

# Bridging Discourse Treebanks with a Unified Rhetorical Structure Parser

Elena Chistova

[chistova@isa.ru](mailto:chistova@isa.ru)

# RST Foundations

RST<sup>1</sup> gives us a structured representation of the discourse flow:

- A document is represented as a constituency tree.
- Elementary Discourse Units (EDUs) are non-overlapping spans forming the leaves of the tree.
- EDUs are connected by specific discourse relations (e.g., CAUSE, ELABORATION, CONTRAST).
- Each relation defines a Nucleus (the central idea) and a Satellite (the secondary information).

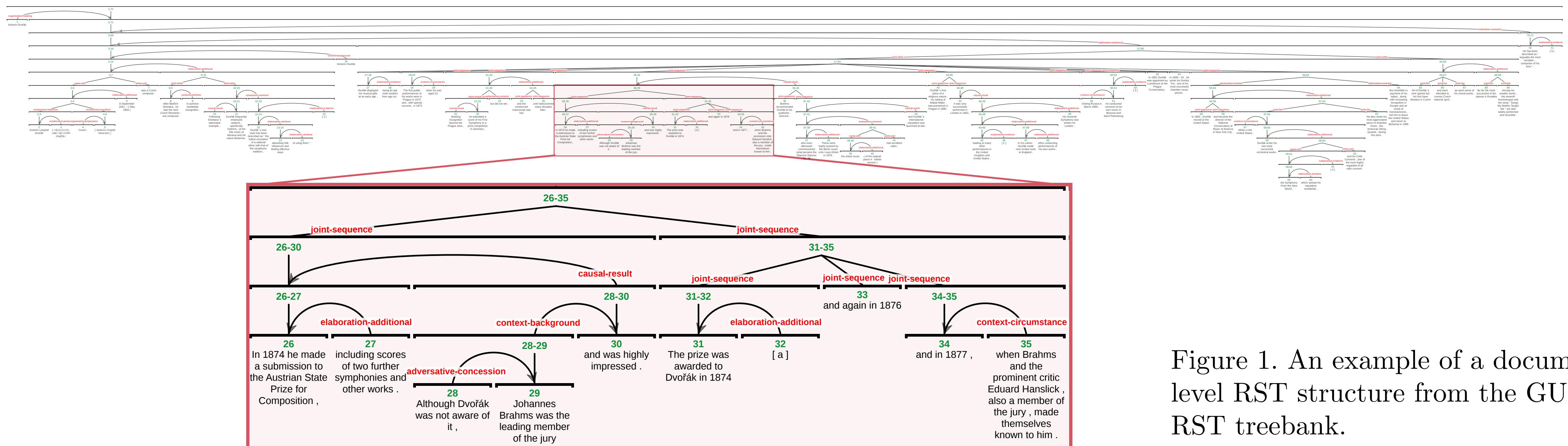


Figure 1. An example of a document-level RST structure from the GUM<sup>2</sup> RST treebank.

<sup>1</sup> Mann, William C., and Sandra A. Thompson. Rhetorical structure theory: A theory of text organization (SCIENCES 1987)

<sup>2</sup> Zeldes, Amir et al. eRST: A Signaled Graph Theory of Discourse Relations and Organization (CL 2025)

# Motivation

The high cost of manual annotation and the diverse interpretation of RST created two primary resource limitations:

## Resource Scarcity

- Creating comprehensive RST Treebanks is extremely time-consuming and expensive.
- This results in small corpora.
- Traditional single-corpus models struggle with data sparsity and lack generalizability.

## Interpretation Variability

- Different treebanks use different sets of relation labels.
- Relation and EDU definitions vary across treebanks.
- A parser trained on one treebank's specific labels cannot easily utilize data from another, wasting valuable annotated discourse data.

