Data Mining for NLP

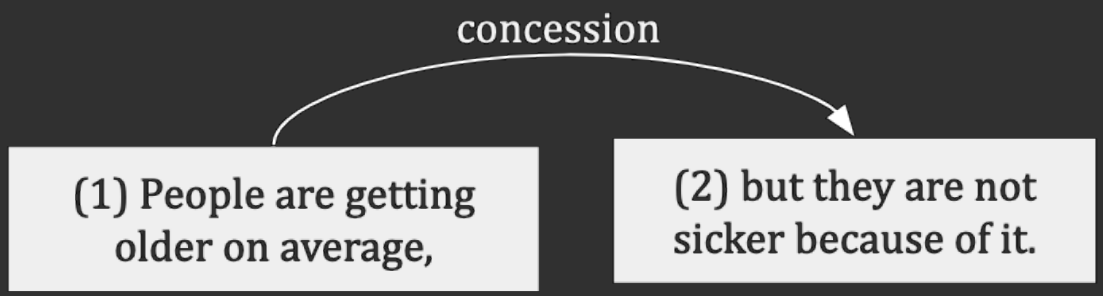
Laurine Huber

LORIA, Université de Lorraine

January 20, 2022

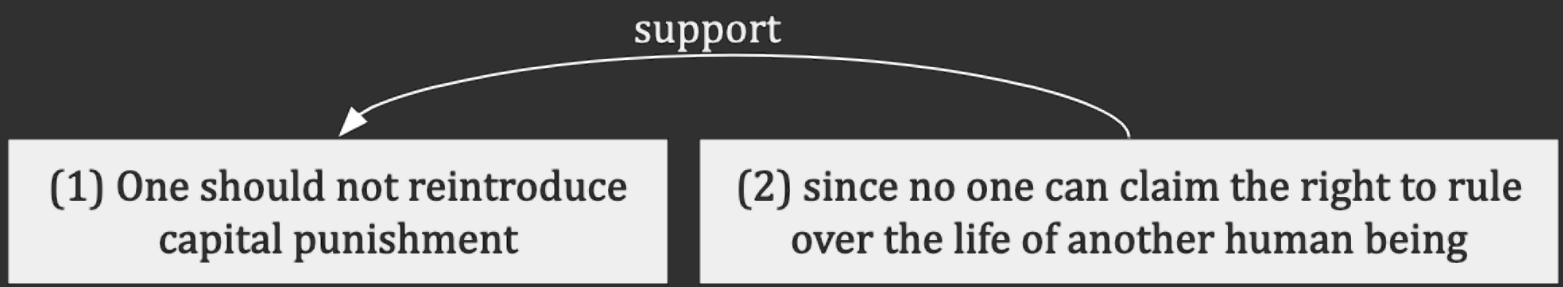
Discourse structure

- Semantic and pragmatic relations between text segments (*reason*, *cause*, *concession* ...)
- Rhetorical Structure Theory [Mann and Thompson, 1988]
- Distinction between nucleus and satellite



Argumentation Structure

- Argumentation relations between text segments (*support*, *attack*, ...)
- Macro-structure of argumentation [Freeman, 2011]
  - Dialogical exchange between a proponent and an opponent
  - Distinction between premise and conclusion



Study a corpus of argumentative texts

Goal: Understand the similarities between discourse and argumentation structures.

- Descriptive: understand linguistic differences between argumentation and discourse structures
- Normative: build bridges between theories; unify annotations

Corpus

- ArgMicroTexts corpus [Peldszus and Stede, 2015] *
- 112 short argumentative texts
- 18 controversial questions

"Should Germany introduce the death penalty?"

1: The death penalty is a legal means that as such is not practicable in Germany.
2: For one thing, inviolable human dignity is anchored in our constitution,
3: and furthermore no one may have the right to adjudicate upon the death of another human being.
4: Even if many people think that a murderer has already decided on the life or death of another person,
5: this is precisely the crime that we should not repay with the same.

* *available online*

Corpus

- ArgMicroTexts corpus [Peldszus and Stede, 2015] *
- 112 short argumentative texts
- 18 controversial questions

"Should Germany introduce the death penalty?"

1: The death penalty is a legal means that as such is not practicable in Germany.
2: For one thing, inviolable human dignity is anchored in our constitution,
3: and furthermore no one may have the right to adjudicate upon the death of another human being.
4: Even if many people think that a murderer has already decided on the life or death of another person,
5: this is precisely the crime that we should not repay with the same.

* *available online*

Corpus

- ArgMicroTexts corpus [Peldszus and Stede, 2015] *
- 112 short argumentative texts
- 18 controversial questions

"Should Germany introduce the death penalty?"

1: The death penalty is a legal means that as such is not practicable in Germany.
2: For one thing, inviolable human dignity is anchored in our constitution,
3: and furthermore no one may have the right to adjudicate upon the death of another human being.
4: Even if many people think that a murderer has already decided on the life or death of another person,
5: this is precisely the crime that we should not repay with the same.

* *available online*

Corpus

- ArgMicroTexts corpus [Peldszus and Stede, 2015] *
- 112 short argumentative texts
- 18 controversial questions

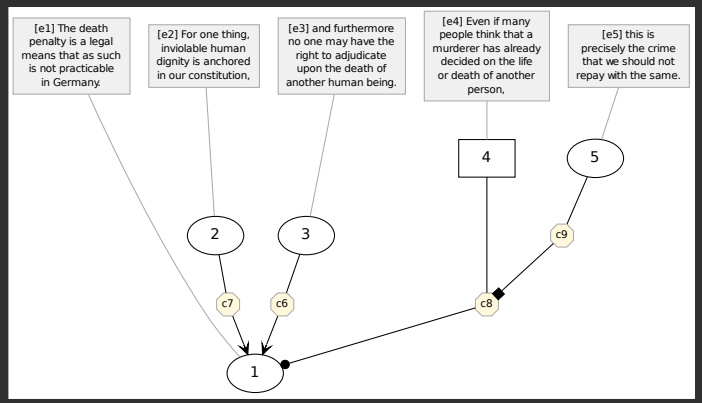
"Should Germany introduce the death penalty?"

1: The death penalty is a legal means that as such is not practicable in Germany.
2: For one thing, inviolable human dignity is anchored in our constitution,
3: and furthermore no one may have the right to adjudicate upon the death of another human being.
4: Even if many people think that a murderer has already decided on the life or death of another person,
5: this is precisely the crime that we should not repay with the same.

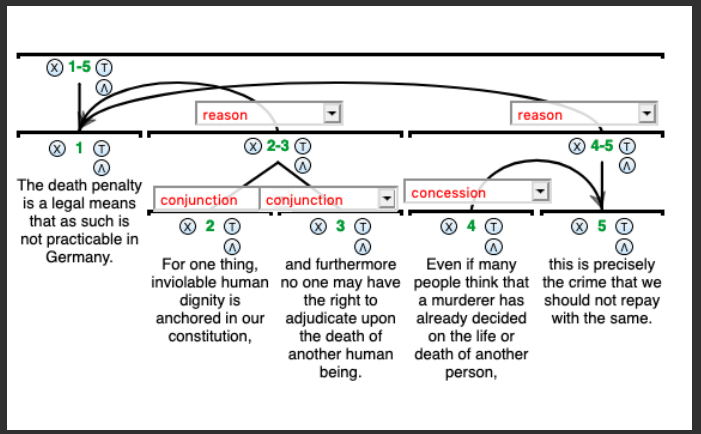
* *available online*

Corpus

- Macro-structure of argumentation [Peldszus and Stede, 2016]
- RST
- (SDRT [Lascarides and Asher, 2007])



(a) ARG annotation



(b) RST annotation

Overview of the approach

Goal: can we align ARG and RST at the subtree level ?

1. Representing ARG and RST structures as trees
2. Building two descriptions of each text
   - ARG and RST descriptions
   - A description is a set of subtrees
3. Aligning set of subtrees that describe almost the same set of texts

Representing ARG and RST structures as trees

Goal: Unify and anonymise the structures.

