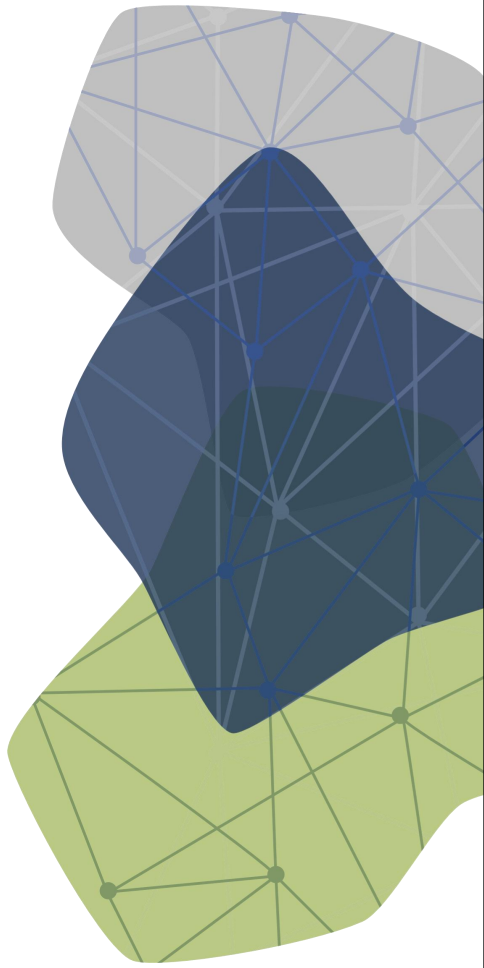
Research Project - 1

Annotating Rhetorical Relations for Discourse parsing

Mehedi Hasan Emon

Advisor Dr. Klabunde

Date: 18.06.2024



LDSL Linguistic Data Science Lab

RUHR UNIVERSITÄT BOCHUM RUB

Overview

- Introduction
- My Research Project
- Conclusion

Introduction

What are Rhetorical Relations?

Rhetorical relations are the logical and functional connections that link different parts of a text. These parts of text, which we will refer to as discourse units, can be clauses, sentences, paragraphs, or larger sections that relate to each other in specific ways. For example, two relative clauses, by a contrast relation.

Example:

Contrast: Ron wears glasses, but he can't read the small print.

Elaboration: Ron wears glasses. They are anti-reflective.

Rhetorical Relations Example?

Let us start with an illustration. Below is a joke from the show Right Wing Comedian performed by the British stand-up comedian Leo Kearse at the Edinburgh Fringe Festival 2018.

a. I've got some sympathy for Trump.

b. He went for a job,

c. tried to throw the interview

d. but accidentally got it

e. and now he hates it.

f. Reminds me of every interview I had for jobs I didn't want when I was on benefits.

Rhetorical Relations Example?

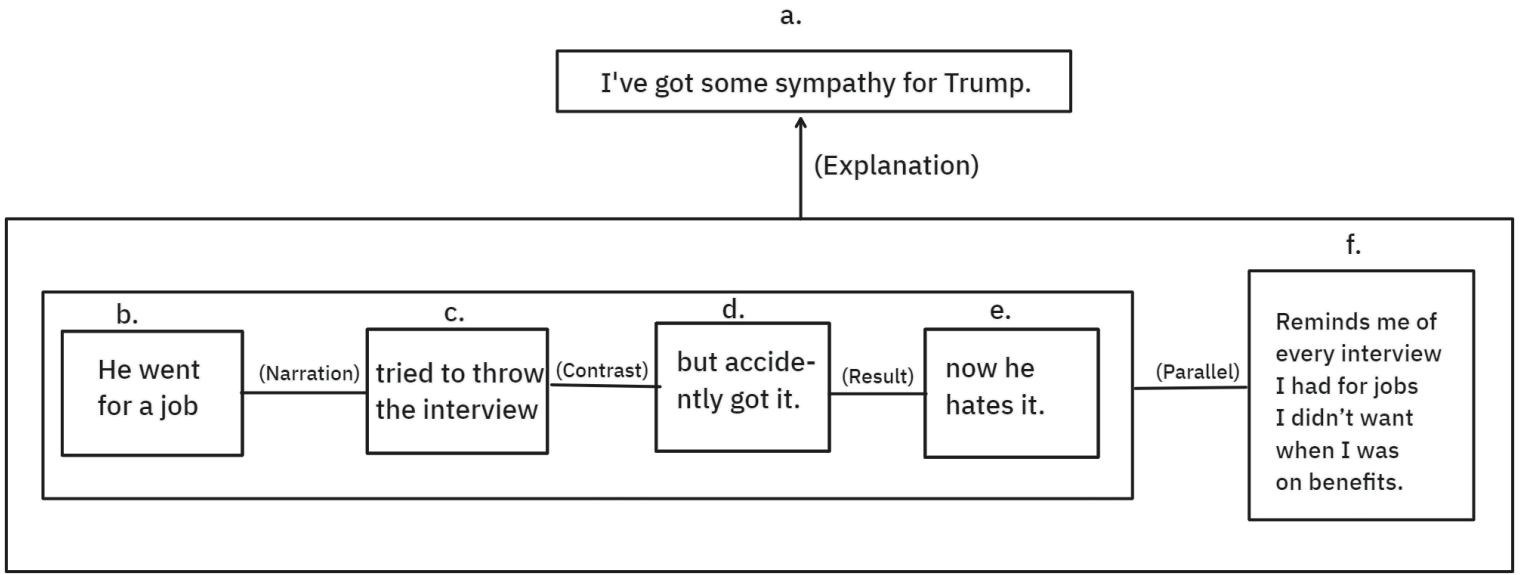


Fig:- Discourse structure for Example text

Type of Rhetorical Relations..

1) Primarily pragmatic relations describe (especially in editorials) the argumentation of the author: What are the claims, and how are they being supported by observations or by other claims?

2) Primarily semantic relations are used when the author describes a (possibly complex) states of a airs in the world; this may involve for example relations of causality among events.

Type of Rhetorical Relations..

3) Textual relations work to organize the text and make its understanding easier by providing orienting information, or repetition.

4) In all three groups above, the relations hold between a nucleus and a satellite, they are thus called mononuclear. In contrast, the fourth group contains multinuclear relations, that connect two or more nucleus.

Type of Rhetorical Relations..

Type	Name of Relations
Primarily pragmatic relations	Background, Antithesis, Concession, Evidence, Reason, Evaluation, Motivation, Enablement
Primarily semantic relations	Circumstance, Condition, Otherwise, Unless, Elaboration, Interpretation, Means, Cause
Textual relations	Preparation, Restatement, Summary
Multinuclear relations	Contrast, Sequence, List, Conjunction, Joint

Some abbreviations..