## Challenge

18 RST treebanks across 11 languages use incompatible annotation schemes

## Goal

Build a single parser using all data while preserving each treebank's native relation inventory

# Data: 18 Diverse Treebanks

- The 18 RST treebanks cover 11 languages: Basque, Chinese, Czech, Dutch, English, French, German, Persian, Portuguese, Russian, and Spanish.
- Data size varies dramatically from eng.gum<sup>1</sup> with over 34,000 EDUs to ces.crdt<sup>2</sup> with 1,345 EDUs.
- Combined, the treebanks feature 96 unique relation classes (`LABEL_NUCLEARITY`), with major differences in distribution and overlap.

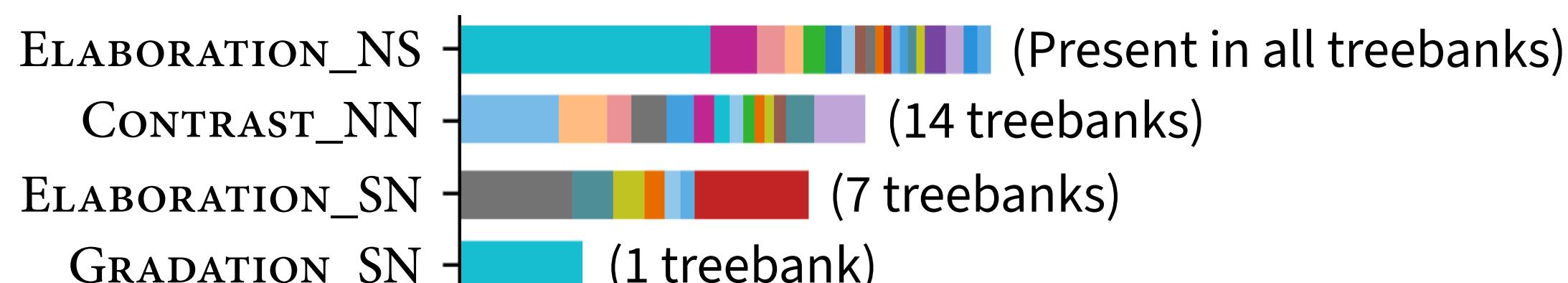


Figure 2. Relation class distribution (log scale).

Treebank	# EDUs	# Classes
ces.crdt	1,345	34
deu.pcc	2,842	37
eng.gum	34,428	27
eng.oll	3,026	35
eng.rstdt	21,789	42
eng.sts	3,208	35
eng.umuc	5,421	46
eus.ert	2,509	31
fas.prstc	5,789	26
fra.annodis	3,307	20
nld.nldt	2,326	45
por.cstn	5,527	38
rus.rrg	25,222	27
rus.rrt	28,247	25
spa.rststb	3,351	43
spa.sctb	744	26
zho.gcdt	9,403	28
zho.sctb	744	26

Table 1. Treebank statistics.

<sup>1</sup> Poláková, Lucie et al. Developing a Rhetorical Structure Theory Treebank for Czech (LREC-COLING 2024)

<sup>2</sup> Zeldes, Amir et al. eRST: A Signaled Graph Theory of Discourse Relations and Organization (CL 2025)

# UniRST Framework

The core end-to-end RST parser architecture is based on DMRST<sup>1</sup> with xlm-roberta-large encoder.  
We propose two strategies for **jointly modeling multiple relation inventories**:

## (a) Multi-Head (MH)

A separate classification layer is used for each distinct relation inventory.

## (b) Masked-Union (MU)

A single classification layer covers all unique labels across treebanks.

A treebank-specific relation mask ( $m_k$ ) enables direct knowledge transfer for overlapping relations (e.g., COMPARISON\_NN).

