Automatic Extraction of Clausal Embedding Based on Large-Scale English Text Data

Iona Carslaw[1,2], Sivan Milton[1,2], Nicolas Navarre[1,2],
Ciyang Qing[2], Wataru Uegaki[2]

[1]School of Informatics, the University of Edinburgh
[2]School of Philosophy, Psychology, & Language Sciences, the University of Edinburgh
{i.c.a.carslaw, s.milton, n.s.navarre}@sms.ed.ac.uk {cqing, w.uegaki}@ed.ac.uk

Introduction

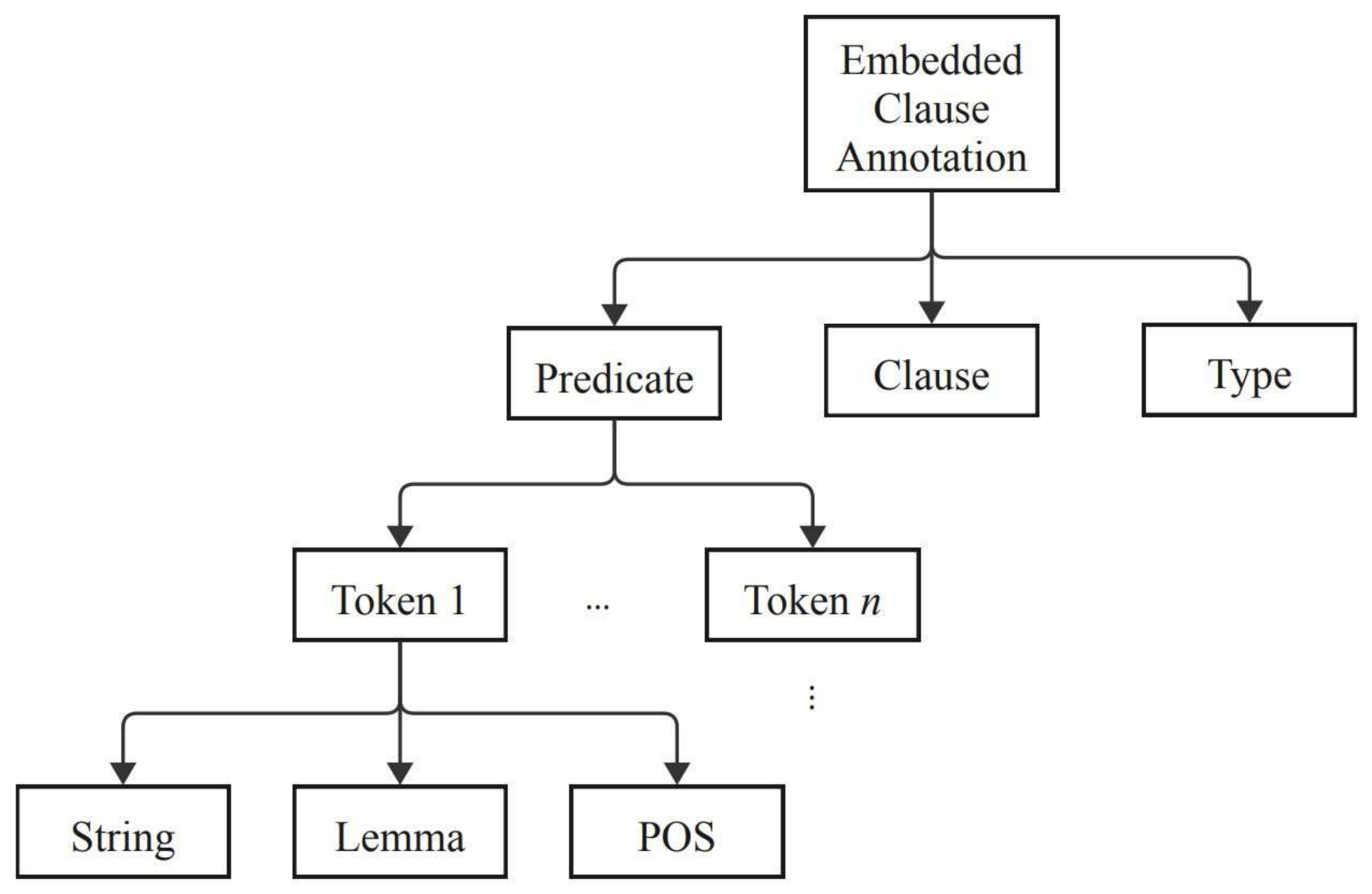
Embedded Clauses (ECs): [clausal complements] selected by embedding predicates