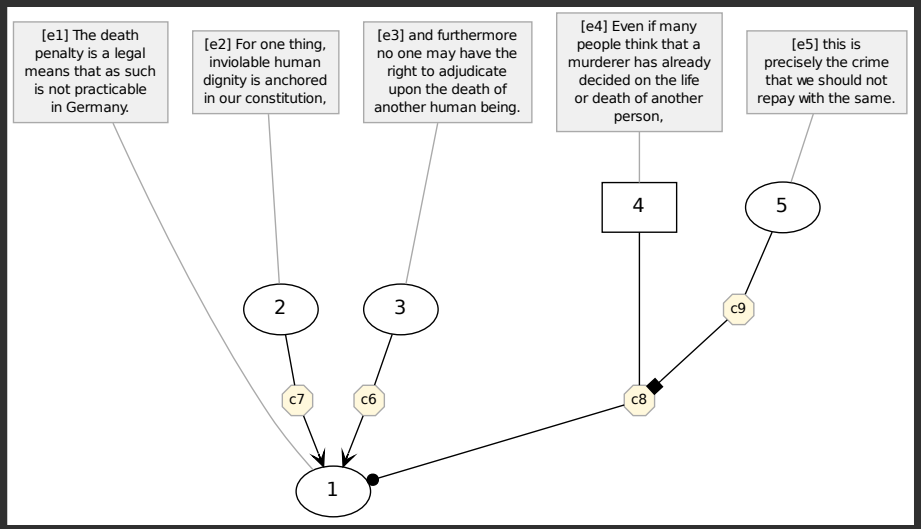
- Transform *ARG* and *RST* structures into labeled trees
- Keep only structure, no text

Representing ARG and RST structures as trees

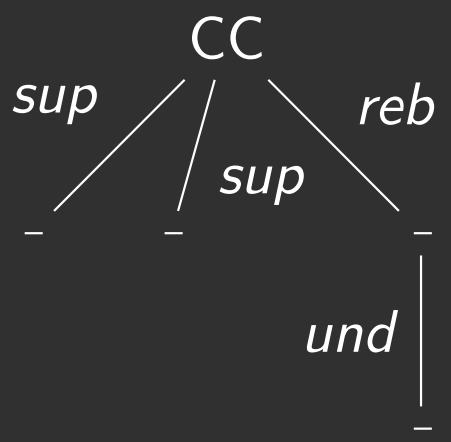
Goal: Unify and anonymise the structures.

- Transform *ARG* and *RST* structures into labeled trees
- Keep only structure, no text

Representing ARG and RST structures as trees : *ARG*



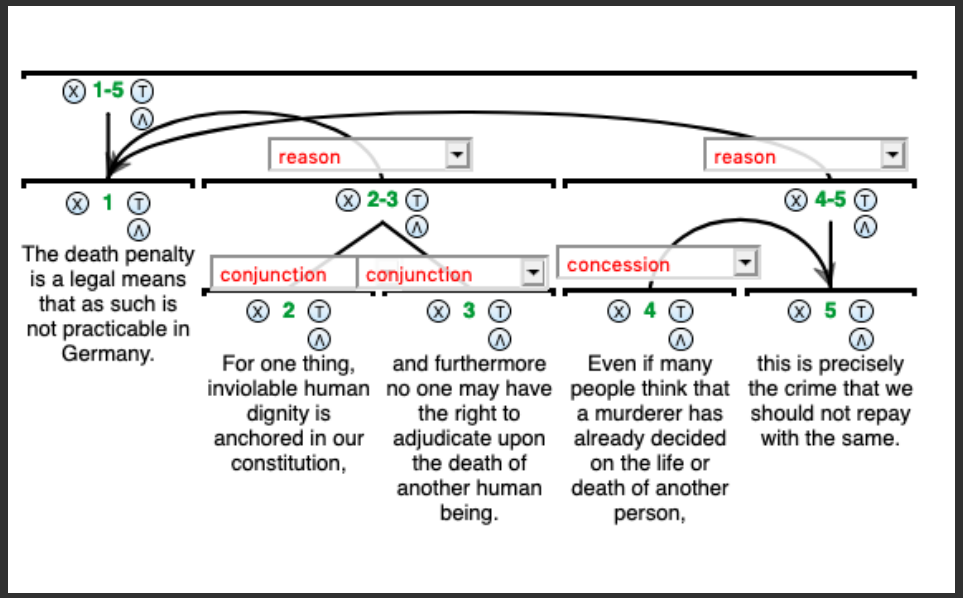
ARG annotation



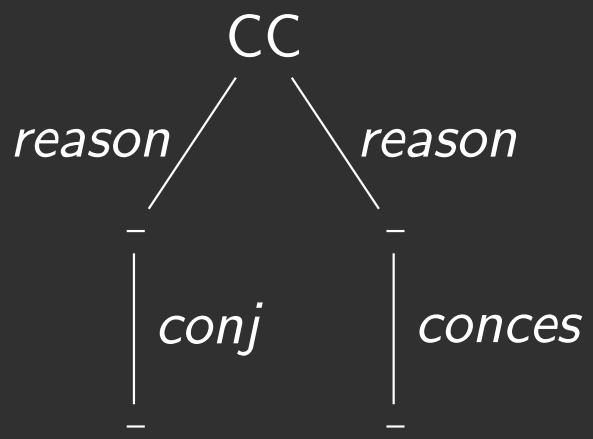
ARG tree derivation

- **Root**: central claim
- **Parent**: conclusion
- **Child**: premisse

Representing ARG and RST structures as trees : *RST*



RST annotation



RST tree derivation

- **Root**: most central nucleus
- **Parent**: nucleus
- **Child**: satellite

Building two descriptions of the corpus

Goal: Produce 2 descriptions of each texts in term of subtrees

1. Extract all subtrees of ARG
2. Extract all subtrees of RST

Frequent subgraph mining: gSpan [Yan and Han, 2002]

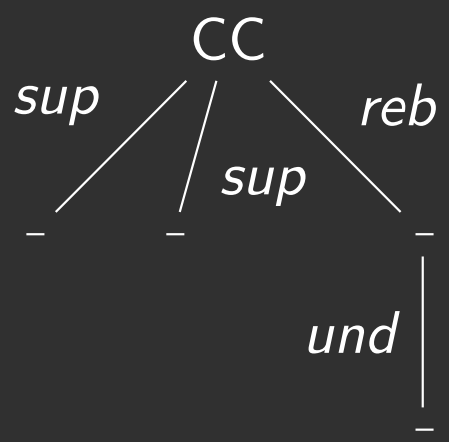
Building two descriptions of the corpus

Goal: Produce 2 descriptions of each texts in term of subtrees

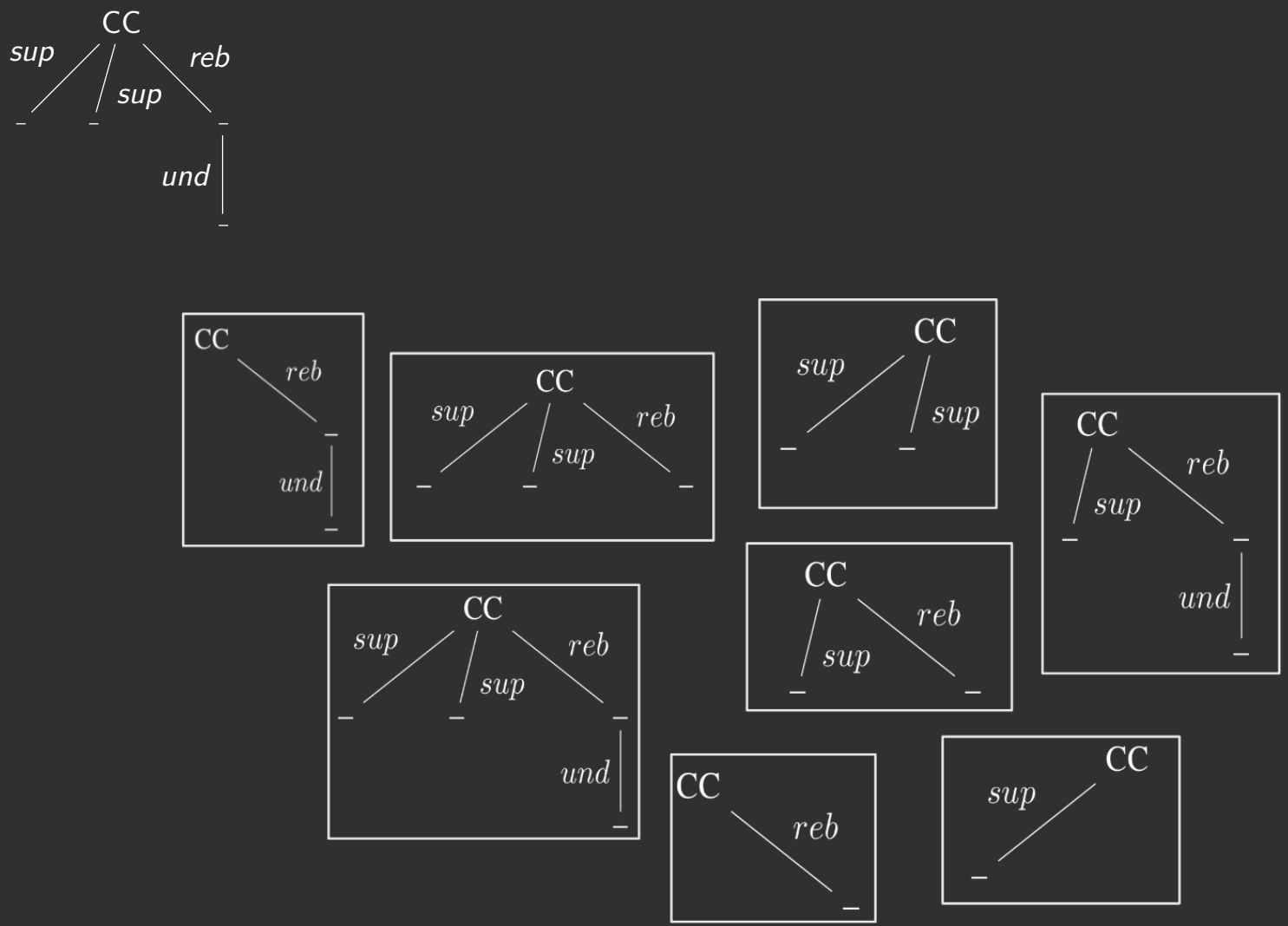
1. Extract all subtrees of ARG
2. Extract all subtrees of RST