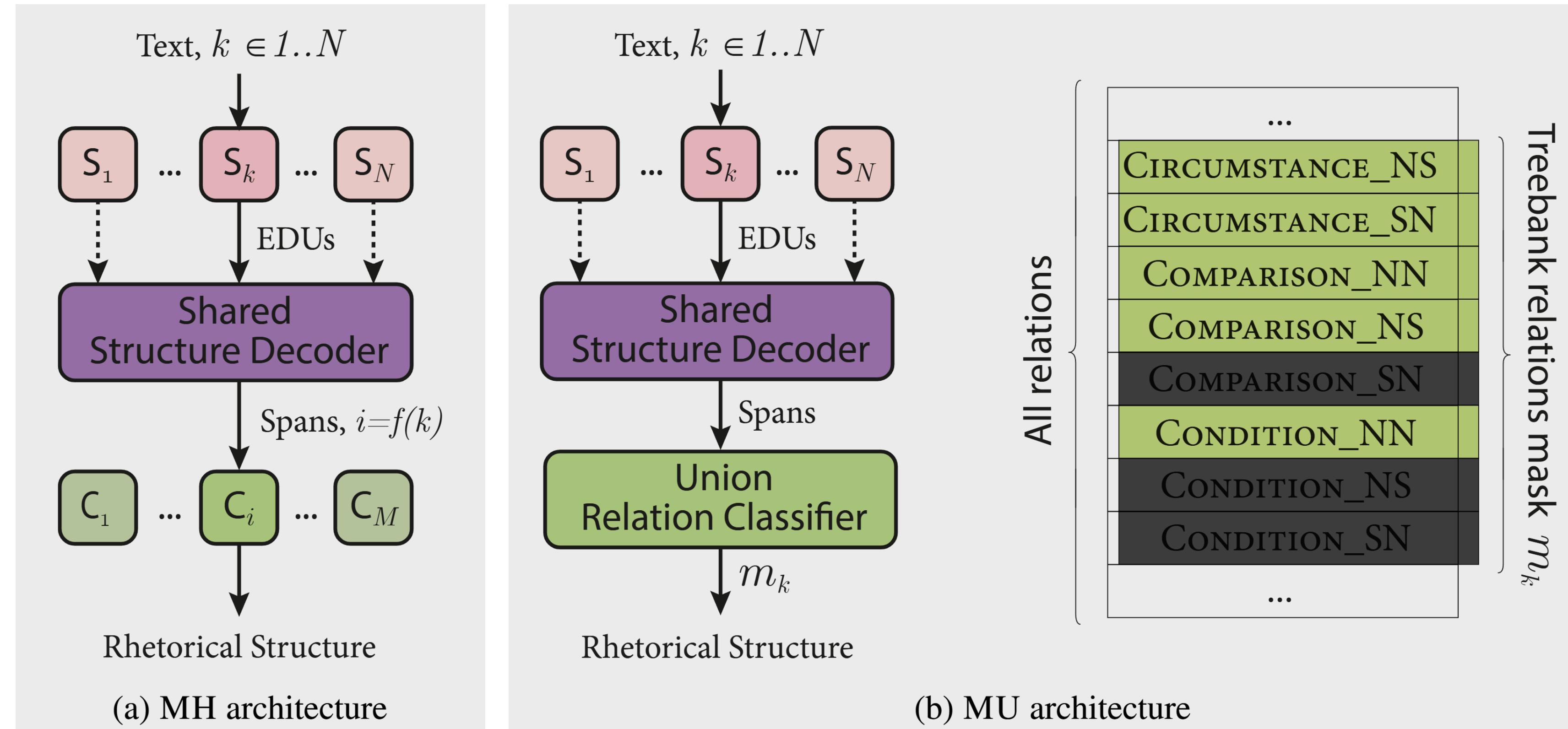


Figure 3. Model variants in the UniRST framework.

<sup>1</sup>Liu, Zhengyuan et al. DMRST: A Joint Framework for Document-Level Multilingual RST Discourse Segmentation and Parsing (CODI 2021)

# Experimental Results: Mono-Treebank Baselines

- We additionally tested in-treebank data augmentation by extracting coherent subtrees:
  - Take a fully annotated document tree.
    - ↓
    - Extract all structurally complete subtrees that contain at least 3 relations and don't break sentences.
      - ↓
      - Add a sample of these smaller subtrees to the training data.
- Data augmentation was beneficial for 10 of the 18 treebanks, particularly smaller ones, but was less effective on the largest corpora.
- These 18 separate models motivate the need for a unified RST parser.

