university of groningen

# Inter annotator agreement in discourse analysis

Nynke van der Vliet

# Presentation overview

- Introduction
- Rhetorical Structure Theory
- Data
- Measures : Cohen's Kappa
- Agreement on segmentation
- Agreement on discourse structure and relation labeling

# Introduction – MTO project

- Modeling Textual Organisation :
  - organization of text into structural units by means of **coherence** (discourse relations) and **cohesion** (lexico-semantic relations)
  - development **Dutch text corpus**, annotated with discourse relations (Rhetorical Structure Theory) and lexical cohesion
  - Automatic **Discourse Parsing**
    - Segmentation
    - Labeling with RST relations

# Rhetorical Structure Theory
## Coherence

- Coherence:

1) John hid Bill's car keys. He was drunk

2) John hid Bill's car keys. He likes pancakes

*A coherent discourse has meaningful connections between its utterances*

# Rhetorical Structure Theory
## Principles

Coherent texts consist of minimal units, which are linked to each other, recursively, through rhetorical relations

- Rhetorical relations also known, in other theories, as coherence or discourse relations

For every part of a coherent text, there is some function, some plausible reason for its presence, evident to readers

- Therefore, there must be some relation holding among the different parts of the text
- There are no gaps

# Rhetorical Structure Theory
## Components

**Units of discourse**

- Texts can be segmented into minimal units, or spans

**Nuclearity**

- Some spans are more central to the text's purpose (nuclei), whereas others are secondary (satellites)
- Based on hypotactic and paratactic relations in language

**Relations among spans**

- Spans are joined into discourse relations
- Subject matter/ Presentational relations

**Hierarchy/recursion**

- Spans that are in a discourse relation may enter into new relations

# Data
## Dutch RST corpus

**Corpus of Dutch of 80 Dutch texts:**

**• 40 expository texts: 20 encyclopedia and 20 popular-scientific texts**

**• 40 persuasive texts: 20 fundraising letters and 20 advertisements**

**RST relation set: Mann&Thompson (32 relations)**

**2 annotation steps : 1. segmentation**

**2. RST annotation**

**Annotation procedure: 2 coders annotate the text independently, after that discussion → 1 final version**

# Data
## Example Encyclopedia text

14 Op het eerste gezicht lijkt het oppervlak van Mercurius erg veel op dat van de Maan.

15 Er zijn grofweg twee typen landschap: hoogland en laagland.

16 In vergelijking met de hooglanden op de Maan zijn er relatief minder inslagkraters in het hoogland gebied van Mercurius.

