish fair
- **create equal (VP/NNS PP PP)** → creating equal, provide equal, create genuinely fair

# Example improvements

- **equal** → same, equality, equals, equally, the, fair, equal rights
- **equal (JJ)** → same, fair, similar, equivalent
- **equal (ADJP)** → necessary, similar, identical, the same, equal in law, equivalent

# Manual Evaluation

- Paraphrases were substituted into a number of sentences containing the original phrase
- Judges were asked if the resulting sentence
  - Preserved the meaning
  - Remained grammatical
- A total of 8,500 judgements were collected over several models

# Experimental conditions

- Tested the baseline model and two syntactically constrained models
- Constraints can apply in two places
  - During the phrase extraction stage
  - When replacing a phrase with its paraphrase in a sentence
- Also re-ranked the results of all of these with a trigram LM

# Training data

- Paraphrase models were all trained in the Europarl corpora
  - 10 bilingual parallel corpora with 30 million words each
  - Total of 315 million English words
- English side parsed with Bikel parser trained on WSJ. 1.3 million sentences parsed in total.
- Same sentences used to train SRILM

# Training data

- Paraphrase models were all trained in the Europarl corpora

- 10 bilingual parallel corpora with 30 million words each

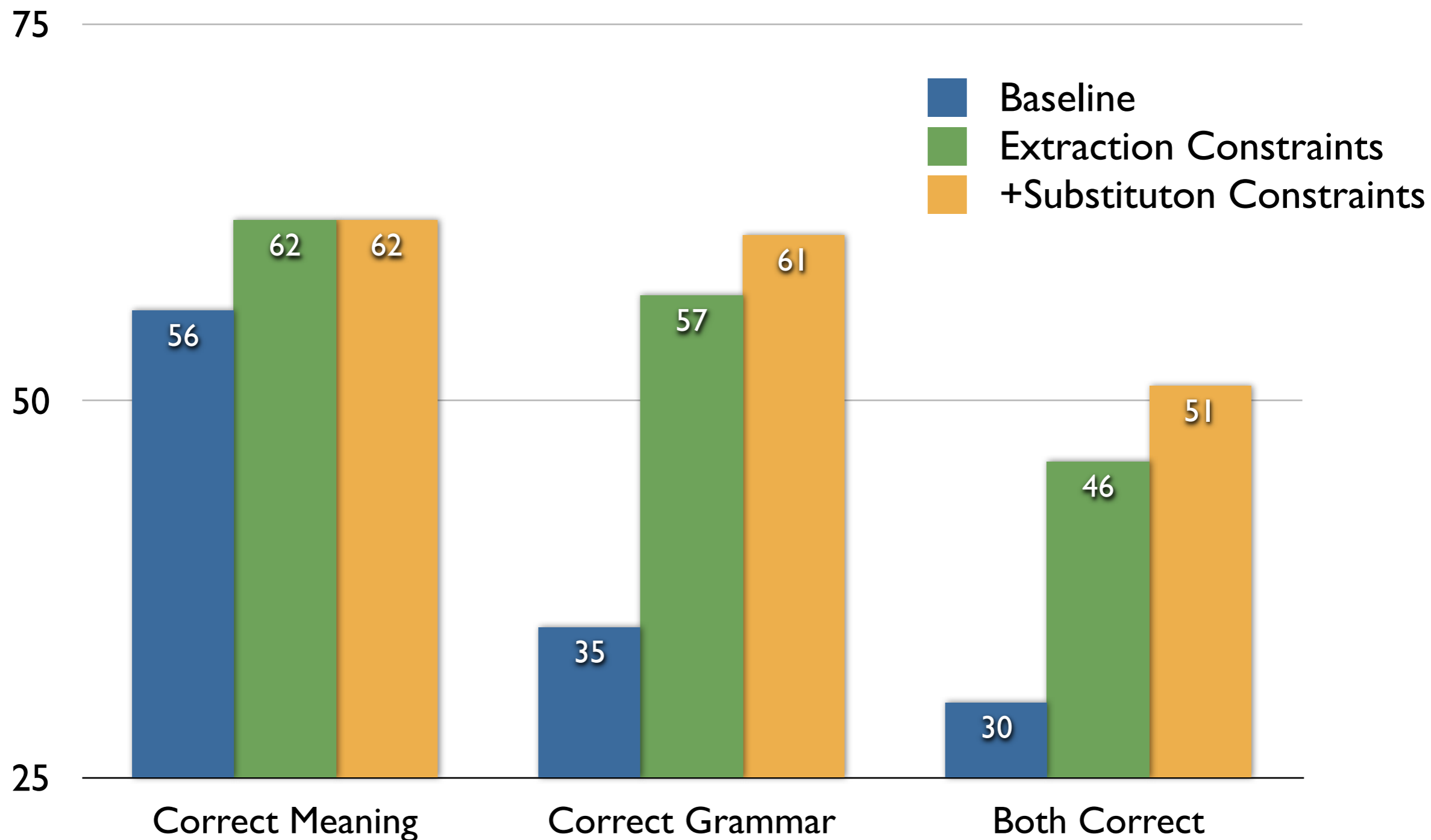
- Total of 315 million English words

- All data and software is available from my web page: <http://cs.jhu.edu/~ccb/>

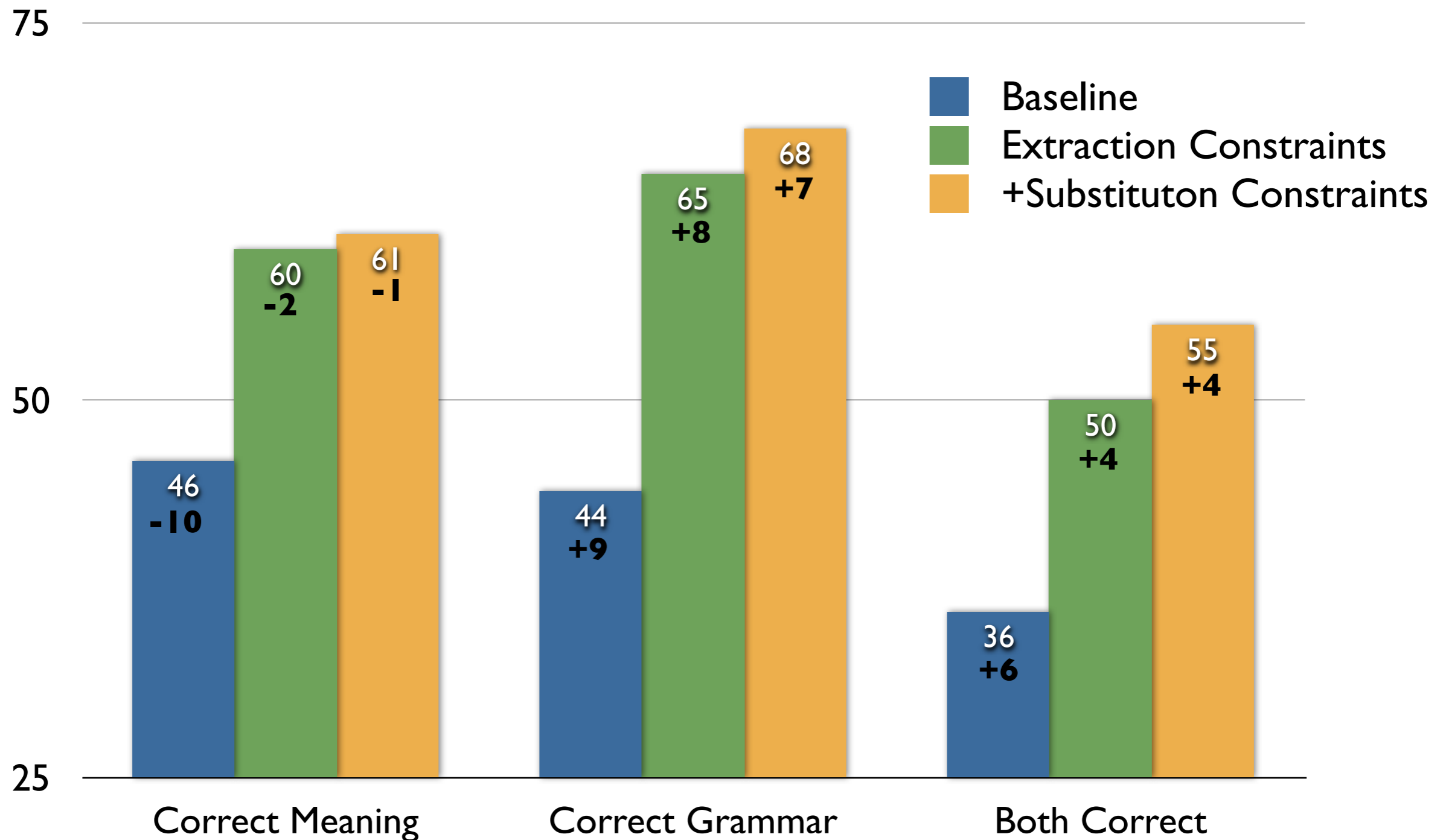
- English parser trained on WSJ. 1.3 million sentences parsed in total.

- Same sentences used to train SRILM

# Initial Results (w/o LM)



# Adding a language model



# Conclusions

- Syntactic constraints reduce errors due to misalignments
- Complex syntactic labels allow us to retain the high coverage of the baseline
- Result in higher paraphrase quality both in terms of grammaticality and in overall quality
  - 24% absolute improvement in correct grammar
  - 19% absolute improvement in overall correctness

# Future Work

- Apply syntactic constraints to paraphrasing techniques that use monolingual corpora
- Extract structural paraphrasing rules

# Thanks!

- Questions?