1) a. Mary ✓hoped/*wondered [that John liked chocolate]. [Declarative]
b. Mary *hoped/✓wondered [whether John liked chocolate]. [Polar Interrogative]
c. Mary *hoped/✓wondered [whether John liked chocolate or cake]. [Alternative Interrogative]
d. Mary *hoped/✓wondered [which chocolate John liked]. [Constituent Interrogative]

Big-picture research goal: a data-driven, scalable approach to analyze clause-embedding phenomena (based on naturalistic data)

Our contributions:
(i) a small-scale English dataset with fine-grained gold standard annotation
(ii) a parser tool to detect and annotate English embedded clauses
(iii) a large-scale extracted set of English embedded clauses

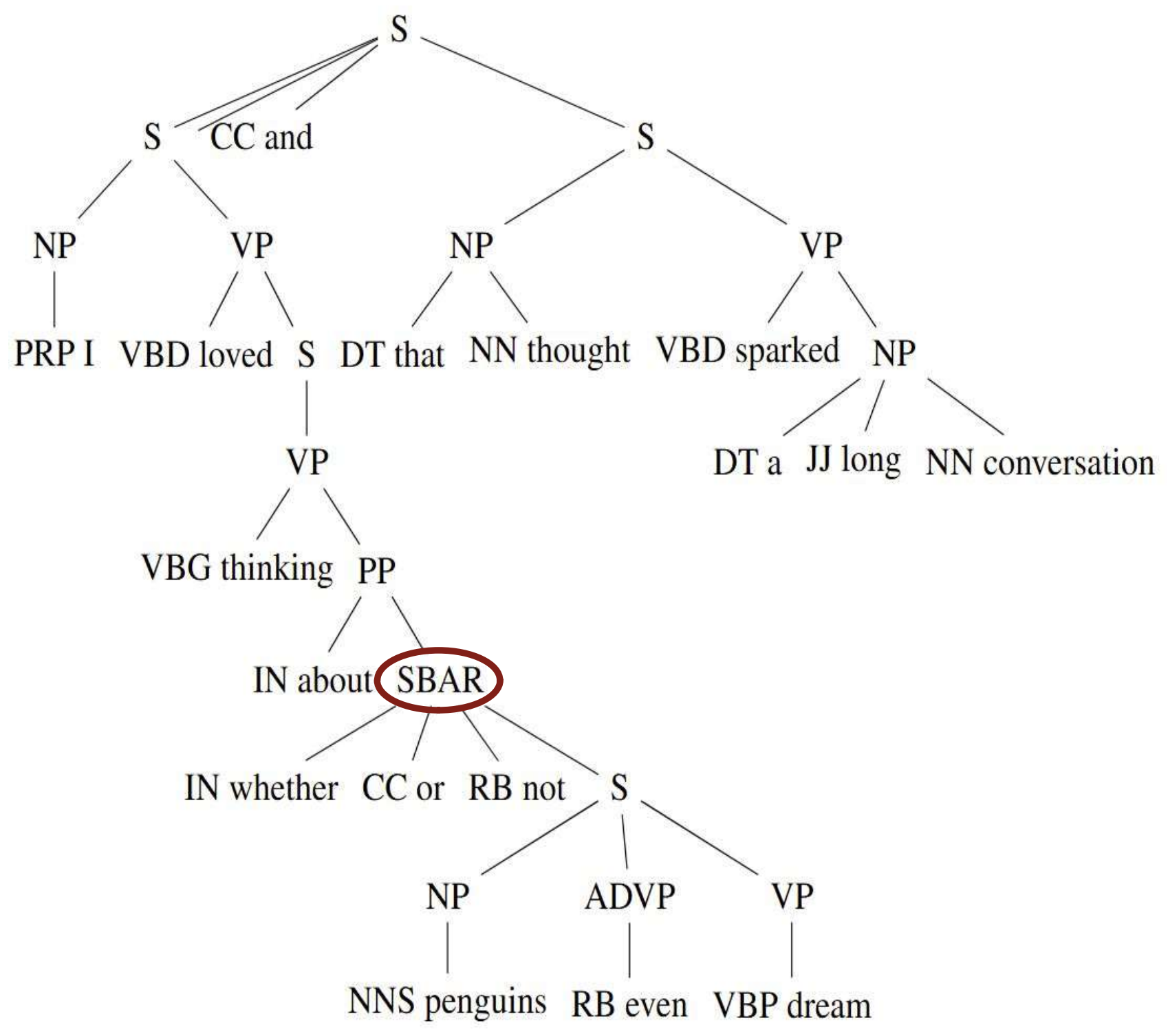
EC Annotation



Parser Tool for Detecting and Annotating Embedded Clauses

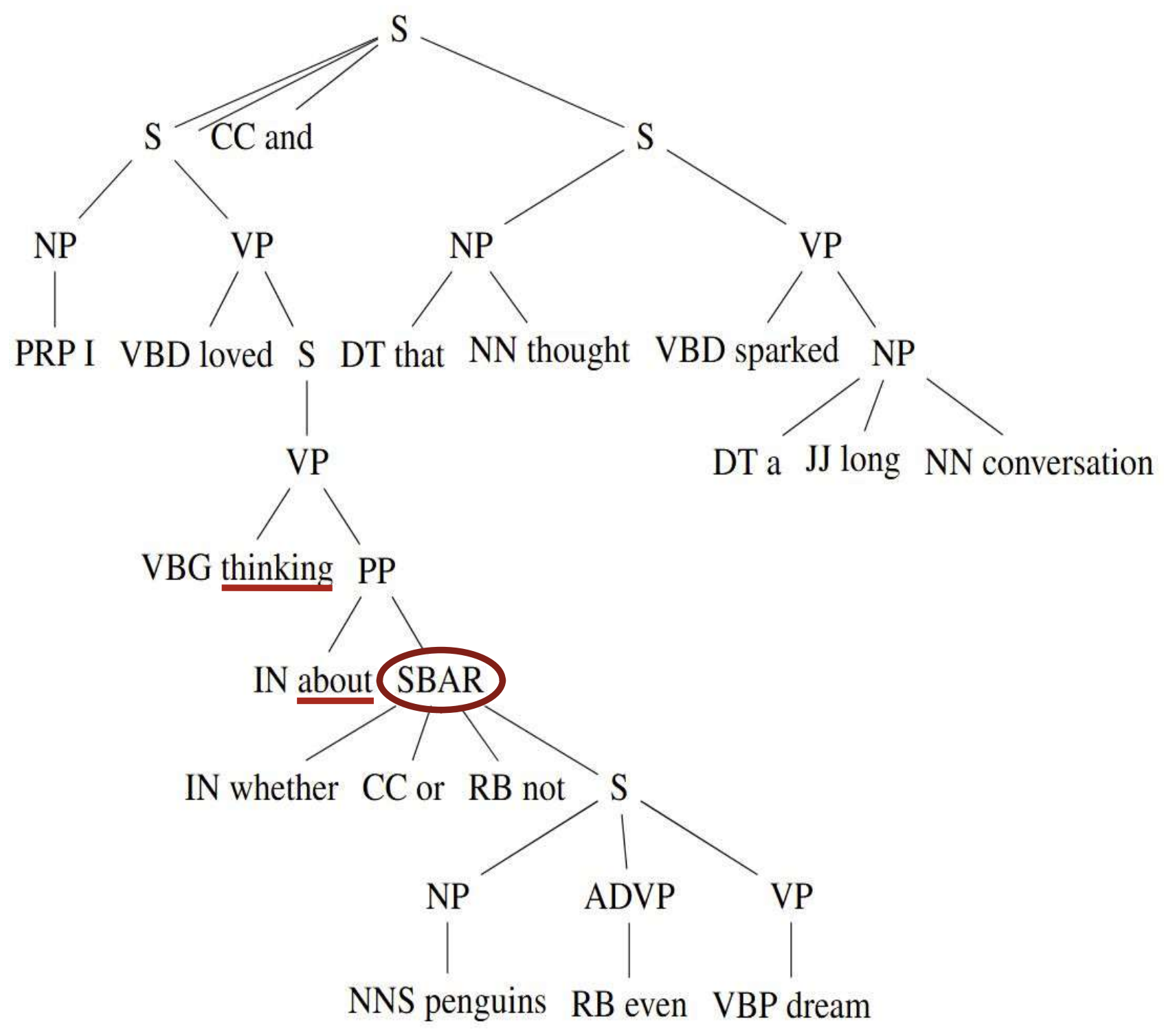
Detection

Find SBAR dominated by a VP



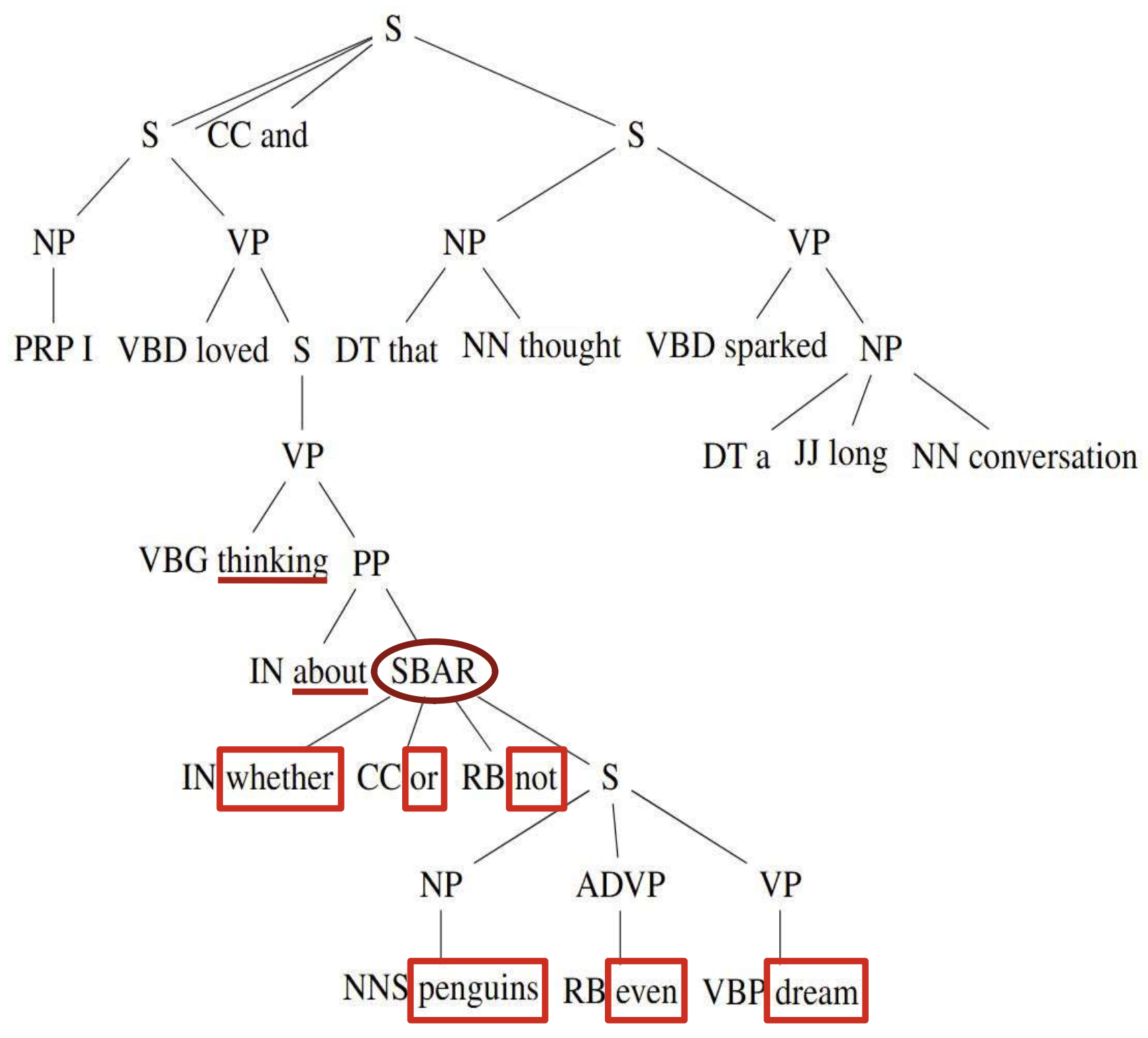