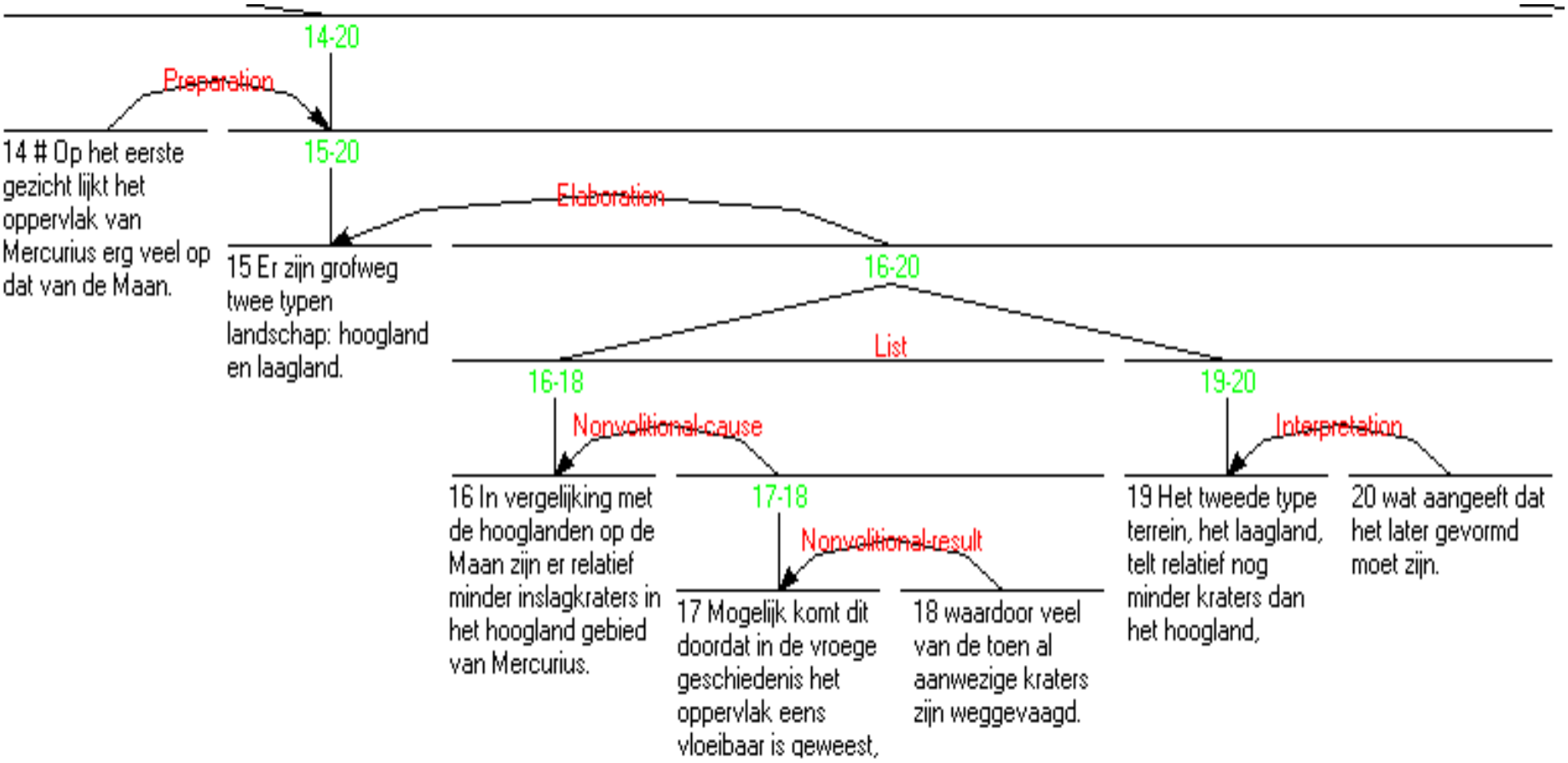
17 Mogelijk komt dit doordat in de vroege geschiedenis het oppervlak eens vloeibaar is geweest,

18 waardoor veel van de toen al aanwezige kraters zijn weggevaagd.

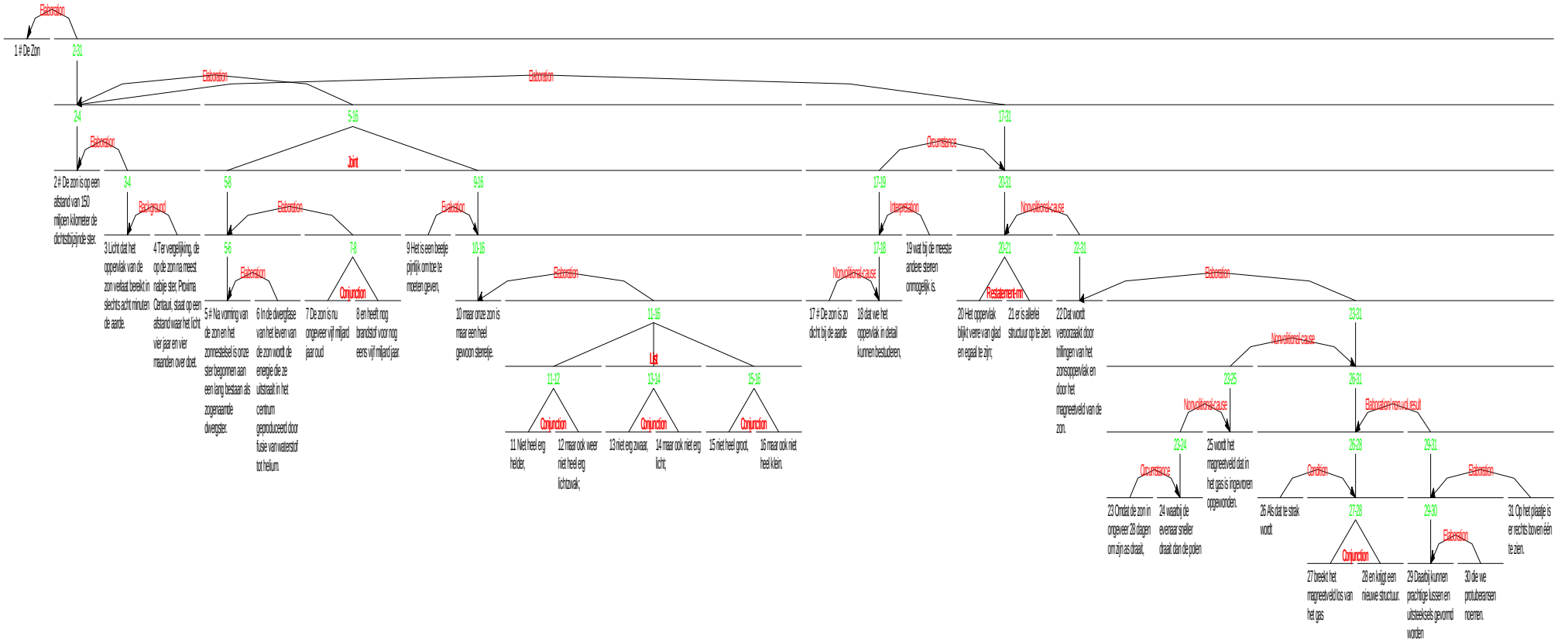
19 Het tweede type terrein, het laagland, telt relatief nog minder kraters dan het hoogland,