Shortcut	Abbreviations
N	Nucleus
S	Satellite
W	Writer
R	Reader
N/S	Usually more elaborate in our version.

Nucleus & Satellite..

Nucleus (N): The main content or primary information that you want to convey.

Satellite (S): The background information that makes it easier for the reader to understand the nucleus. Without the background information in the satellite, it would be difficult to comprehend the nucleus.

Example:- [Before he moved to Bochum,]S [he lived in Dhaka.]N

What are Discourse units?

Discourse Units (DUs) are the building blocks of a text, representing the smallest segments that convey a complete thought or idea. Thus, a “basic discourse unit” is a text segment with linguistic properties which are used to construe both semantic representations (interpretations, inferences) and the text and context models at stake. (Degand & Simon, 2009)

Early studies have demonstrated that discourse parsing can benefit various downstream NLP tasks, including sentiment analysis, summarization, question answering and machine translation evaluation. (Hu and Wan 2023)

List of Rhetorical Relations..

Background:-

N/S: Understanding Satellite (S) makes it easier for Reader (R) to understand the content of Nucleus (N); without the background information in S, it would be dificult to comprehend N. In a text, S mostly but not always precedes N. A Background S at the beginng of the text often serves to introduce the topic of the text.

Example:

[Until 1984, Burkina Faso was called Obervolta.]S [According to an EMNID poll, many Europeans today believe that they are two different countries.]N

List of Rhetorical Relations..

Input Sentence:-

[Until 1984, Burkina Faso was called Obervolta.]S [According to an EMNID poll, many Europeans today believe that they are two different countries.]N

DU1: Until 1984, Burkina Faso was called Obervolta.

DU2: According to an EMNID poll, many Europeans today believe that they are two different countries.

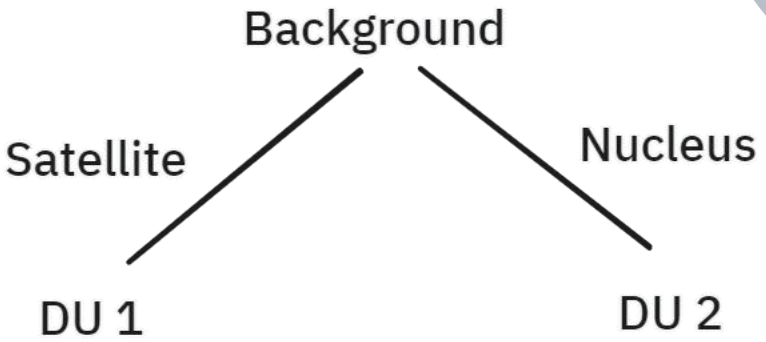


Fig -2 : RST Tree Structure

List of Rhetorical Relations..

Antithesis:-

N/S: The contents of N and S are not compatible often on the level of their evaluation. it involves the juxtaposition of contrasting ideas in parallel structures.

Typical connectives: but; neg - rather; neg - instead; ...

Example:

Attempt of a city to sell off real estate [At one point they seemed to have succeeded.]S [But the buyer didn't pay.]N (maz-6193, PCC)

List of Rhetorical Relations..

Concession:-

N/S: W concedes S and implicitly confirms that S and N are usually not compatible; in the current instance, involves acknowledging a point from the opposing view, often to demonstrate a balanced and fair argument.

Typical connectives: although; but; still; despite; ..

Example:

[Sanitary facilities nowadays are a standard on big campgrounds.]S [But Radewege has a hard time with upgrading to such standards.]N (maz-6488, PCC)

[Although it is expensive,]S [the quality justifies the price.]N

List of Rhetorical Relations..

Elaboration:-

N/S: S provides details or more information on the state of affairs described in N . N precedes S in the text. Typical relations between N and S are set::element, whole::part, abstraction::instance, procedure::step.

Typical connectives: in particular; for example; ...

Example:

[Diepensee will relocate.]N [No question about that.]S (maz-6993, PCC)

[Today thousands of visitors want to sense the atmosphere of a historical classroom,]N [as it exists only in very few places in Germany.]S (maz-6728, PCC)

List of Rhetorical Relations..

Cause/effect:-

N/S: The state/event in N is being caused by the state/event in S.

Typical connectives: because; since; therefore; ...

Example:

[Mayer Jochen Wagner reacted with surprise, too.]N [After all, just on Monday the community council had agreed to expand the village of Diepensee.]S (maz-6993, PCC)

[Because it rained heavily,]S [the football match was postponed.]N

List of Rhetorical Relations..

Contrast:-

N: Exactly two nuclei. Both are of equal importance for W's purposes.The contents are comparable yet not identical. They differ in aspects that are important to W.

Typical connectives: on the other hand; yet; but; ...

Example:

[My first car was small.]N [The second was already a sizable limousine.]N

[I love summer.]N [On the other hand, my brother prefers winter.]N

List of Rhetorical Relations..

Summary:-

N/S: N consists of more than one DUs. S succeeds N in the text and repeats the information given in N, but in a shorter form.

Typical connectives: in short; ...

Example:

[John spent the entire weekend meticulously cleaning his car, inside and out. He washed, waxed, and polished every surface until it gleamed in the sunlight.]N [In short, John dedicated his entire weekend to thoroughly cleaning his car, leaving no surface untouched.]S

My Research Project I

- Data Collection
- Preprocessing
- Annotate RR With annotation tool
- Analyse the results
- Realize automatic discourse parsing

Data Collection..

Sports Report (Cricket):-

Example report [0:1]:-

England pace bowler Jofra Archer could play in this year's T20 World Cup but will not play test cricket until 2025, according to England managing director Rob Key.The 29-year-old has not played for England in any format since March 2023 because of an elbow injury. Archer's England career has been plagued by injuries and he has not played a test match since the tour of India in 2021. Jofra's been out at Sussex's pre-season in India and bowled quickly out there. He bowled really well. He's now just gone back to the Caribbean, where he will play a little bit of club cricket, stuff like that. It's all about getting himself ready for that T20 World Cup. He'll hopefully play the Pakistan T20 series (in May), but it's all fingers crossed with Jofra at the moment.

Source:- https://www.thedailystar.net/sports/cricket/news/jofra-archer-targeting-t20-world-cup-3582941

Data Collection..