Frequent subgraph mining: gSpan [Yan and Han, 2002]

Building two descriptions of the corpus: subtrees extraction

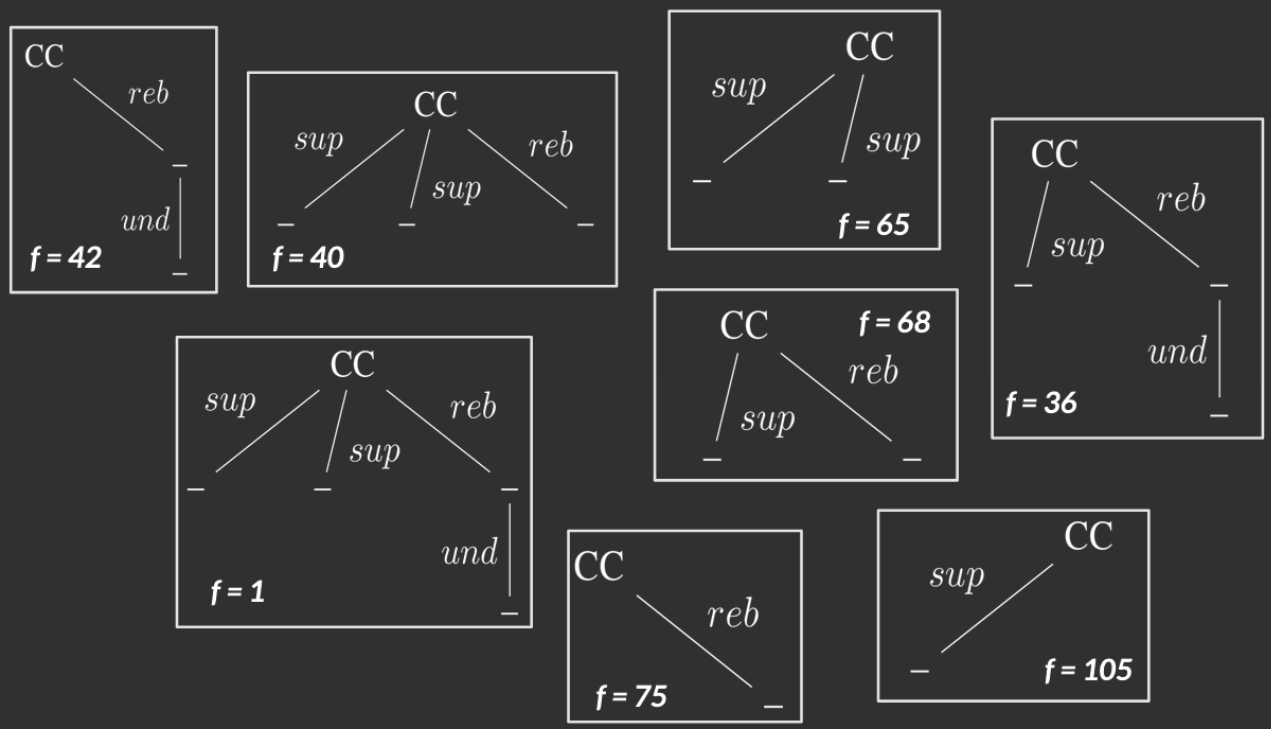


Building two descriptions of the corpus: subtrees extraction



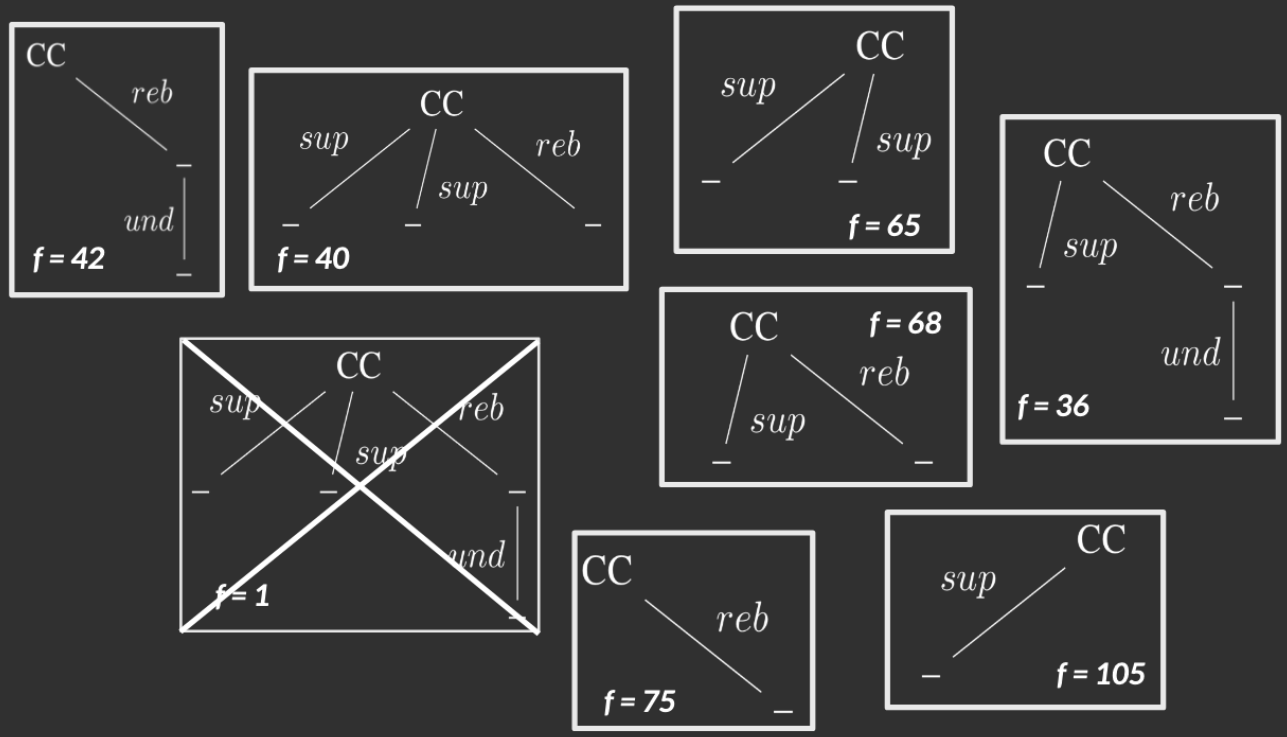
Building two descriptions of the corpus: subtrees extraction

- f is the frequency of occurrence of subtrees in the corpus



Building two descriptions of the corpus: subtrees extraction

- keep subtrees with $f \geq 2$



Redescription mining

Goal: Find an ARG description and a RST description that characterize almost the same set of objects