20 wat aangeeft dat het later gevormd moet zijn.

# Data
## Example Encyclopedia text



# Data
# a whole EE text



# Measures

## Why compute inter-annotator agreement?

- Show that subjective coding distinctions in the coding scheme can be understood and applied by people other than the coding developers
- Reproducibility/Inter-coder variance: different coders require to code in the same way

# Measures
## Model Reliability study

- A reliability study involves a set of **items** (markables), a set of **categories,** and a set of **coders**:
    - **Items** $\{i \mid i \in I\}$, with cardinality i
    - **Categories** $\{k \mid k \in K\}$with cardinality k
    - **Coders** $\{c \mid c \in C\}$with cardinality c

# Measures
## percentage/observed agreement

*The percentage of judgements on which two analysts agree when coding the same data independently*

$$\text{agr}_i = \begin{cases} 1 \text{ if the two coders assign } i \text{ to the same category} \\ 0 \text{ if the two coders assign } i \text{ to different categories} \end{cases}$$

Observed agreement over the values $\text{agr}_i$ for all items $i \in I$ is then:

$$A_o = \frac{1}{\mathbf{i}} \sum_{i \in I} \text{agr}_i$$

But:

- not correct for chance agreement (no comparability, biased)
- not correct for distribution of items among categories

# Measures
## Chance-corrected Coefficients (1)

- *Observed agreement (Ao) :* proportion of items on which 2 coders agree
- *Expected agreement (Ae):* probability of 2 annotators agreeing on any category

$$S, \pi, \kappa = \frac{A_o - A_e}{1 - A_e}$$

Ae(S) => uniform distribution : $\sum_{k \in K} \frac{1}{k} \cdot \frac{1}{k} = k \cdot \left(\frac{1}{k}\right)^2 = \frac{1}{k}$

Ae(π) => same distribution for each coder: $\sum_{k \in K} \left(\frac{n_k}{2i}\right)^2 = \frac{1}{4i^2} \sum_{k \in K} n_k^2$

Ae(K) => separate distribution for each coder: $\sum_{k \in K} \frac{n_{c_1 k}}{i} \cdot \frac{n_{c_2 k}}{i} = \frac{1}{i^2} \sum_{k \in K} n_{c_1 k} n_{c_2 k}$

# Measures
## Chance-corrected Coefficients (2)

***Example (Artstein and Poesio, 2008)***

| | | CODER A | | |
|---|---|---|---|---|
| | | STAT | IREQ | TOTAL |
| CODER B | STAT | 20 | 20 | 40 |
| | IREQ | 10 | 50 | 60 |
| | TOTAL | 30 | 70 | 100 |