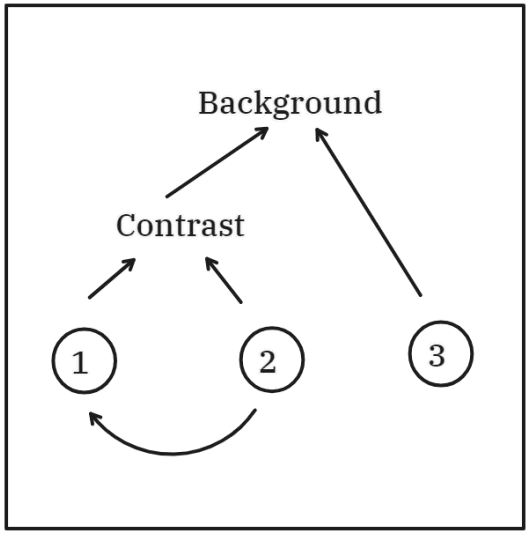
Sports Report (Cricket):-

Example report [0:1].sentence(0):-

England pace bowler Jofra Archer could play in this year's T20 World Cup but will not play test cricket until 2025, according to England managing director Rob Key.

RST study:-

Background(Contrast([England pace bowler Jofra Archer could play in this year's T20 World Cup but]$_1$, will not play test cricket until 2025]$_2$), [according to England managing director Rob Key.]$_3$)



RST Parsing Tree

LDSL
Linguistic Data Science Lab

Preprocessing..

To annotate sports reports for rhetorical study using Inception, we need to preprocess the dataset. Here are the steps that are necessary to follow:

1) **Clean the dataset** - removing extra spaces, special characters, or correct any formatting inconsistencies.

2) **Split Sentences-** Tokenize report into individual sentences. This can be done using natural language processing (NLP) libraries such as NLTK or SpaCy.

3) **Discourse Parsing**:- For our project, I will handle it manually, but we can also consider using tools such as SpaCy, the Discourse Parser from the Natural Language Toolkit (NLTK), or specialized tools like Rhetorical Structure Theory (RST) parsers for discourse parsing.

Fig-3 : Inception annotation tools

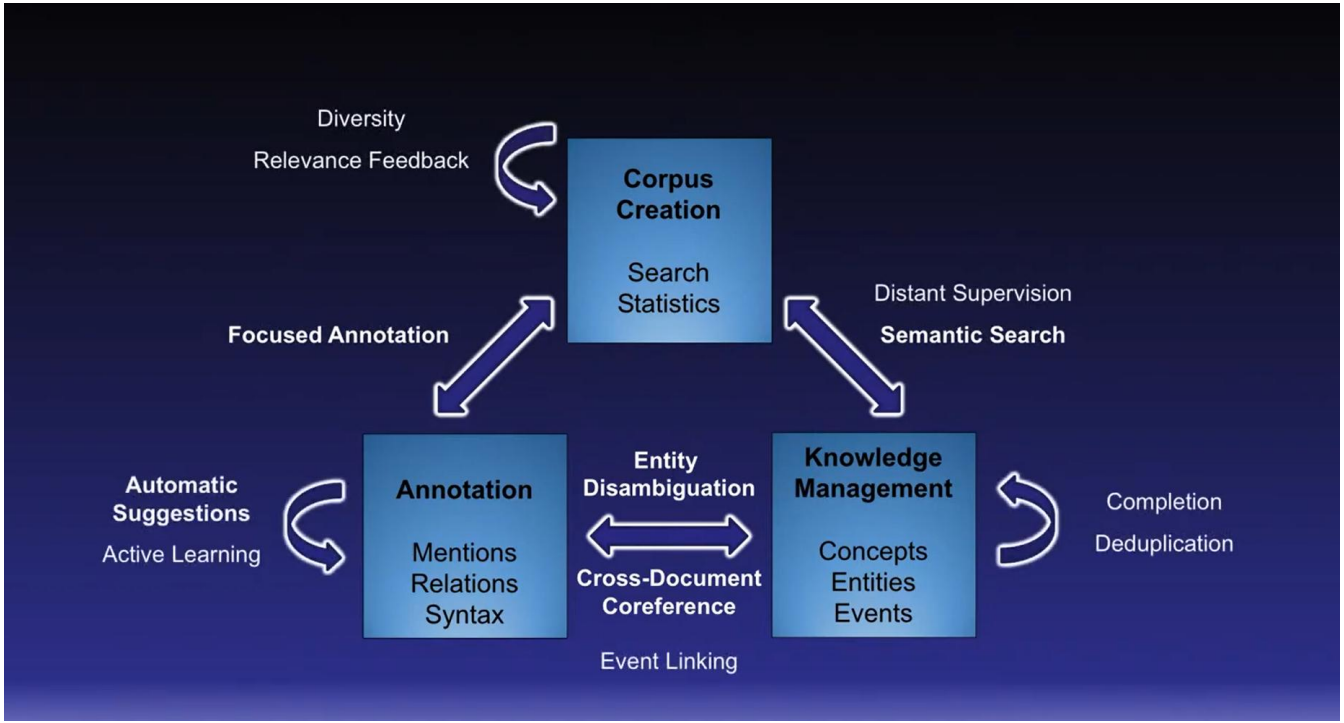
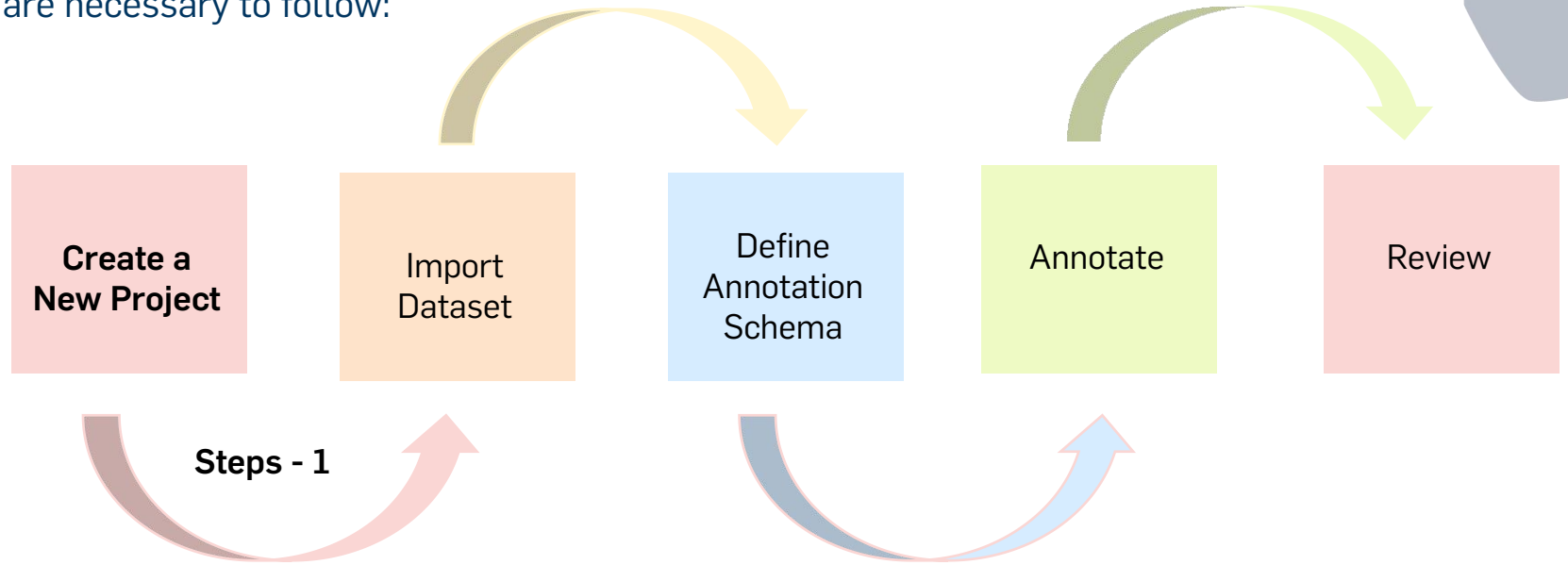