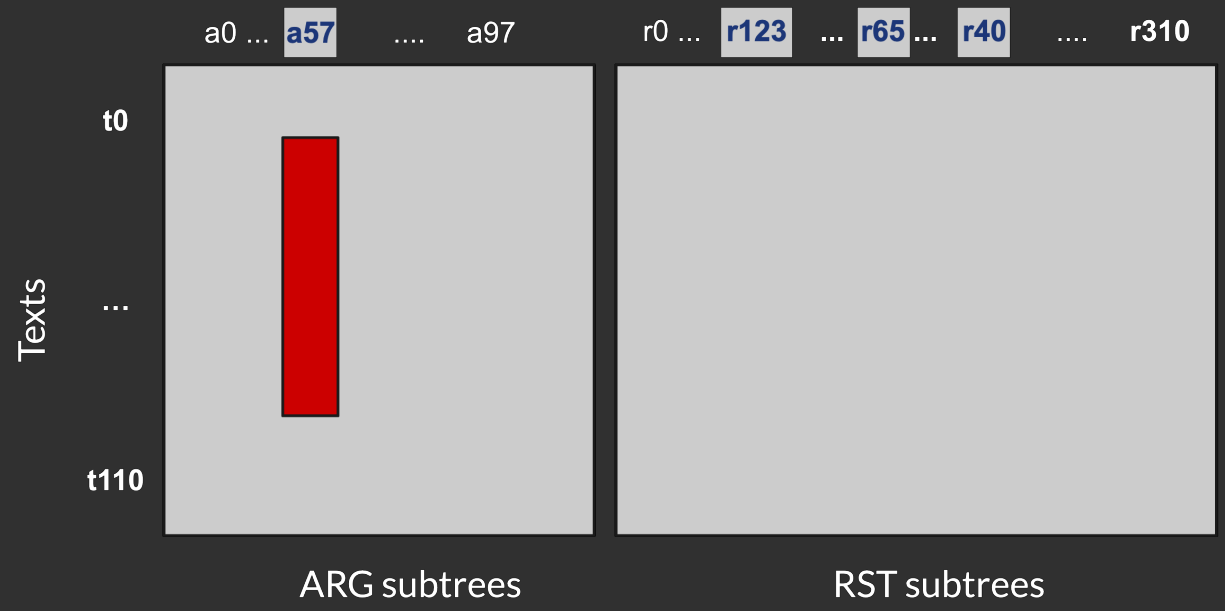
- Two different descriptions of the each text
  - $ARG = \{a0, a1, ..., a98\}$
  - $RST = \{r0, r1, ..., r311\}$
- A set of objects: a set of texts from the corpus
- A text t_i is described by
  - a subset of ARG
  - a subset of RST

Redescription mining

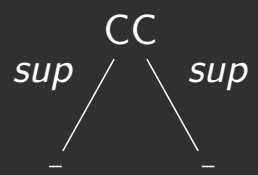
Goal: Find an ARG description and a RST description that characterize almost the same set of objects

- Two different descriptions of the each text
  - $ARG = \{a0, a1, ..., a98\}$
  - $RST = \{r0, r1, ..., r311\}$
- A set of objects: a set of texts from the corpus
- A text t_i is described by
  - a subset of ARG
  - a subset of RST

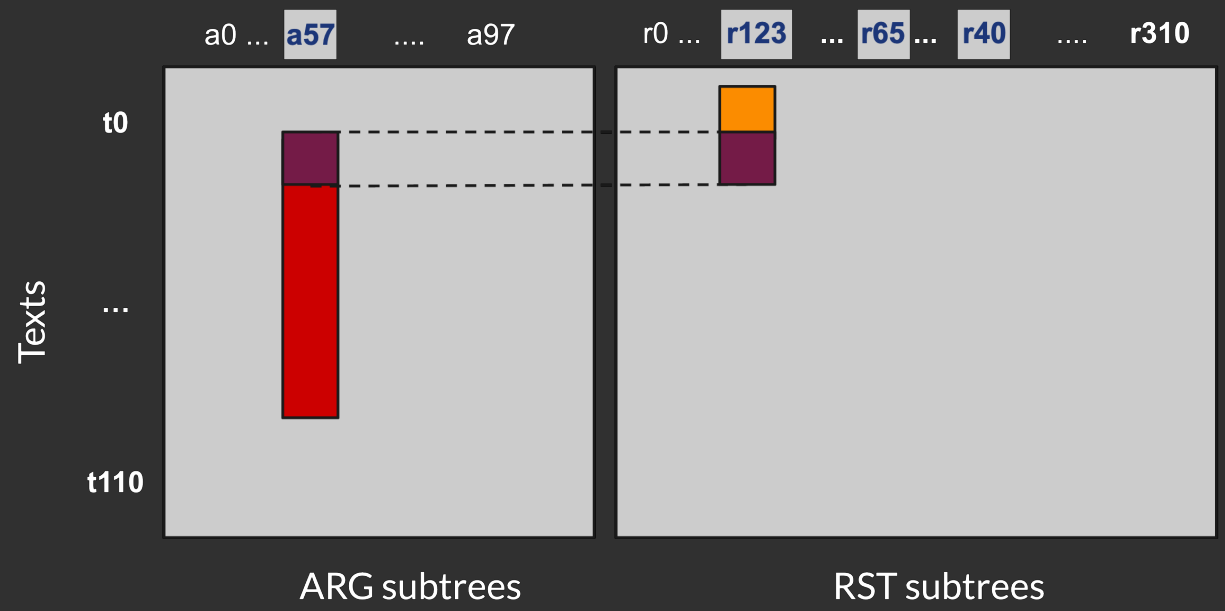
Redescription mining



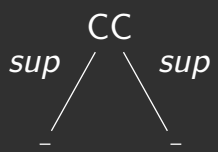
$$Rd1 : a57 \leftrightarrow \varnothing$$