The value of different coefficients applied to the data from Table 1.

| Coefficient | Expected agreement | Chance-corrected agreement |
|---|---|---|
| $S$ | $2 \times (\frac{1}{2})^2 = 0.5$ | $(0.7 - 0.5)/(1 - 0.5) = 0.4$ |
| $\pi$ | $0.35^2 + 0.65^2 = 0.545$ | $(0.7 - 0.545)/(1 - 0.545) \approx 0.341$ |
| $\kappa$ | $0.3 \times 0.4 + 0.6 \times 0.7 = 0.54$ | $(0.7 - 0.54)/(1 - 0.54) \approx 0.348$ |

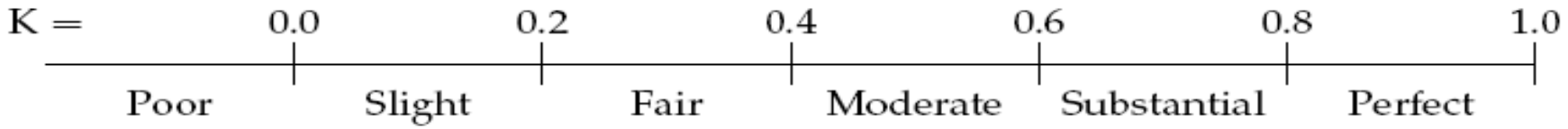
# Measures
## The interpretation of Cohen's K

- -1 < K <1
- K > 0,8 → good reliability

0.67 < K < 0.8 → allowing tentative conclusions

(Carletta(1996), based on Krippendorf (1980))

- Landis & Koch (1977):



# Measures
## Other measures

- More than two annotators
  - Fleiss Multi-π
  - Multi-K
- Weigthed Agreement Coefficients
  - Krippendorff's α
  - Cohen's Kw

# Inter-annotator Agreement on segmentation (1)

- Segmentation: dividing a text into non-overlapping elementary discourse units (clause-level)
- Classification task:
  - Items: all the words in a text / all EDUs?

  (see Marcu(1999) and Carletta (1997))
  - Categories: Boundary or Non-boundary
  - Coders: IB and NV.

# Inter-annotator Agreement on segmentation (2)

**Results for 5 encyclopedia texts and 5 fundraising letters**

| | Items | | EDU | | | word | | |
|---|---|---|---|---|---|---|---|---|
| ***file*** | ***edus*** | ***words*** | ***Ao*** | ***Ae*** | ***K*** | ***Ao*** | ***Ae*** | ***K*** |
| *EE16* | 35 | 310 | *0,9714* | 0,9714 | 0,0000 | 0,9968 | 0,8022 | *0,9837* |
| *EE17* | 30 | 351 | *1,0000* | 1,0000 | 1,0000 | 1,0000 | 0,8437 | *1,0000* |
| *EE18* | 27 | 309 | *0,9630* | 0,9630 | 0,0000 | 0,9968 | 0,8432 | *0,9794* |
| *EE19* | 38 | 402 | *0,8684* | 0,8767 | -0,07 | 0,9876 | 0,8390 | *0,9228* |
| *EE20* | 28 | 271 | *0,9643* | 0,9643 | 0,0000 | 0,9963 | 0,8176 | *0,9798* |
| *FL16* | 28 | 338 | *1,0000* | 1,0000 | 1,0000 | 1,0000 | 0,8480 | *1,0000* |
| *FL17* | 33 | 295 | *0,9697* | 0,9697 | 0,0000 | 1,0000 | 0,8039 | *0,9827* |
| *FL18* | 22 | 191 | *0,9545* | 0,9545 | 0,0000 | 0,9948 | 0,8002 | *0,9738* |
| *FL19* | 23 | 204 | *1,0000* | 1,0000 | 1,0000 | 1,0000 | 0,7999 | *1,0000* |
| *FL20* | 35 | 288 | *1,0000* | 1,0000 | 1,0000 | 1,0000 | 0,7865 | *1,0000* |