Fig - 4: Inception key features

Annotate RR With Inception tools..

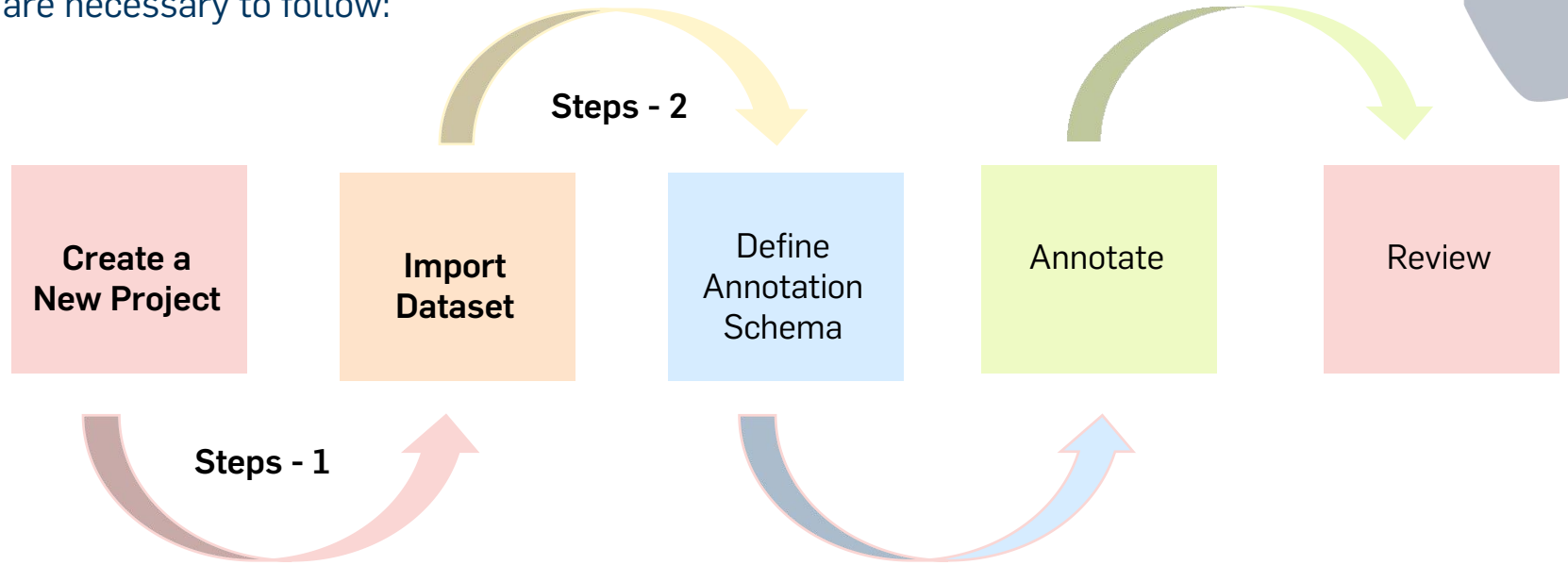
To annotate sports reports for rhetorical study using Inception tools, Here are the steps that are necessary to follow:



LDSL
Linguistic Data Science Lab

Annotate RR With Inception tools..

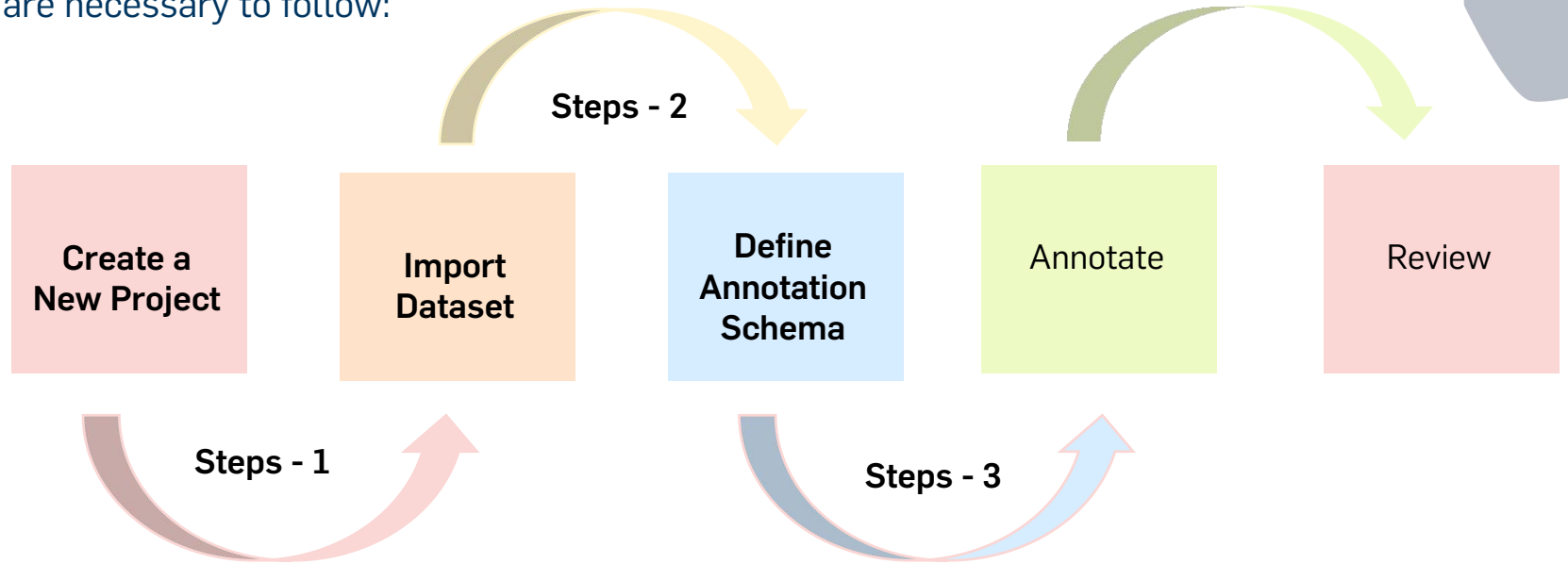
To annotate sports reports for rhetorical study using Inception tools, Here are the steps that are necessary to follow:



LDSL
Linguistic Data Science Lab

Annotate RR With Inception tools..