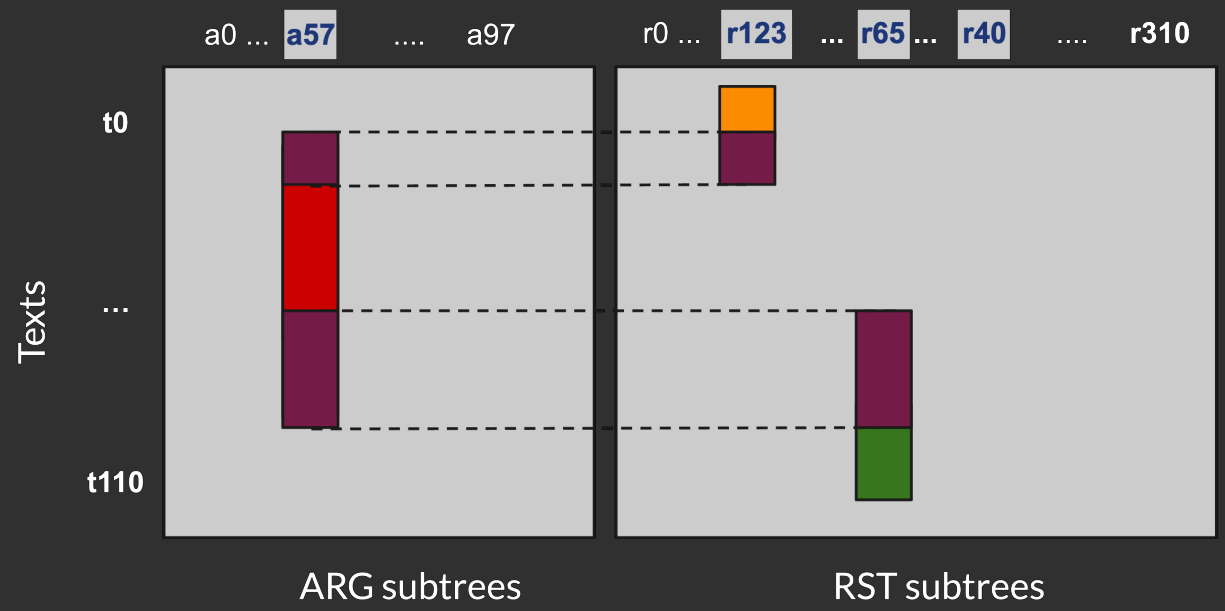
Redescription mining



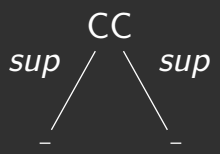
$$Rd1 : a57 \longleftrightarrow r123$$

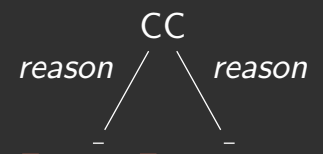


Redescription mining

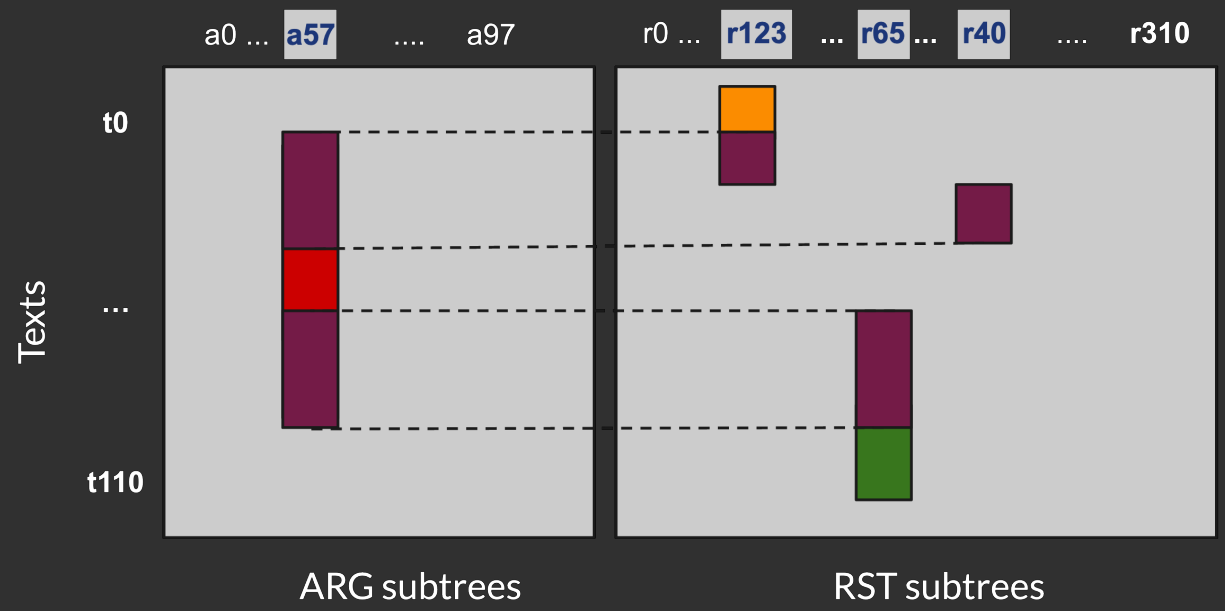


$$Rd1 : a57 \longleftrightarrow r123 \vee r65$$

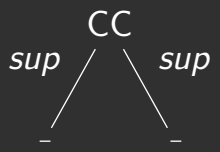


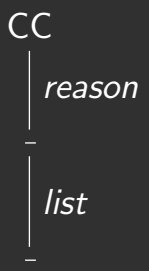


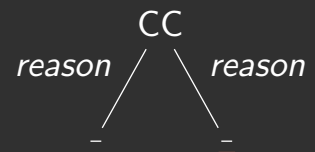
Redescription mining

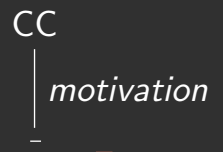


$$Rd1 : a57 \longleftrightarrow r123 \vee r65 \vee r40$$









Redescription mining

- A redescription is pair of queries
  - *qArg* a logical formulae over the *Arg* subtrees
  - *qRst* a logical formulae over the *Rst* subtrees
- *qArg* and *qRst* should describe **almost** the same set of texts
- *"Almost"*: given a similarity threshold calculated with Jaccard index

$$Jacc(qArg, qRst) = \frac{supp(qArg \wedge qRst)}{supp(qArg \vee qRst)}$$

Experiment setup

- Algorithm: ReRemi
- Conjunctions and disjunctions allowed
- Length of the query limited to 4
- Output: 35 redescriptions

Results

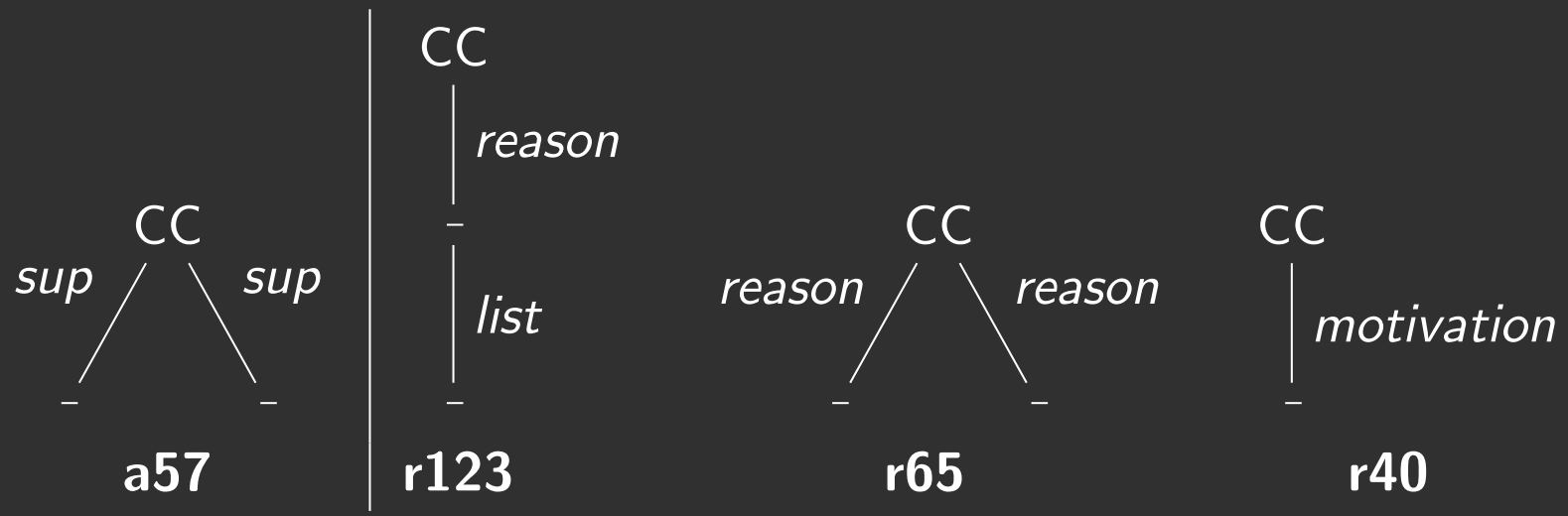
id	q1	q2	J(q1,q2)	# texts
*Rd*1	a57	r123 ∨ r65 ∨ r40	0.691	54
*Rd*2	a58	r61 ∨ r119 ∨ r125	0.351	13
*Rd*3	a23 ∨ a59	r125	0.3	8

3 over 35 obtained redescriptions

aX and rX correspond to *ARG* and *RST* subtrees respectively.