# Inter-annotator Agreement on RST analysis (1)

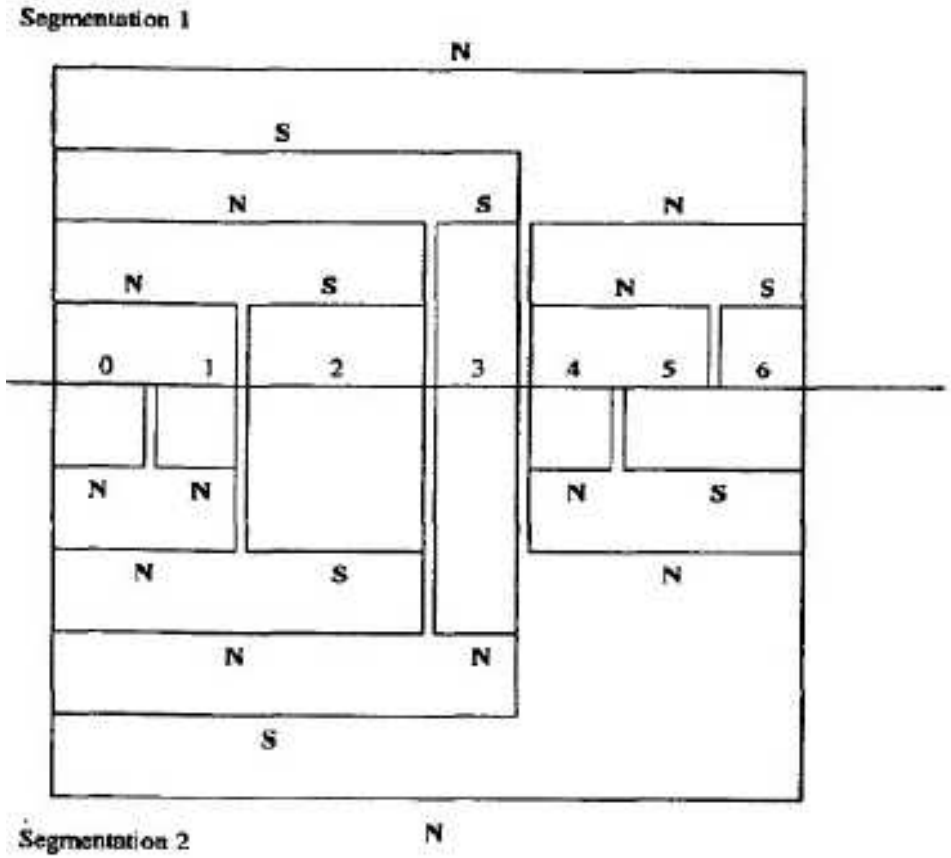
- RST analysis: Build a discourse tree from the EDUs of a text (spans, nuclearity and relations)
- Decisions at one level in the discourse tree affect decisions at other levels → NOT independent !
- Problem: *How to calculate agreement on hierarchical annotation?*

# Inter-annotator Agreement on RST analysis (2)

- Marcu et al.(1999): map hierarchical structures into sets of units that are labeled with categorical judgments, using Kappa
  - Spans (categories : Y/N)
  - Nuclearity (categories :N/S/none)
  - Relation labeling (categories: set of 32 RST relations)
- Den Ouden(2004): agreement on the levels of hierarchies, using Kappa