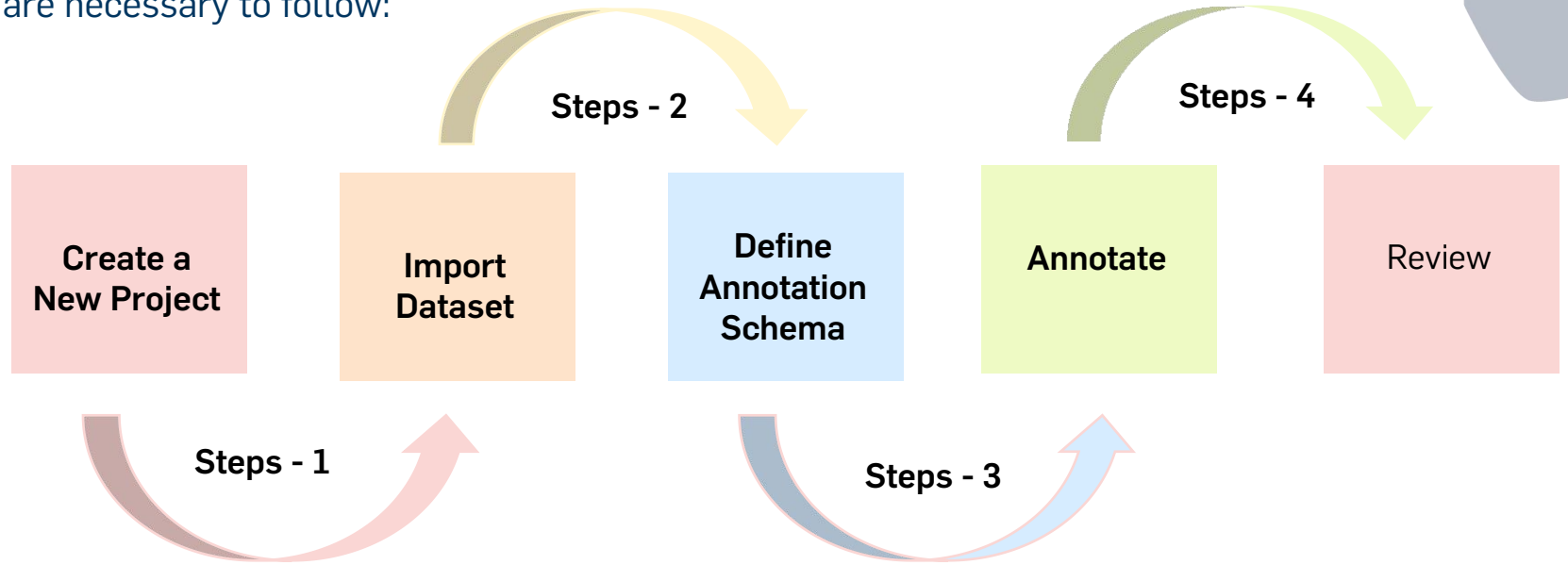
To annotate sports reports for rhetorical study using Inception tools, Here are the steps that are necessary to follow:



LDSL
Linguistic Data Science Lab

Annotate RR With Inception tools..

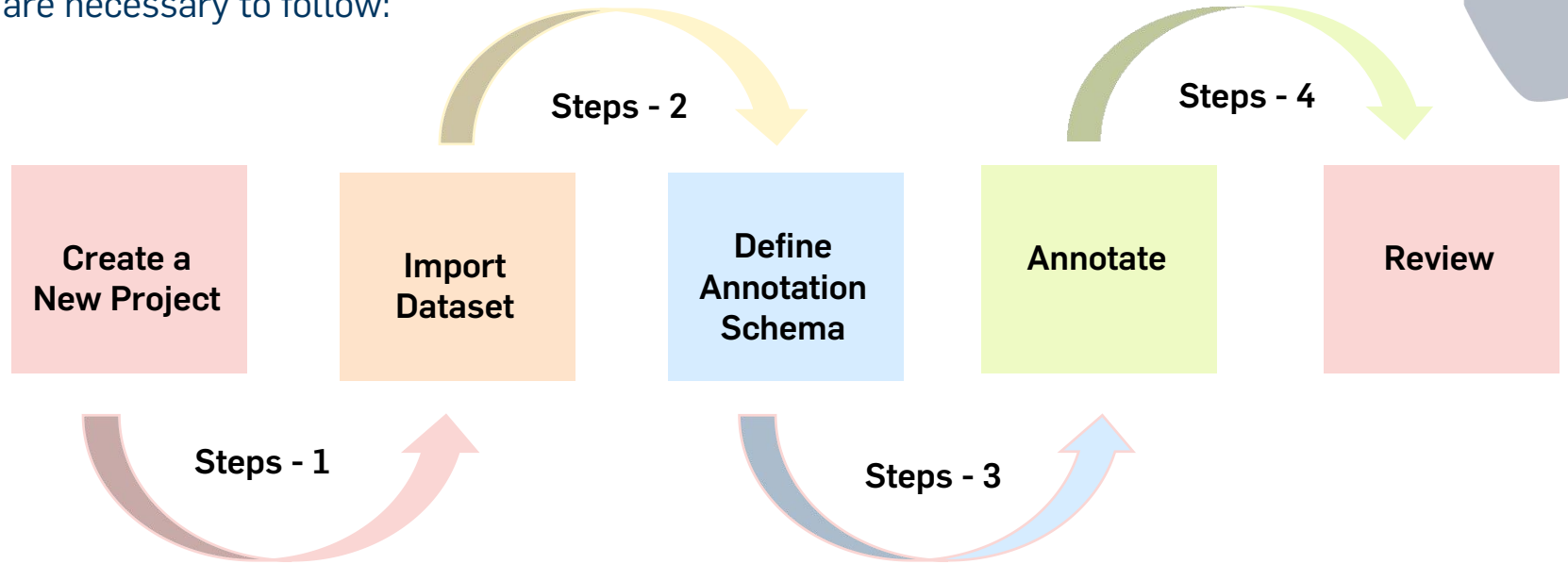
To annotate sports reports for rhetorical study using Inception tools, Here are the steps that are necessary to follow:



LDSL
Linguistic Data Science Lab

Annotate RR With Inception tools..