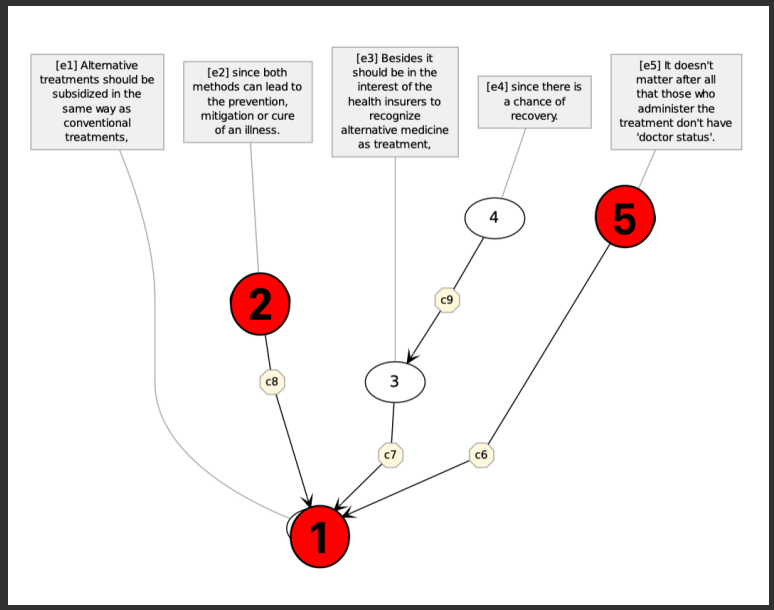
Results

$$Rd1 : a57 \longleftrightarrow r123 \vee r65 \vee r40$$

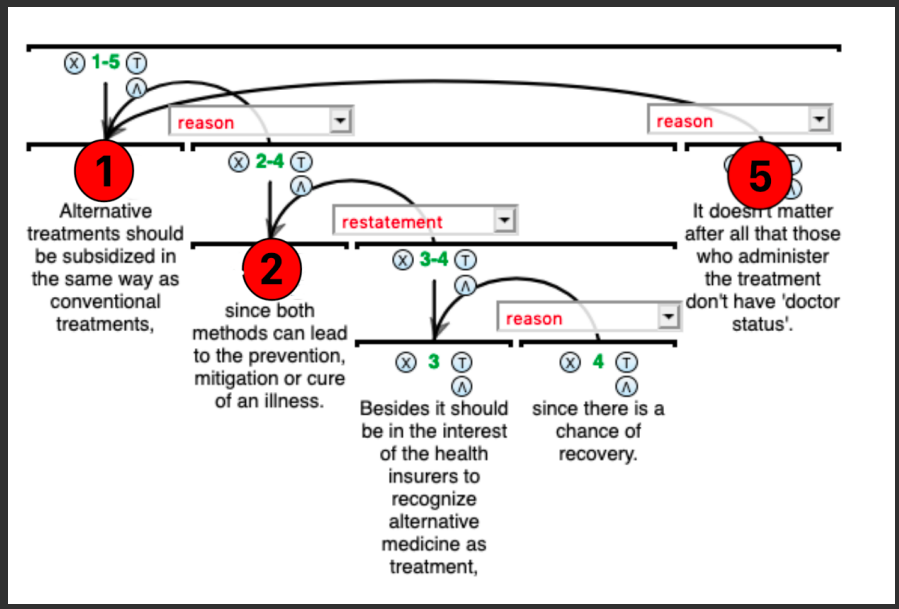


RST is more fine grained than ARG

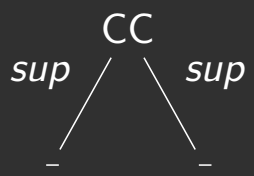
Well captured information

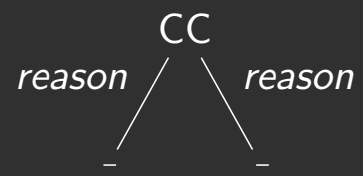


(a) ARG annotation

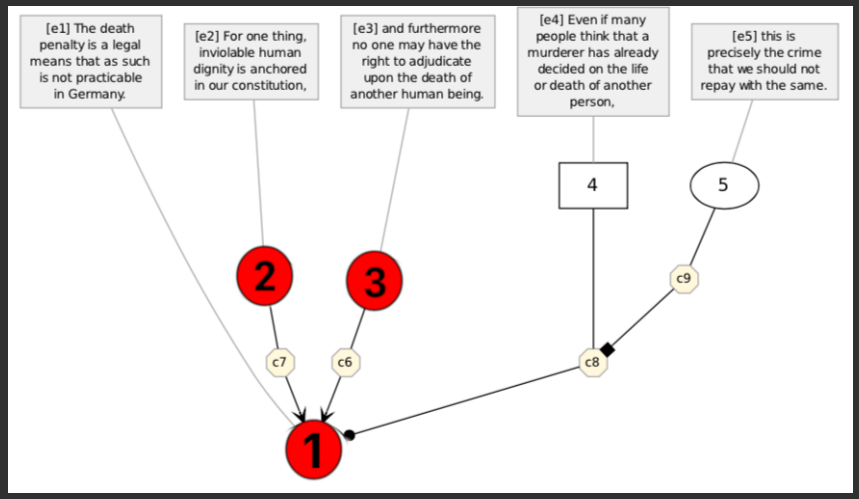


(b) *RST* annotation

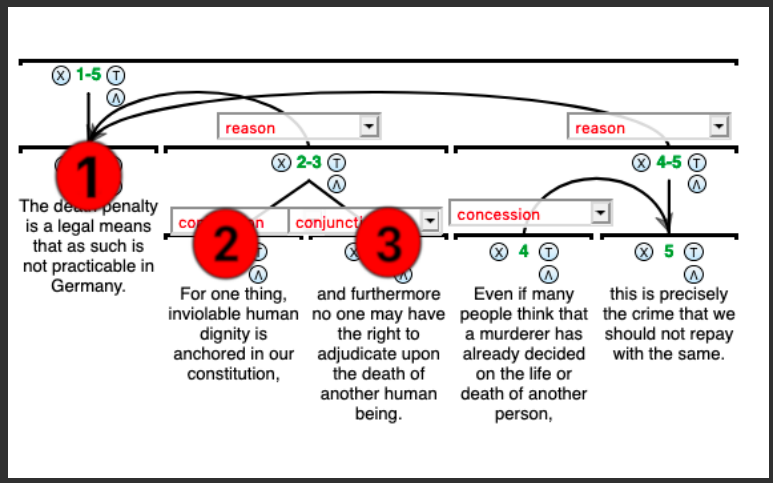




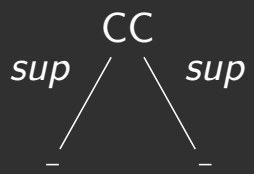
Anonymization lead to wrong captured patterns