Treebank	Best	Full	F1
ces.crdt			12.3
deu.pcc			21.2
eng.gum			47.4
eng.oll			24.0
eng.rstdt			52.9
eng.sts			18.0
eng.umuc			24.0
eus.ert			22.7
fas.prstc			34.4
fra.annodis			28.6
nld.nlldt			27.3
por.cstn			43.8
rus.rrg			46.0
rus.rrt			42.3
spa.rststb			32.8
spa.sctb			27.3
zho.gcdt			42.3
zho.sctb			29.6

Table 2. Best mono-treebank baseline performance (mean from multiple runs). Highlighted scores indicate augmentation.

# Experimental Results: UniRST

- Masked-Union with separate segmentation performs the best.

	Segmentation	Gold Seg	End-to-End
MH	Single	46.6	40.8
	Multiple	47.6	41.6
MU	Single	47.8	42.3
	Multiple	48.3	42.5

Table 3. Performance (Full F1) of the UniRST variants.

- UniRST outperforms 16 of 18 treebank-specific models.
- Gains are most prominent for smaller treebanks:
  - ces.crdt (Czech): +14.5 F1
  - spa.sctb (Spanish): +13.5 F1
  - zho.sctb (Chinese): +11.1 F1

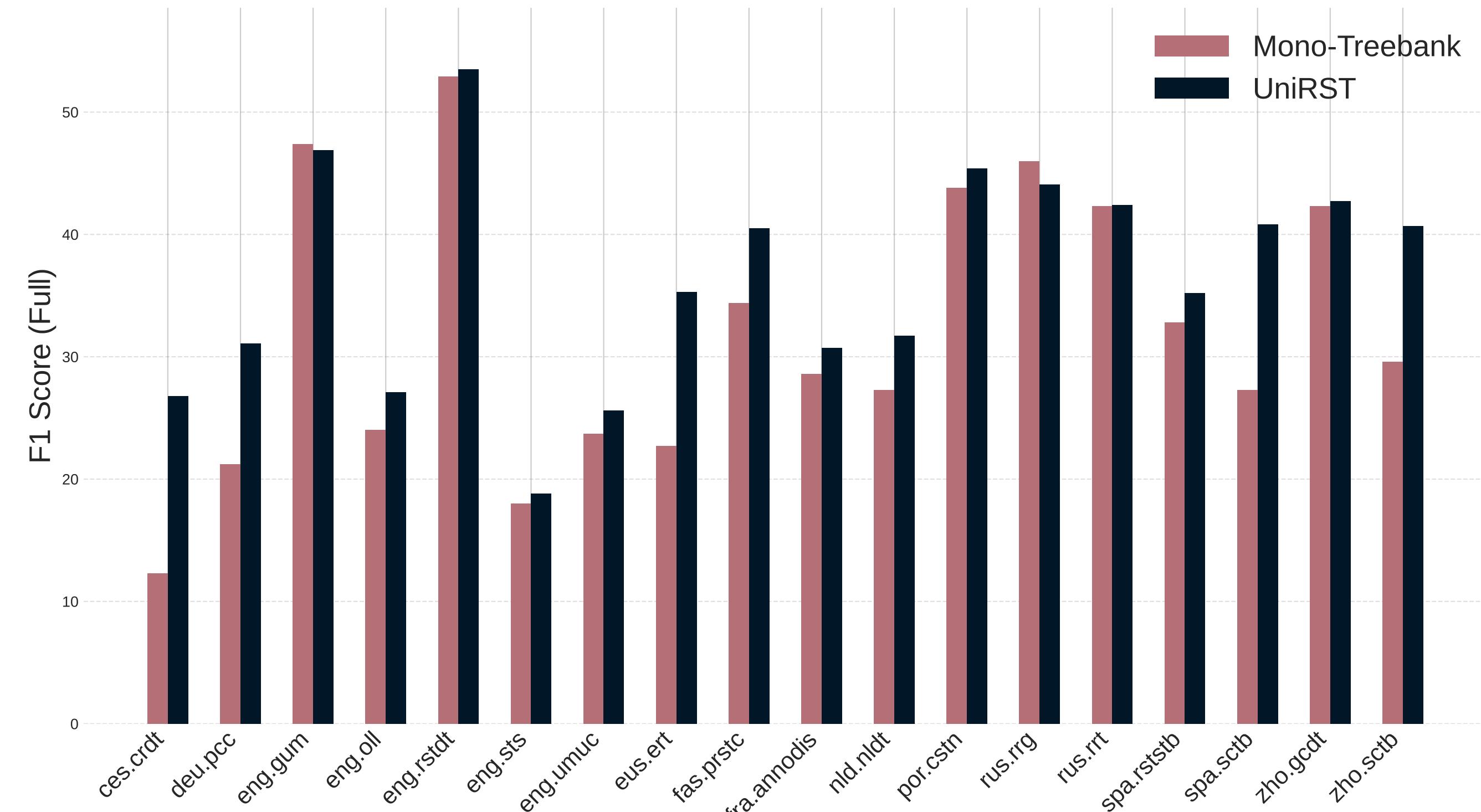


Figure 4. UniRST outperforms mono-treebank baselines.

# Apply UniRST

- Open-source `isanlp_rst` Python library
- End-to-end RST parsing for 11 languages
- Export results to  
.rs3 (RSTWeb), .png, and .pdf