To annotate sports reports for rhetorical study using Inception tools, Here are the steps that are necessary to follow:



LDSL
Linguistic Data Science Lab

Analyse the results..

Based on the annotation results of our sports report, we can explore various analytical approaches to gain deeper insights and present comprehensive findings...

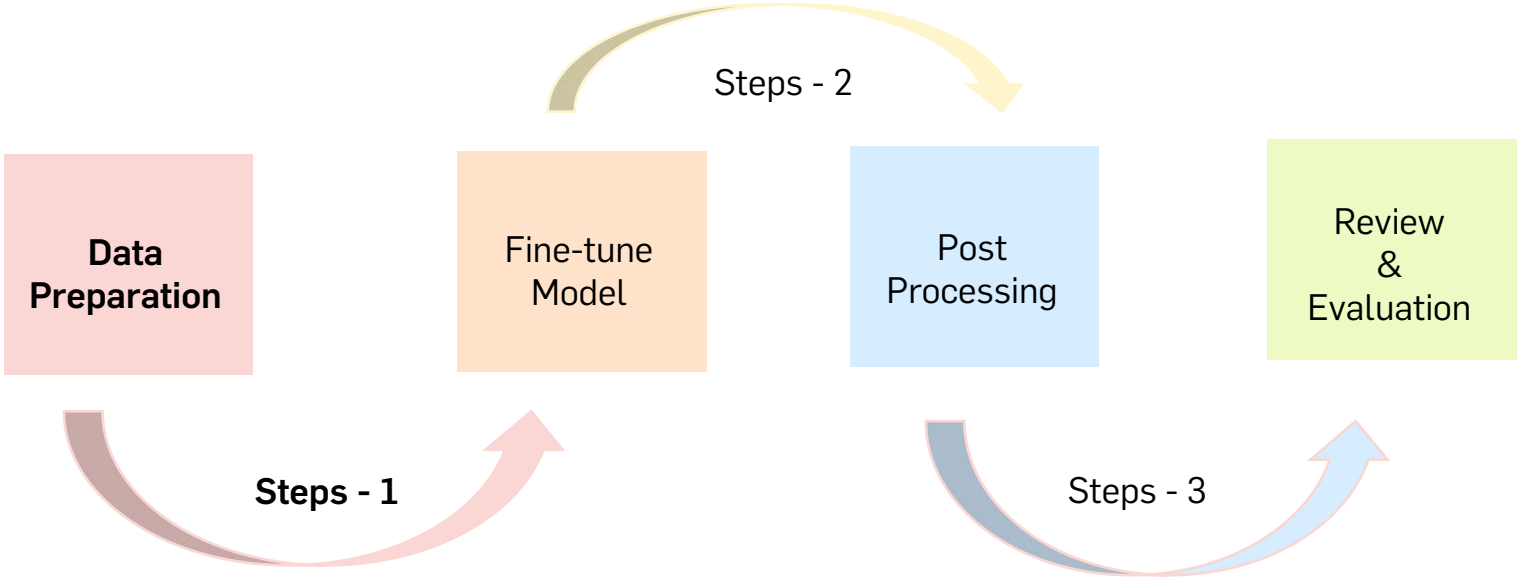
1) **Quantitative Analysis** - Frequency of each rhetorical relation, calculating proportions.

2) **Qualitative Analysis-** Identify Patterns and Themes and common contexts such as where "Contrast" or "Cause" relations appear.

3) **Visualizations**:- Represent our findings into visualizations such as frequency of RR into a bar chart and proportions into pie chart.

Realizing Automatic Discourse Parsing (Hu and Wan, 2023)

Some steps that are relevent to realize automatic discourse parsing according to Hu and Wan (2023): RST discourse parsing as text-to-text generation:-



LDSL
Linguistic Data Science Lab

Data Prepration..

Linearization:

The linearization is carried out from the bottom up according to **postorder traversal**.

[[Goverment leading was not intended to be a way] Nucleus span [[to obfuscate spending figures,] Nucleus joint [[hide fraudulent activity] Nucleus joint [[or provide large subsidies] Nucleus joint] Nucleus joint] Satellite Elaboration]

