

What Counts as a Discourse Unit?

Towards Drawing the Boundary Between Annotation Error and Human Label Variation

Maximilian Krupop, Steffen Frenzel, Manfred Stede

Agenda

- 1 | What counts as a Discourse Unit?**
- 2 | What do models and annotators show?**
- 3 | When is a boundary difference HLV?**

Discourse units depend on framework

PDTB:

- select local relation and its **argument spans**

[**ARG1** "Of course it's better to sell private," he says],

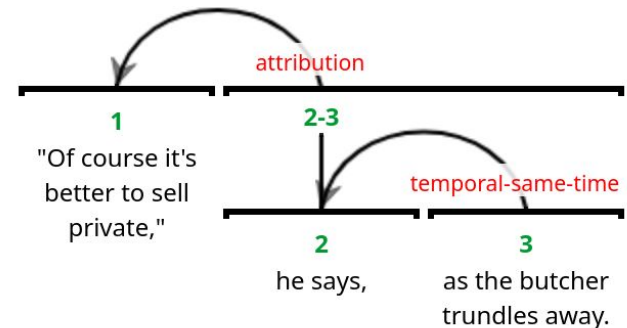
[**CONN** as]

[**ARG2** the butcher trundles away]

Temporal.Synchronous

RST-DT:

- segment complete text into **EDUs**
- connect EDUs and larger spans in document-wide tree



Discourse units depend on framework

PDTB:

- select local relation and its **argument spans**

[**ARG1** "Of course it's better to sell private," he says],

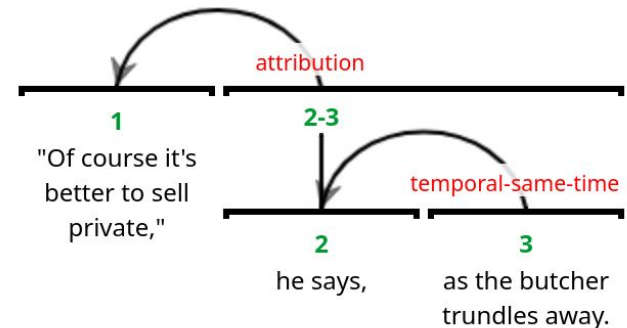
[**CONN** as]

[**ARG2** the butcher trundles away]

Temporal.Synchronous

RST-DT:

- segment complete text into **EDUs**
- connect EDUs and larger spans in document-wide tree



EDU boundaries are already analyses

- Elementary **D**iscourse **U**nits are usually clause-like, but should form *“complete, distinct unit of information that the subsequent discourse may connect to.”* (Stede 2011, p. 89)
- segmentation granularity can depend on downstream task
 - later analyses inherit chosen boundaries!

Text → *EDU segmentation* → *RST*

Guidelines specify boundaries and structures in RST

Construction	Carlson & Marcu (2001)	Stede, Taboada & Das (2017)
Attribution complement	[He said] [that the car won.]	[He said that the car won.]
Phrasal adjunct	[We stayed home] [because of the rain.]	[We stayed home because of the rain.]

- same constructions, different segmentation rules
 - permitted relation set
 - presence of strong discourse markers

2 | What do models and annotators show?

Automatic segmentation struggles with complex prose

- eduseg_de (XLM-RoBERTa-base) results for German text

	APA-RST	Hannah Arendt, <i>Franz Kafka</i>
boundary F1	0.943	0.776
text profile	short news text avg. 14.91 toks/sent. avg. 1.51 EDUs/sent.	long, syntactically complex essay avg. 33.58 toks/sent. avg. 2.87 EDUs/sent.

Human agreement also drops on complex prose

- same annotation setup, 2 human annotators

	APA-RST	Hannah Arendt, <i>Franz Kafka</i>
Human boundary F1	.957	.807

- scores show **where disagreement increases** but **not what disagreements mean**

3 | When is a boundary difference HLV?

Data overview

Material	Annotations	ToDo
Arendt, Kafka (DE)	2 human annotations	diagnose
Arendt, Human Condition (EN)	2 human annotations	diagnose
Arendt, Human Condition (DE)	2 human annotations	diagnose
PCC (DE)	2 human annotations	needs harmonization
Arendt, Kafka (EN)	1 human annotation	collect second
Salinger, Catcher (EN)	1 human annotation	collect second

German $\approx$ 3,500 EDUs, English $\approx$ 1,400 EDUs

- prerequisite: independently double-annotated data
- research goal: automate/model this process



Annotation error under same setup

- English, Hannah Arendt: *The Human Condition*

Annotator A (annotation error)

[To these preoccupations and perplexities,] [this book does not offer an answer.]

Annotator B (= adjudicated reference)

[To these preoccupations and perplexities, this book does not offer an answer.]