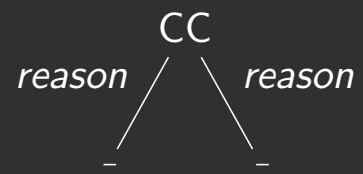


(a) *ARG* annotation



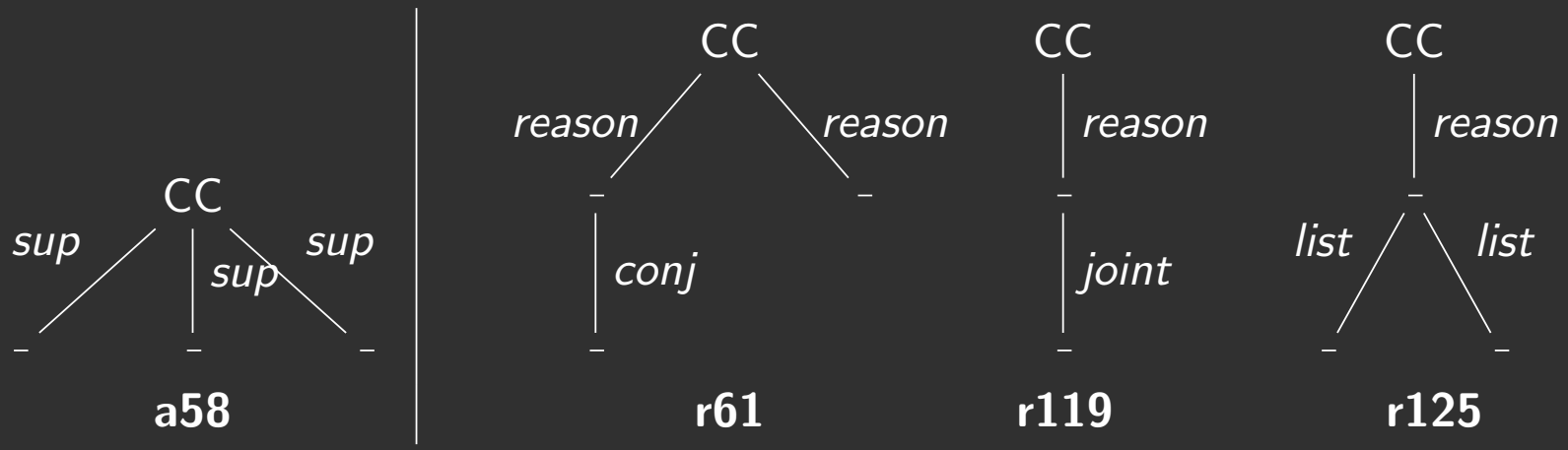
(b) *RST* annotation





Results

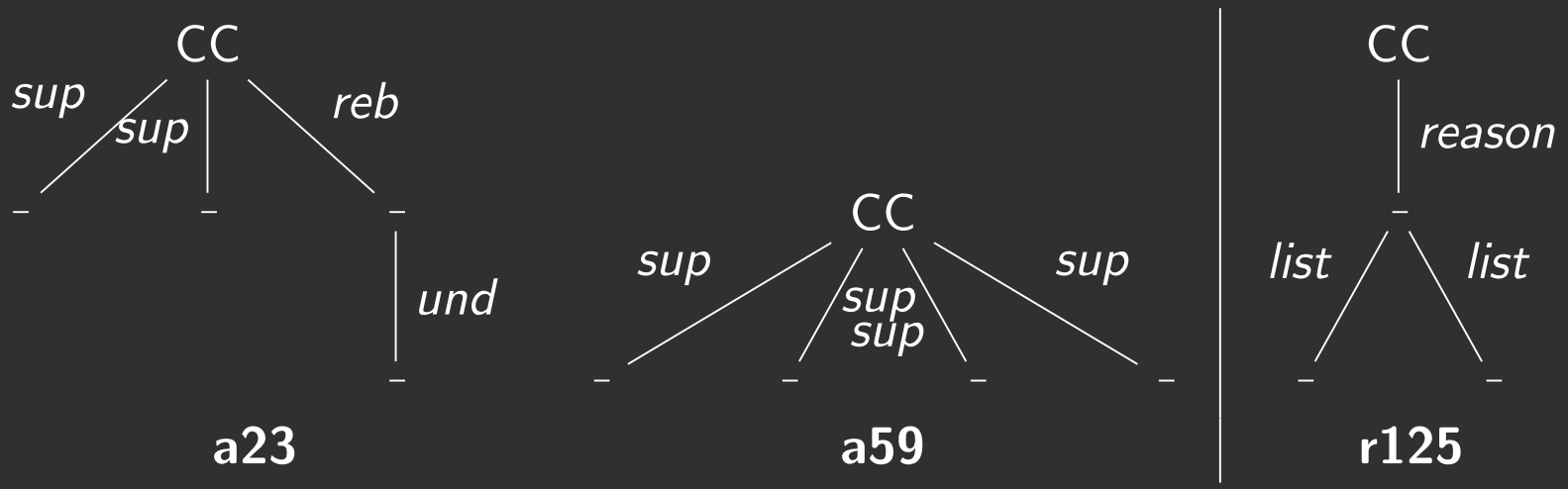
$$Rd2 : a58 \longleftrightarrow r61 \vee r119 \vee 125$$



Rd2 is a specialization of Rd1

Results

$$Rd3 : a23 \vee a59 \longleftrightarrow r125$$



2 $\neq$ ARG representations of the one RST subtree

Conclusion

- Turn a linguistic problem into a Data Mining problem
- Systematic, generic and automatic comparison
- Understand the links between $\neq$ theories

Joint work with *Yannick Toussaint, Charlotte Roze, Mathilde Dargnat and Chloé Braud*, presented at ArgMining 2019.

Other interesting questions ?

- Can we find argumentative patterns specific to arguments that are in favor of or in opposition to a stance.
- Can we use data mining on argumentative patterns to classify between pro and cons arguments.

For and against arguments

Should shopping malls generally be allowed to open on holidays and Sundays? $\longrightarrow$ NO

1. Supermarket employees and people who work in shopping centres also have the right to a Sunday off work.
2. Likewise public holidays should remain what they are: for some a day of introspection, for others a paid day off that is not taken away from the annual paid leave proper.
3. Hence it is good when shops are not open on Sundays and public holidays.
4. People, however, who work during the week and on Saturdays then have a problem: everyone else can shop weekdays, but they can't.
5. For those people the late opening hours, which meanwhile already extend to 12:00 midnight, present a good alternative.

For and against arguments

Should shopping malls generally be allowed to open on holidays and Sundays? $\longrightarrow$ YES

1. Well, I as an employee find it very practical to be able to shop at least on weekends.
2. Sure, other people have to work in the shops on the weekend,
3. but they can have days off during the week and run errands at their leisure while I'm stuck in the office.
4. Plus, the state wants me to spend my money,
5. and how am I supposed to do that when the shops aren't open when I'm off work?