EDU1: Goverment leading was not intended to be a way

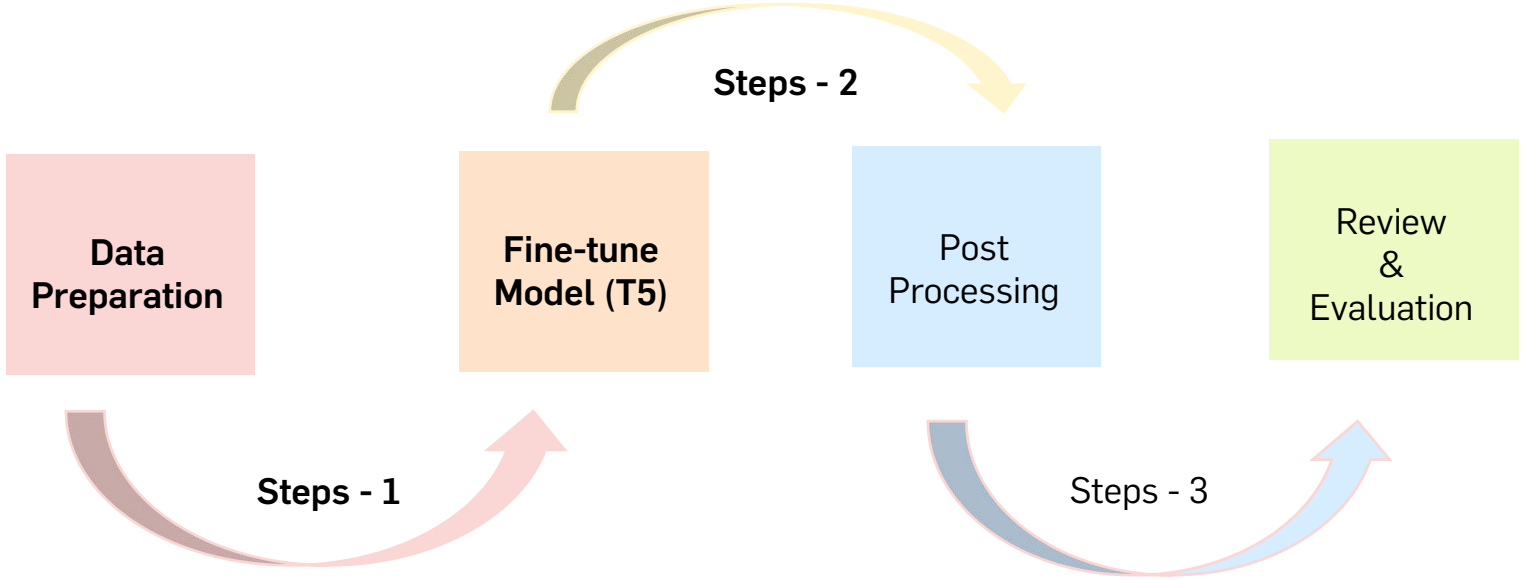
EDU2: to obfuscate spending figures,

EDU3: hide fraudulent activity

EDU4: or provide large subsidies

Realizing Automatic Discourse Parsing (Hu and Wan, 2023)

Some steps that are relevent to realize automatic discourse parsing according to Hu, X. and Wan, X. (2023): RST discourse parsing as text-to-text generation:-



LDSL
Linguistic Data Science Lab

Fine-tune Model (T5)..

Seq2Seq Training:

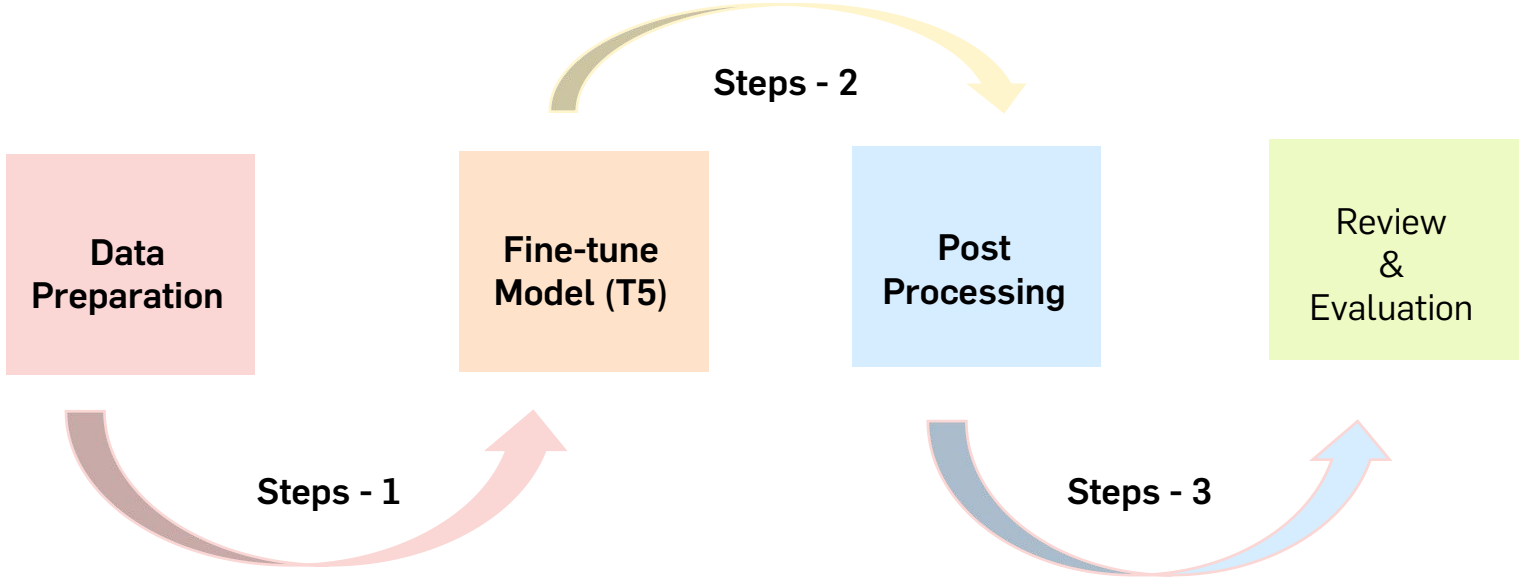
As the input and new output of the task are both sequences, RST parsing can thus be trained or finetuned on any generation model as a text-to-text generation task. Mathematically, the model calculates the **conditional probability** of a target output sequence and finds the most probable sequence.

Constrained Decoding:

A seq2seq model normally generates the target output token-by-token according to the probability distribution and creates **format errors**, To address this problem, they introduce a constrained decoding method. it dynamically modifies the candidate vocabulary set in beam search according to the current generated sequence and follows the designed formats.

Realizing Automatic Discourse Parsing (Hu and Wan, 2023)

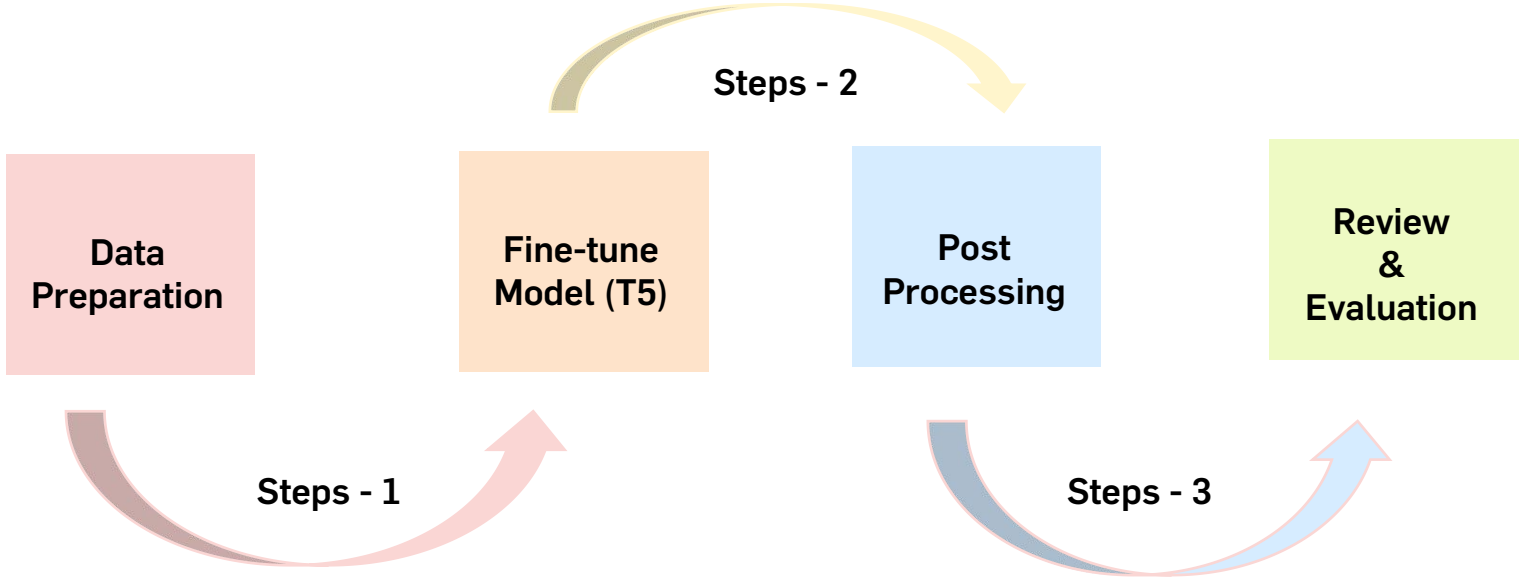
Some steps that are relevent to realize automatic discourse parsing according to Hu, X. and Wan, X. (2023): RST discourse parsing as text-to-text generation:-