Scan for GitHub Repo

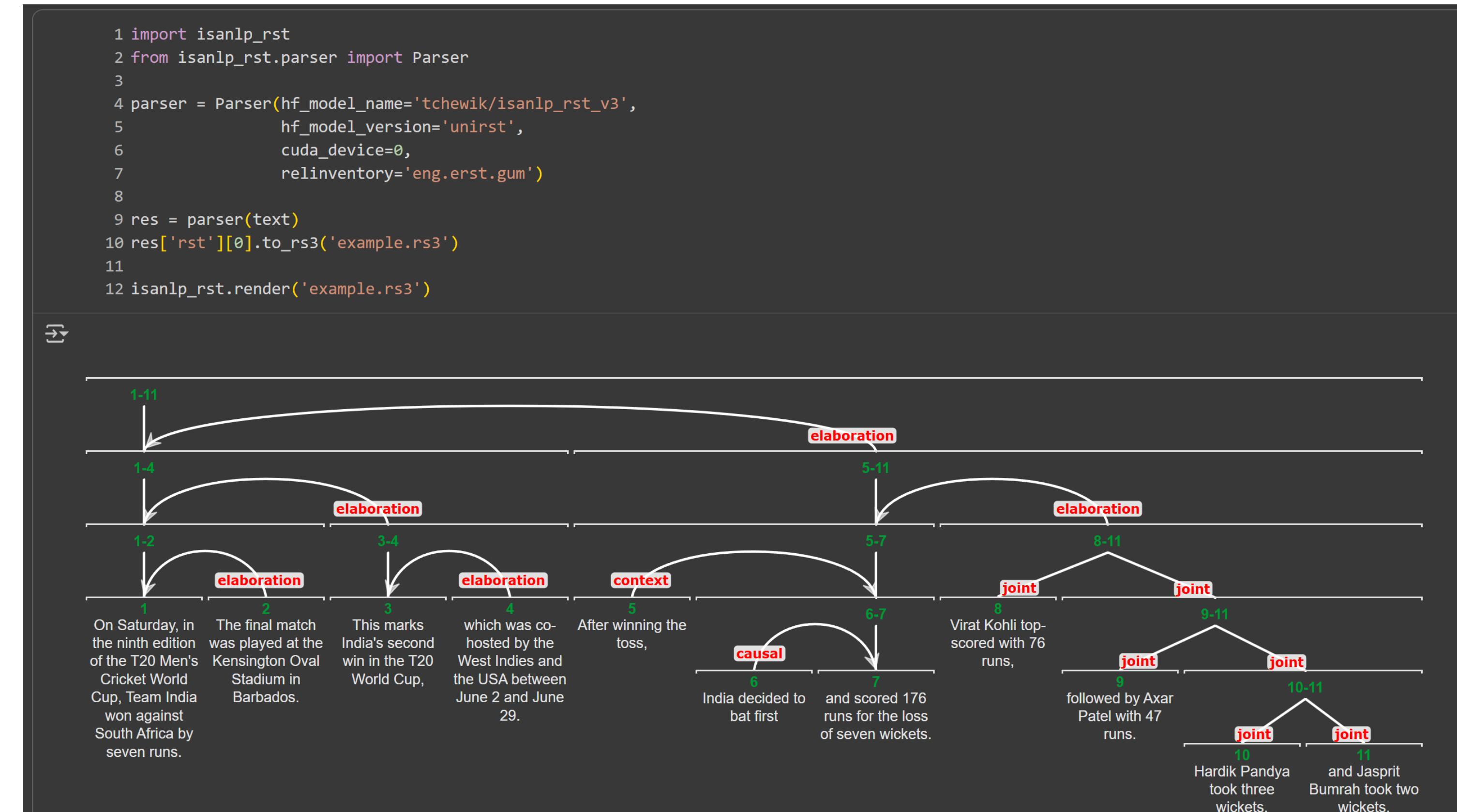
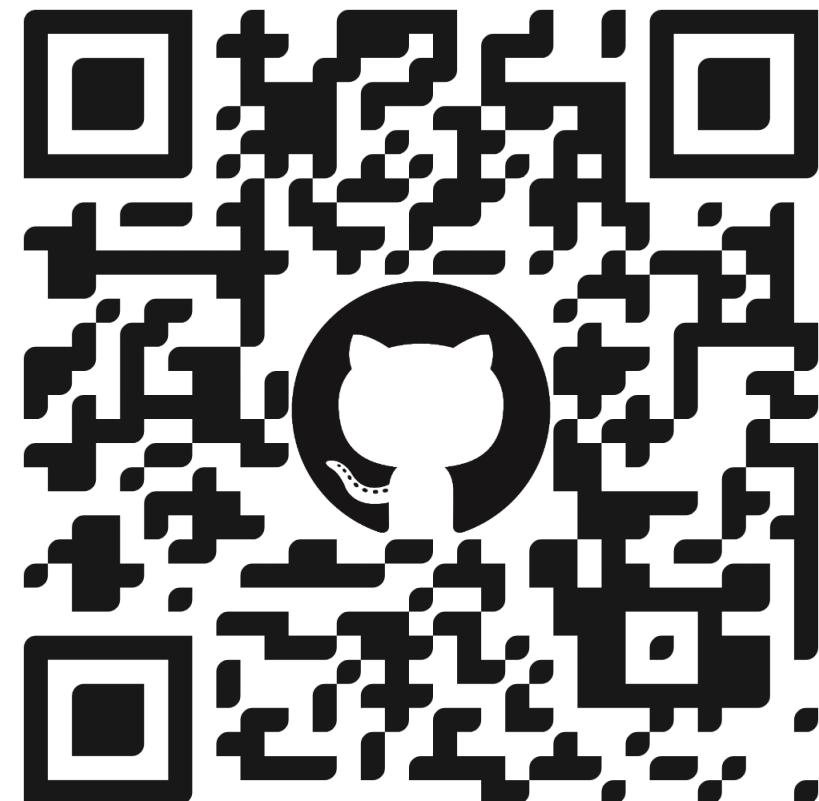


Figure 5. UniRST usage.

# Conclusion

- We introduced UniRST: the first end-to-end discourse parser that jointly handles 18 treebanks across 11 languages without collapsing their native label sets.
- The Masked-Union approach proved most effective, enabling parameter sharing while respecting corpus-specific annotations.
- Our unified multilingual parser outperformed 16 of the 18 strong mono-treebank baselines.
- The model is available as a part of the easy-to-use `isanlp_rst` library.

# Thank you for your attention!

paper & code