# Inter-annotator Agreement on RST analysis (3)

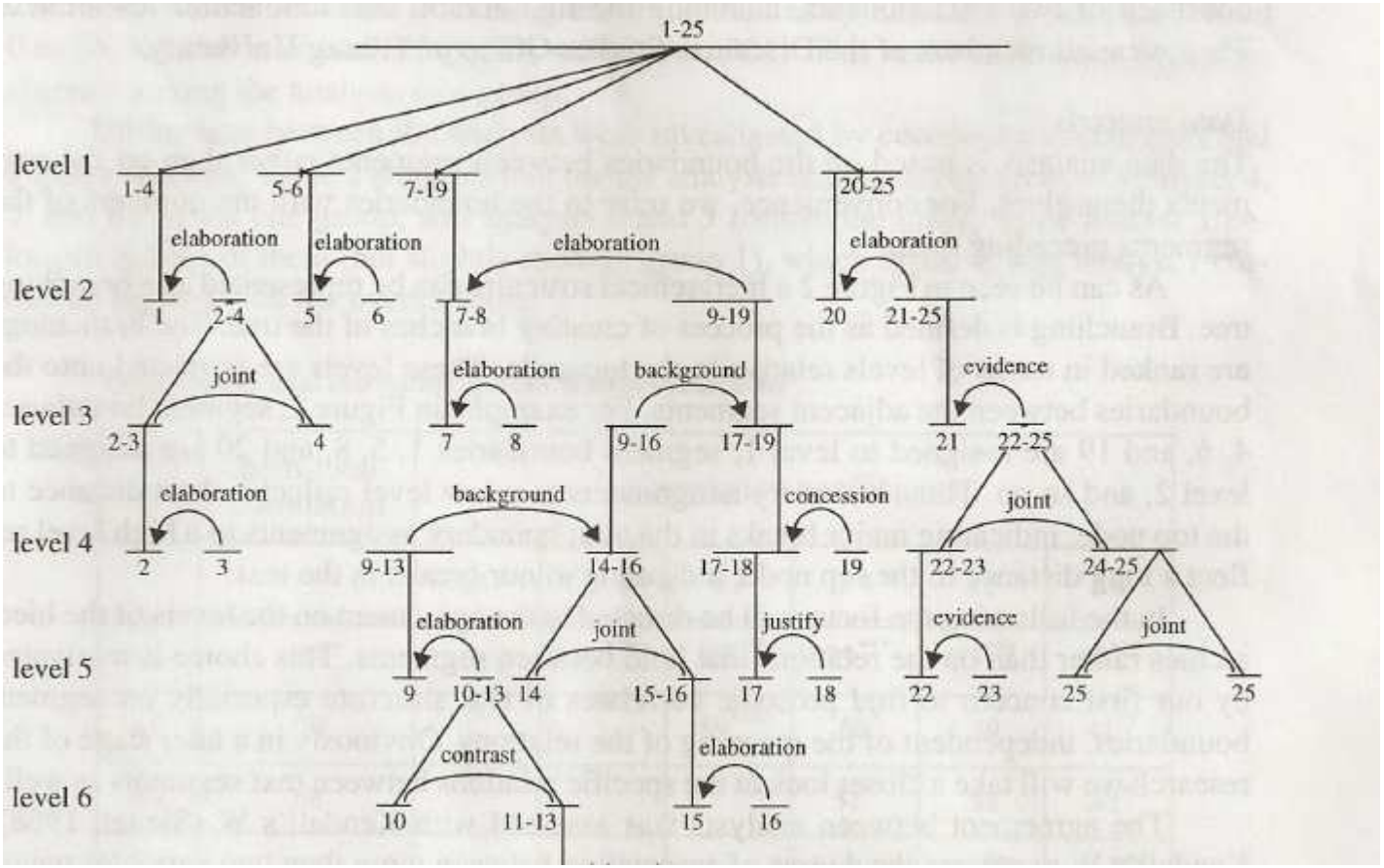
- Marcu(1999) → categorical judgments of spans and nuclearity




| Segment | Segmentation 1 | Segmentation 2 |
|---|---|---|
| [0,0] | none | N |
| [0,1] | N | N |
| [0,2] | N | N |
| [0,3] | S | S |
| [0,4] | none | none |
| [0,5] | none | none |
| [0,6] | N | N |
| [1,1] | none | N |
| [1,2] | none | none |
| [1,3] | none | none |
| [1,4] | none | none |
| ... | | |
| [4,4] | none | N |
| [4,5] | N | none |
| [4,6] | N | N |
| [5,5] | none | none |
| [5,6] | none | S |
| [6,6] | S | none |

# Inter-annotator Agreement on RST analysis (4)

- Den Ouden(2004) → categorical judgments of levels




| boundary | level |
|---|---|
| 1 | 2 |
| 2 | 4 |
| 3 | 3 |
| 4 | 1 |
| 5 | 2 |
| 6 | 1 |
| 7 | 3 |
| 8 | 2 |
| 9 | 5 |
| 10 | 6 |