Predicate Identification

Find VP parent tokens until EC



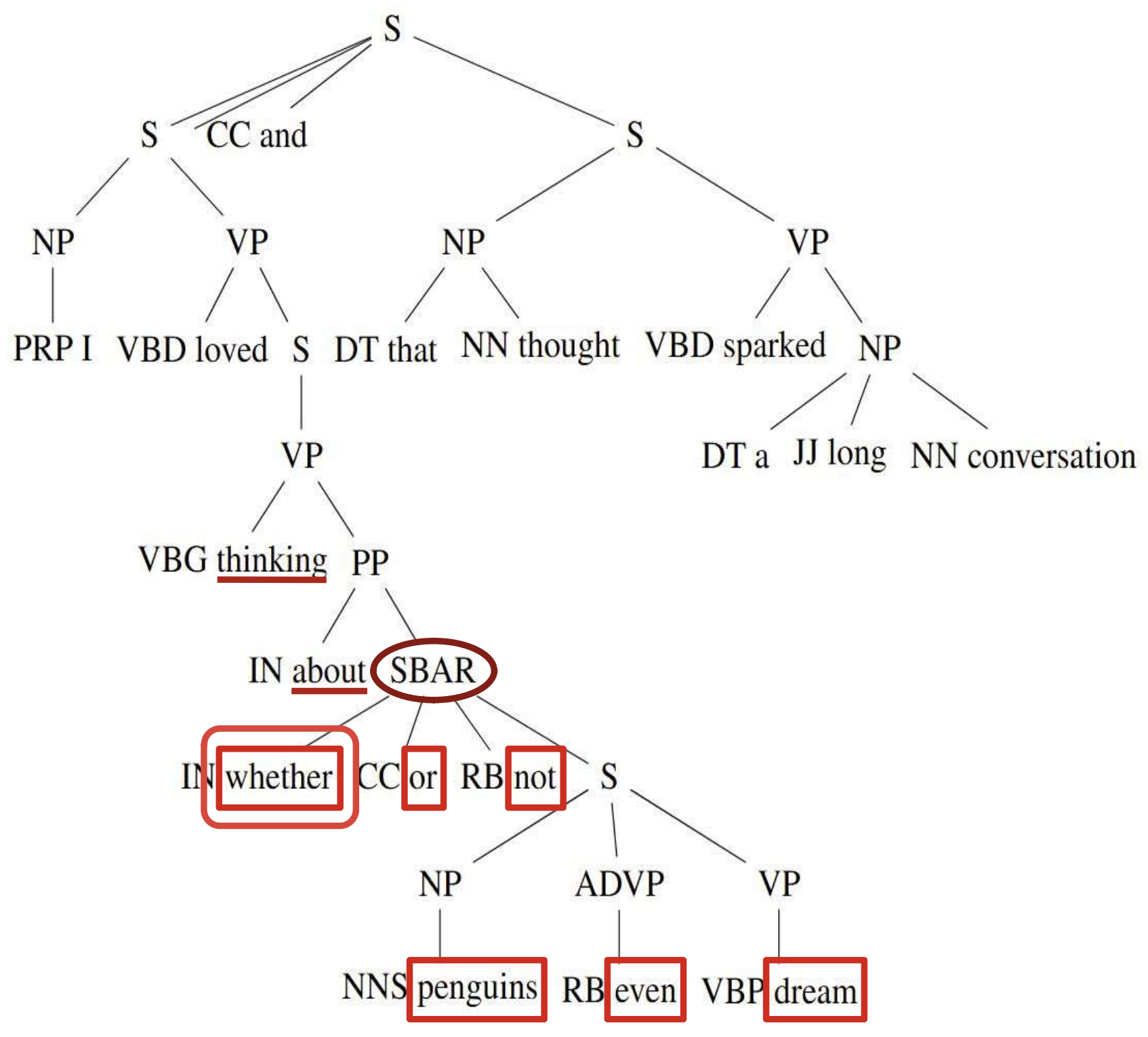
Clause Identification

Find the tokens in SBAR



Typing

Find the complementizer



Sentence: I loved thinking about whether or not penguins even dream and that thought sparked long conversation.