For and against arguments

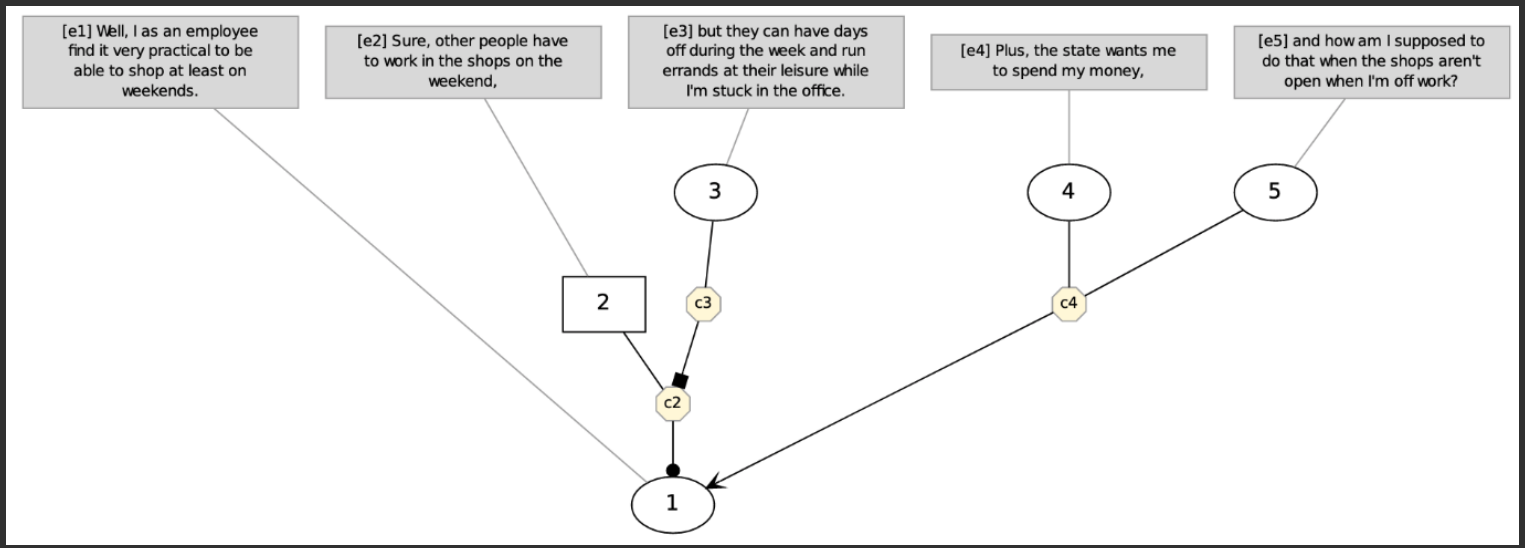


Figure: Arg annotation of CON argument

For and against arguments

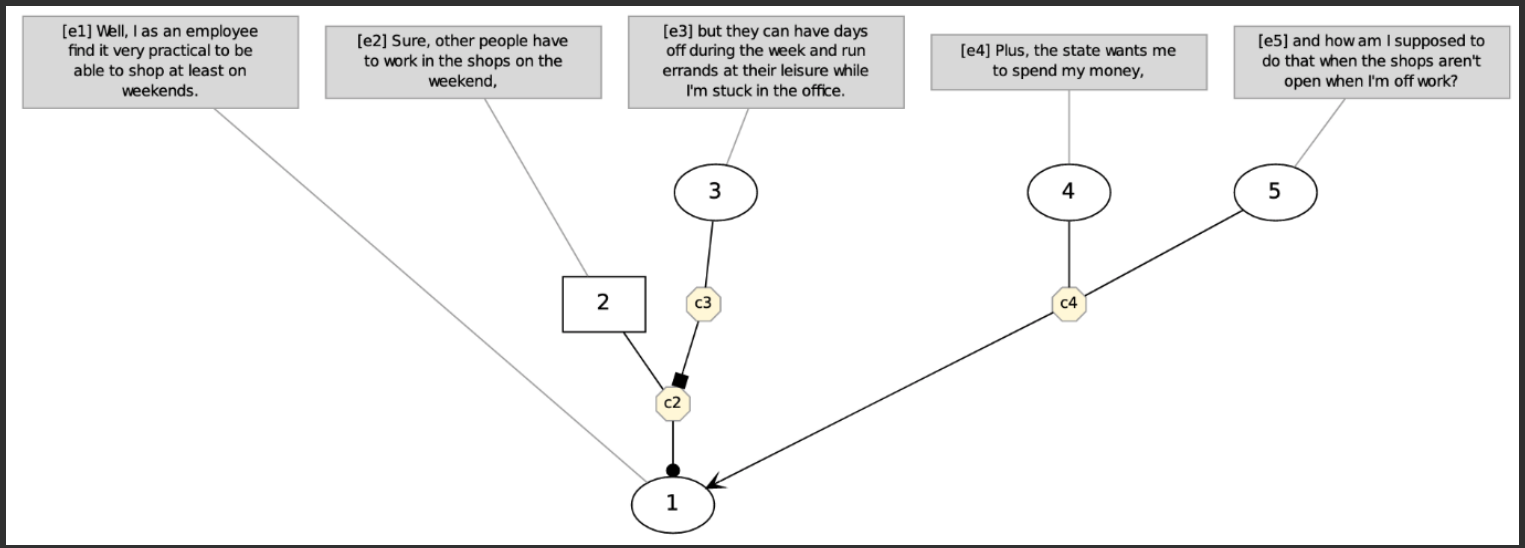
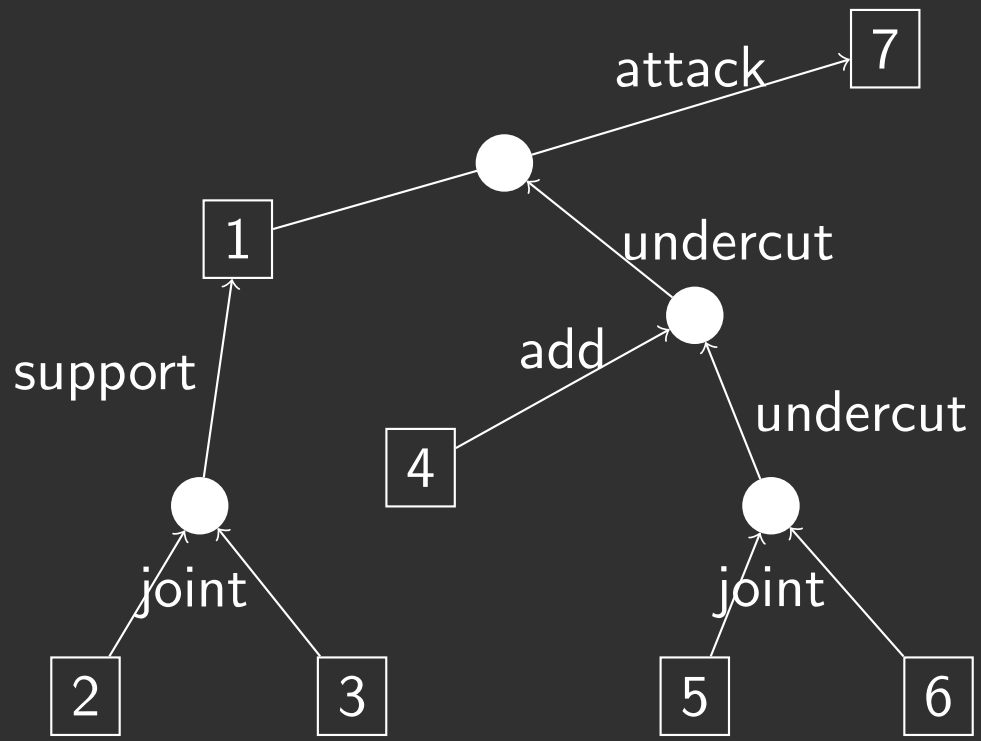


Figure: Arg annotation of PRO argument

From ARG patterns to Formal Contexts

Should we continue to separate our waste for recycling?

1. [It's annoying and cumbersome to separate your rubbish properly all the time.]
2. [Three different bin bags stink away in the kitchen
3. and have to be sorted into different wheelie bins.]
4. [But still Germany produces way too much rubbish]
5. [and too many resources are lost
6. when what actually should be separated and recycled is burnt.]
7. [We Berliners should take the chance and become pioneers in waste separation!]



From ARG patterns to Formal Contexts

	$A = \{a_1, ..., a_{86}\}$
$T = \{t_1, ..., t_{112}\}$	$(t, a) \in I$

- $T = \{t_1, ..., t_{112}\}$ is the set of micro texts
- $A = \{a_1, ..., a_{86}\}$ is the set of ARG subgraphs/patterns
- I is the incidence relation indicating that a text contains an ARG pattern

Description of the annotation

```
<?xml version='1.0' encoding='UTF-8'?>
<arggraph id="micro_b001" topic_id="waste_separation" stance="pro">
  <edu id="e1"><![CDATA[Yes, it's annoying and cumbersome to separate your rubbish properly all the time.]]></edu>
  <edu id="e2"><![CDATA[Three different bin bags stink away in the kitchen and have to be sorted into different wheelie bins.]]></edu>
  <edu id="e3"><![CDATA[But still Germany produces way too much rubbish]]></edu>
  <edu id="e4"><![CDATA[and too many resources are lost when what actually should be separated and recycled is burnt.]]></edu>
  <edu id="e5"><![CDATA[We Berliners should take the chance and become pioneers in waste separation!]]></edu>
  <adu id="a1" type="opp"/>
  <adu id="a2" type="opp"/>
  <adu id="a3" type="pro"/>
  <adu id="a4" type="pro"/>
  <adu id="a5" type="pro"/>
  <edge id="c6" src="e1" trg="a1" type="seg"/>
  <edge id="c7" src="e2" trg="a2" type="seg"/>
  <edge id="c8" src="e3" trg="a3" type="seg"/>
  <edge id="c9" src="e4" trg="a4" type="seg"/>
  <edge id="c10" src="e5" trg="a5" type="seg"/>
  <edge id="c1" src="a1" trg="a5" type="reb"/>
  <edge id="c2" src="a2" trg="a1" type="sup"/>
  <edge id="c3" src="a3" trg="c1" type="und"/>
  <edge id="c4" src="a4" trg="c3" type="add"/>
</arggraph>
```

Pro and con contexts

The complete context can be divided based on for and against arguments.

- 46 texts **for** the claim (T_{pro}) / 86 attributes
- 42 texts **against** the claim (T_{con}) / 77 attributes

	$A = \{a_1, ..., a_{86}\}$
T_{pro}	$(t, a) \in I$

	$A = \{a_1, ..., a_{77}\}$
T_{con}	$(t, a) \in I$

Project

- The project is based on three contexts: complete, and the subcontexts for and against.
- These put in relation microtexts and argumentation structures. Some are texts for a given claim while others are against, e.g., *Should Germany introduce the death penalty?*

The original micro texts and arg. structures are found here:
`https://github.com/peldszus/arg-microtexts-multilayer`

Project

- There is an additional file (`arg_attr_patterns.json`) that contains the correspondence between the structure identifier arg and the corresponding structure:
- json file $\{k : v\}$ with k the identifier, and v the structure (a character string) in the following format:
  - $t \# arg\_id$: the first line contains the structure identifier (arg_id)
  - $v\ v\_id\ v\_label$: each line starting with v describes a node v_id and its label v_label (no label on the nodes with "_")
  - $e\ src\_id\ trg\_id\ e_label$: each line starting with e describes an arc between src_id and trg_id (two previously defined nodes) and its label e_label

Project

- **The goal** of the project is to classify microtexts w.r.t. the argumentation structures that they contain.
- The classification should rely on the hypotheses (for, against, falsified generalizations) that you'll mine, and an analysis is expected. **Reference to the descriptions given in the page above is required!**
- You should sample a few examples (at least twice, with about 10% each) for testing your classifier. **This is an exploratory project!**

You'll have to form groups of a maximum of 3, and write a **short report** with your findings and analysis.

Deadline: 20 of February 2022!

References I


Freeman, J. B. (2011).
Argumentation Structure: Representation and Theory.
Springer, Dordrecht.


Lascarides, A. and Asher, N. (2007).
Segmented Discourse Representation Theory: Dynamic Semantics With Discourse Structure.
In Bunt, H. and Muskens, R., editors, *Computing Meaning*, volume 3, pages 87–124. Springer Netherlands, Dordrecht.


Mann, W. and Thompson, S. (1988).
Rhetorical structure theory: Towards a functional theory of text organization.
TEXT, 8:243–281.

References II


Peldszus, A. and Stede, M. (2015).
An annotated corpus of argumentative microtexts.
In *Proceedings of the First European Conference on Argumentation: Argumentation and Reasoned Action*, volume 2, pages 801–816, Lisbon, Portugal.


Peldszus, A. and Stede, M. (2016).
Rhetorical structure and argumentation structure in monologue text.
In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 103–112, Berlin, Germany. Association for Computational Linguistics.

References III


Yan, X. and Han, J. (2002).
gSpan: graph-based substructure pattern mining.
In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 721–724, Maebashi City, Japan. IEEE.