# Inter-annotator Agreement on RST analysis (3)

Results for 2 Encyclopedia Texts and 2 fundraising letters: Spans,Nuclearity and Relations

| | **Items** | | Spans | Nuclearity | | Relations |
|---|---|---|---|---|---|---|
| ***Files*** | ***segments*** | ***I =n(n+1)/2**** | ***Ks*** | ***Kn*** | ***Items*** | ***Kr*** |
| FL18 | 22 | 253 | 0,85 | 0,77 | 8 | 0,38 |
| FL19 | 23 | 276 | 0,86 | 0,83 | 10 | 0,66 |
| EE17 | 30 | 465 | 0,91 | 0,83 | 16 | 0,46 |
| EE19 | 36 | 666 | 0,89 | 0,84 | 22 | 0,79 |

* all possible spans that range over the units in the text

# Inter-annotator Agreement on RST analysis (4)

Results for 2 Encyclopedia Texts and 2 fundraising letters: Levels

| *Files* | *seg* | *Kl* | *Klw* |
|---|---|---|---|
| FL18 | 22 | 0,28 | 0,43 |
| FL19 | 23 | -0,04 | 0,25 |
| EE17 | 30 | 0,42 | 0,64 |
| EE19 | 36 | 0,42 | 0,58 |

Dab=0 ↔ a =b
Dab=0,5 ↔ a-b = (-)1
Dab=1 ↔ a-b =>1 or <-1

**Cohen's Kw**

$$\kappa_w = 1 - \frac{\mathrm{D_o}}{\mathrm{D_e}}$$

$$\mathrm{D_o^{\kappa w}} = \frac{1}{\mathbf{d}_{\max}} \frac{1}{\mathbf{i}} \sum_{i \in I} \mathrm{disagr}_i = \frac{1}{\mathbf{d}_{\max}} \frac{1}{\mathbf{i}} \sum_{i \in I} \mathbf{d}_{k(c_1,i)k(c_2,i)}$$

$$\mathrm{D_e^{\kappa w}} = \frac{1}{\mathbf{d}_{\max}} \frac{1}{\mathbf{i}^2} \sum_{j=1}^{\mathbf{k}} \sum_{l=1}^{\mathbf{k}} \mathbf{n}_{c_1 k_j} \mathbf{n}_c$$

Each pair of categories is associated with a weight : $\mathbf{d}_{k_a k_b}$

# Inter-annotator Agreement on RST analysis (5)

- Problems with RST annotation method (Marcu et al, 1999):
  - Violation of independence assumption: data points over which the kappa coefficient is computed are not independent
  - None-agreements: K will be artificially high because of agreement on non-active spans.
  - Hierarchy: no difference in low-high level agreements

# Questions ?

# Discussion

- Reidsma & Carletta (2008) : Even a K measure of 0.8 does not guarantee that what looks like good performance really is. At the very least, computational linguists should look for any patterns in the disagreement among coders and assess what impact they will have.

→ How useful is the K measure?

- What other measures that take the dependency of annotations in a discourse tree into account could be used to compute inter-annotator agreement of RST annotation?

# segmentation guidelines (MTO)

- EDUs:
  - Clauses and sentences
  - adjunct clauses, with either finite or non-finite verbs
  - coordinated clauses (also elliptical conjuncts)
  - non-restrictive relative clauses (marked by a comma), also as embedded EDUs
  - fragments functioning as complete utterances (signalled with fullstop)
- Not EDUs:
  - complements of attributive and cognitive verbs

# References


- Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics (survey article). Computational Linguistics 34(4): 555-596, 2008.
- Carletta, J. C., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., and Anderson, A. (1997). *The Reliability of a Dialogue Structure Coding Scheme*. Computational Linguistics, 23(1), 13-31.
- Daniel Marcu, Magdalena Romera, and Estibaliz Amorrortu (1999). *Experiments in Constructing a Corpus of Discourse Trees: Problems, Annotation Choices*, Issues. The Workshop on Levels of Representation in Discourse, pages 71-78, Edinburgh, Scotland, July 1999.
- Ouden, H. den (2004). *Prosodic realizations of discourse structure*. PhD Thesis, University of Tilburg, Tilburg.
- Reidsma, D. and Carletta, J. (2008) *Reliability Measurement without Limits*. Computational Linguistics 34(3): 319-326. This paper has also been cited as "Reliability Measurement: There's no safe limit", which was the title under which it was initially submitted.