- dependent prepositional phrase, not complete & distinct information unit
- are both analyses defensible? **no** → **annotation error**

Boundary differences can be genuine HLV

- English, Hannah Arendt: *The Human Condition*

Annotator A: reformulation as separate EDU

[and they may be of great relevance to man in the singular,] [that is, to man in so far as he is not a political being,]

reading: separately interpretable elliptical reformulation

Annotator B: reformulation attached

[and they may be of great relevance to man in the singular, that is, to man in so far as he is not a political being,]

reading: local clarification of “*man in the singular*”

- shared setup, both analyses defensible? **yes** → **HLV**

Boundary differences can be genuine HLV

- English, Hannah Arendt: *The Human Condition*

[There may be truths beyond speech,]

[and they may be of great relevance to man in the singular,]

[that is, to man in so far as he is not a political being,]

[whatever else he may be.]

Conclusion & outlook

Diagnosis	Response
scheme-induced difference	document or harmonize
annotation error	correct
HLV	preserve, even if one reference is adjudicated

Why does human agreement fall to .807 F1 on complex prose?

- **working hypothesis:** nested constructions make otherwise clear principles interact
- **next step:** stress-test guideline instructions with more texts

Outlook

Analyse difficult boundary cases

- focus on sentence-internal candidate boundaries
- test attribution cues, interrupted host units
- flag likely annotation errors and HLV candidates for review

Re-evaluate existing models

- compare predictions with each human segmentation
- check whether deviations match human-supported alternative

Model alternative segmentations

- boundary probabilities as first baseline

Thank you for your attention.

References

EDU definition and segmentation guidelines

Carlson, L., & Marcu, D. (2001). *Discourse tagging reference manual* (ISI-TR-545). University of Southern California Information Sciences Institute.

Stede, M. (2011). *Discourse processing*. Morgan & Claypool.

Stede, M., Mamprin, S., & Peldszus, A. (2016). Diskurssegmentierung. In M. Stede (Ed.), *Handbuch Textannotation: Potsdamer Kommentarkorpus 2.0* (pp. 23–44). Universitätsverlag Potsdam.

Stede, M., Taboada, M., & Das, D. (2017). *Annotation guidelines for rhetorical structure* (Version 1.0). University of Potsdam & Simon Fraser University.

References

Data, corpora, and empirical studies I

Arendt, H. (2018). *The human condition* (2nd ed., enlarged). University of Chicago Press.

Arendt, H. (2018). *Sechs Essays* (B. Hahn, Ed.). Göttingen: Wallstein Verlag.

Carlson, L., Marcu, D., & Okurowski, M. E. (2002). *RST Discourse Treebank* (LDC2002T07). Linguistic Data Consortium.

Demberg, V., Scholman, M. C. J., & Torabi Asr, F. (2019). How compatible are our discourse annotation frameworks? Insights from mapping RST-DT and PDTB annotations. *Dialogue & Discourse*, 10(1), 87–135.

Frenzel, S., Krupop, M., & Stede, M. (2026). Discourse segmentation of German text with pretrained language models. *Journal for Language Technology and Computational Linguistics*, 39(1), 1–31.

References

Data, corpora, and empirical studies II

Prasad, R., Webber, B., Lee, A., & Joshi, A. (2019). *Penn Discourse Treebank Version 3.0* (LDC2019T05). Linguistic Data Consortium.

Seemann, H. J., Shahmohammadi, S., Stede, M., & Scheffler, T. (2024). Discourse-level features in spoken and written communication. In *Proceedings of KONVENS 2024*.

Shahmohammadi, S., & Stede, M. (2024). Discourse parsing for German with new RST corpora. In *Proceedings of KONVENS 2024*, 65–74.

Sidarenka, U., Peldszus, A., & Stede, M. (2015). Discourse segmentation of German texts. *Journal for Language Technology and Computational Linguistics*, 30(1), 71–98.

Hewett, F. (2023). APA-RST: A text simplification corpus with RST annotations. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse*, 173–179.

Appendix

Annotation error

- German, Hannah Arendt: *Franz Kafka*

[Das einzige, was den Leser in Kafkas Werk lockt und verlockt, ist die Wahrheit selbst,]

[und diese Verlockung ist Kafka in seiner stillosen Vollkommenheit]

[– jeder »Stil« würde durch seinen eigenen Zauber von der Wahrheit ablenken –]

[bis zu dem unglaublichen Grade geglückt, daß seine Geschichten auch dann in Bann schlagen,]

[wenn der Leser ihren eigentlichen Wahrheitsgehalt erst einmal nicht begreift.]

- two instructions conflict:
Mark an independent proposition ↔ *Do not create incomplete fragments*
- are both analyses defensible? **no** (given guideline rules) → **annotation error**

Guidelines specify boundaries and structures

Attribution complement

RST-DT

[He said] [that the car won.]

Stede, Taboada & Das (2017)

[He said that the car won.]

Embedded relative clause

[The red car,]₁

[which my friend had bought last week,]₂

[narrowly won the race.]₃

RST-DT: ₂ embedded, ₁ + ₃ SAME-UNIT

annotation setup = boundary rules + permitted structures