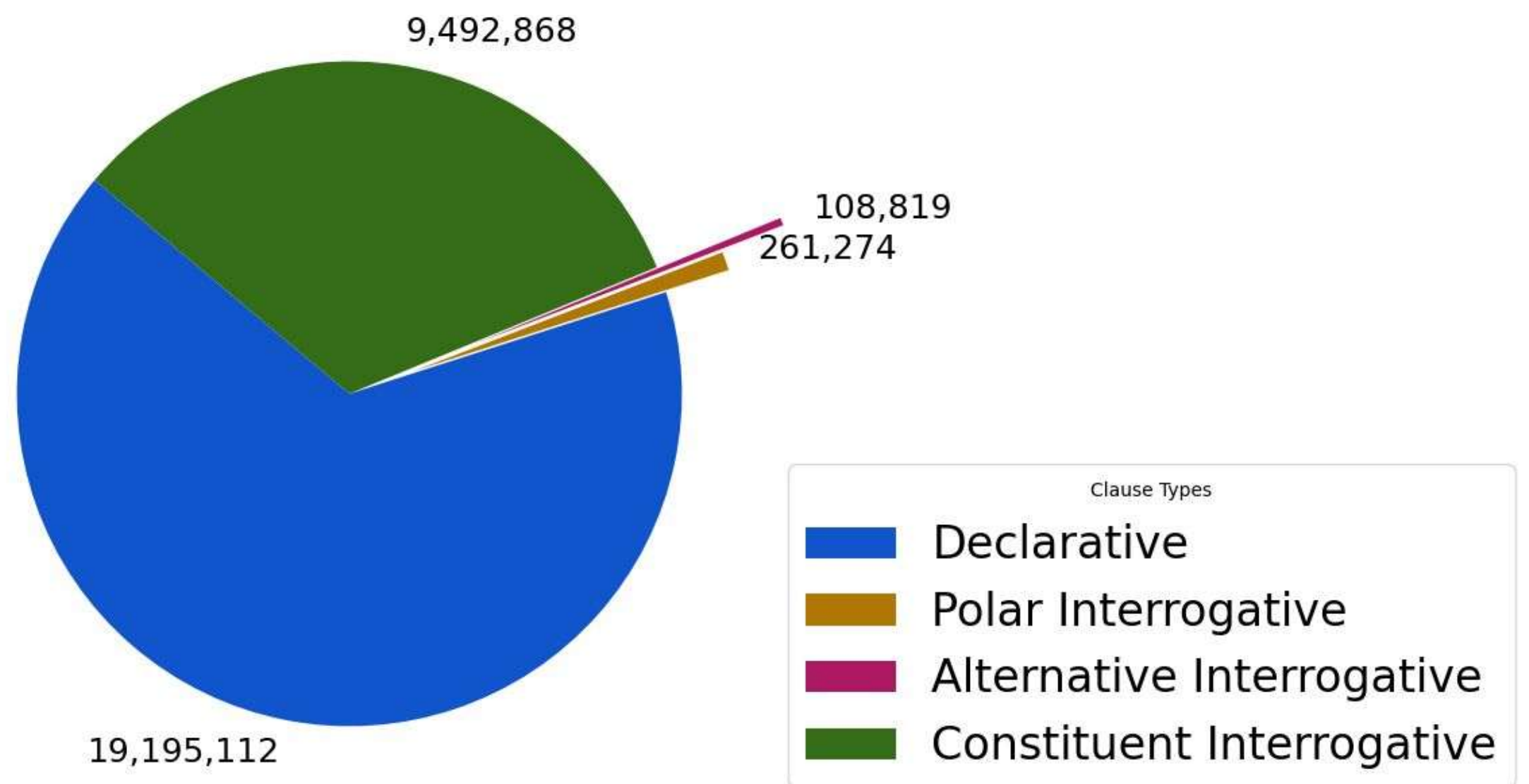
Large-Scale Dataset

We evaluate the accuracy of our parser tool on our manually annotated Golden Embedded Clauses Set (GECS).

Detection	Predicate Identification	Clause Identification	Typing
0.91	0.91	0.87	0.96

Note: Detection is given as an F1 score, while the rest are given as accuracy scores on the true positives set from the Detection task.

Using our parser tool, we detect and annotate embedded clauses from a subset of Dolma (Soldaini et al., 2023). As a result, we have a large-scale dataset of 29,000,000 embedded clause examples with the following statistical breakdown:



Case Study: Emotive Factive Predicates

Past research has found that emotive factive predicates (e.g. *be happy*, *be glad*) are not able to embed polar or alternative interrogative clauses (Karttunen, 1977; Abels, 2004; Sæbø, 2007).

2) a. *Mary was happy [whether John liked chocolate].
b. *Mary was happy [whether John liked chocolate or cake].

Within the large-scale dataset, we search for emotive predicates, with the following results:

# Declarative	Polar Interrogative	Alternative Interrogative	Constituent Interrogative
175,479	159	134	47,877

We find a statistical breakdown matching the above generalisation. However, we also find some genuine counter-examples:

3) a. In the post you talk about your child's health issues and in the end ask if people are happy with [whether they've exercised or not].
b. You might be surprised about [whether there's hope for future scooters].

THE UNIVERSITY of EDINBURGH
NLP UKRI CENTRE FOR DOCTORAL TRAINING
UK Research and Innovation