LDSL
Linguistic Data Science Lab

Realizing Automatic Discourse Parsing (Hu and Wan, 2023)

Some steps that are relevent to realize automatic discourse parsing according to Hu, X. and Wan, X. (2023): RST discourse parsing as text-to-text generation:-



LDSL
Linguistic Data Science Lab

Review & Evaluation..

Evaluation metrics:

F1-scores of unlabeled (Span) and labeled (Nuclearity and Relation). They use 18 rhetorical relation labels.

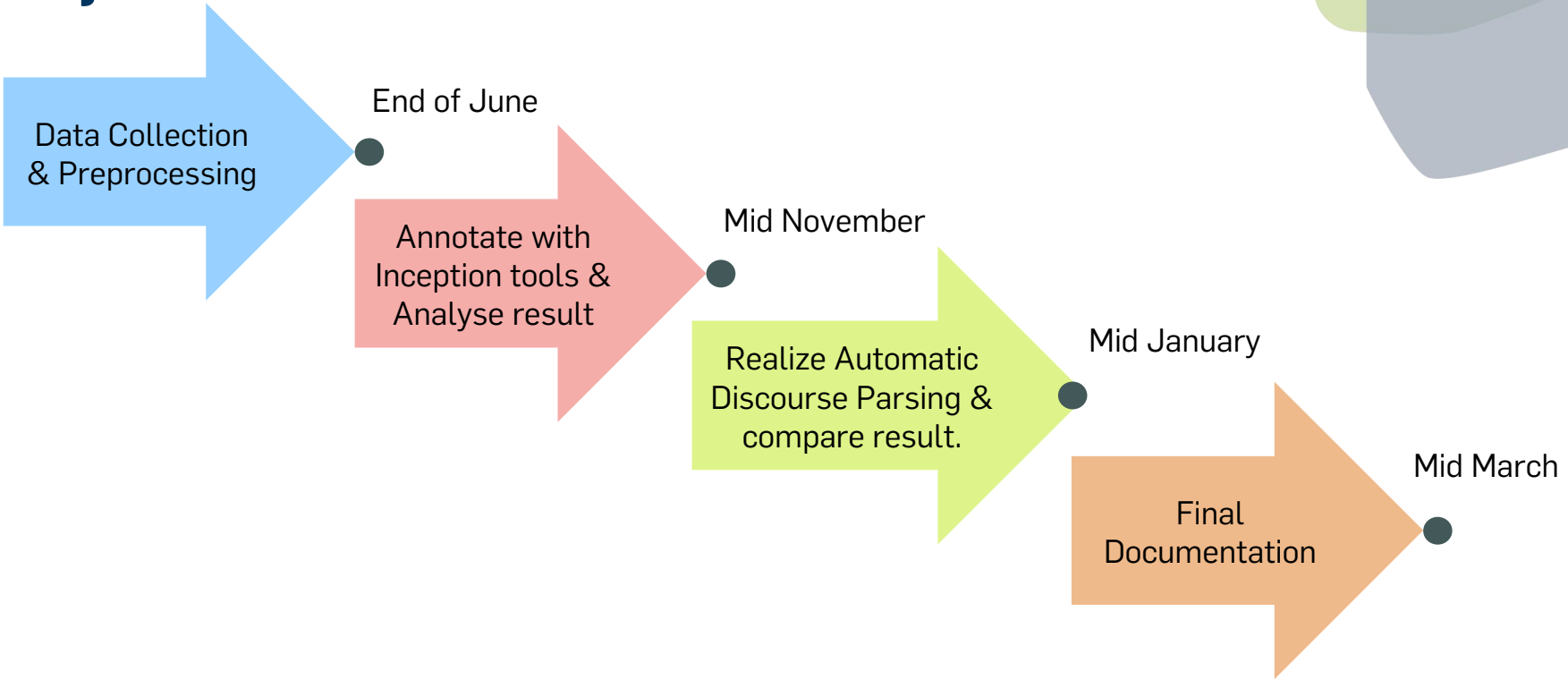
Conclusion

- Expected Challenges
- Project Timeline
- Reference

Expected Challenges..

- **Data collection for sports report (cricket)**
- **Inception tools**
  - choose the right relations for our dataset
  - define annotation schema
- **Automatic Discourse Parsing (Hu and Wan, 2023)**
  - Linearization
  - Finetuning of the T5 models

Project Timeline..



LDSL
Linguistic Data Science Lab

Reference..

MANN, WILLIAM & Thompson, Sandra. (1988). Rethorical Structure Theory: Toward a functional theory of text organization. Text. 8. 243-281. 10.1515/text.1.1988.8.3.243.

Hu, Xinyu & Wan, Xiaojun. (2023). RST Discourse Parsing as Text-to-Text Generation. IEEE/ACM Transactions on Audio, Speech, and Language Processing. PP. 1-12. 10.1109/TASLP.2023.3306710.

Stede, M. (2017). Annotation Guidelines for Rhetorical Structure.

Jasinskaja, Katja & Karagjosova, Elena. (2020). Rhetorical Relations. 10.1002/9781118788516.sem061.

Klie, Jan-Christoph & Bugert, Michael & Boullosa, Beto & Eckart de Castilho, Richard & Gurevych, Iryna. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation.

Research Project - 1

Annotating Rhetorical Relations for Discourse parsing

Thank you for your attention

Mehedi Hasan Emon