Annotating Multiparty Discourse: Challenges for Agreement Metrics

Nina Wacholder, Smaranda Muresan,
Debanjan Ghosh, Mark Aakhus

{ninwac|debanjan.ghosh|aakhus}@rutgers.edu
smara@ccls.columbia.edu

8th Linguistic Annotation Workshop (LAW VIII)
COLING 2014, Dublin Ireland
Aug. 23-24, 2014

Long term research goal

- Build computational model(s) of argumentation in multi-party discourse based on Argumentative Discourse Units (ADUs) :

- ADUs: elementary units of argumentation related to each other in a theory of discourse) (Peldzsus and Stede 2013)

Key Challenge:
How to build models that reflect natural variation in human judgments about ADUs

- Fuzzy boundaries (Grosz and Sidner 1986; Krippendorff, 2004; Artstein and Poesio 2008)

- Multiple (>2) annotators (Artstein and Poesio 2008; Bayerl and Paul 2011) makes it hard to achieve 'good' IAA

Outline of the rest of the paper

- Argumentative Discourse Units in Pragmatic Argumentation Theory (PAT) (Van Eemeren et al. 1993)
- Report on annotation study
- Limited usefulness of state-of-the-art IAA for understanding fuzzy boundaries and variation in annotator behavior
- Our approach: Identify ADUs that are harder/ easier to annotate based on overlapping annotations
- Conclusion

Outline of the rest of the paper

- **Argumentative Discourse Units in Pragmatic Argumentation Theory (PAT) (Van Eemeren et al. 1993)**
- Report on annotation study
- Limited usefulness of state-of-the-art IAA for understanding fuzzy boundaries and variation in annotator behavior
- Our approach: Identify ADUs that are harder/ easier to annotate based on overlapping annotations
- Conclusion

Relational Theory of Discourse

Pragmatic Argumentation Theory (PAT) (Van Eemeren et al. 1993)

Argumentative Discourse Units in Pragmatic Argumentation Theory (PAT) (slightly adapted from Annotation Guidelines)

- **Target: A Target is (a part of) a *prior action* that has been called out by a subsequent action.**
- **Callout: A Callout is (a part of) a *subsequent action* that selects (a part of) a prior action and marks and comments on it in some way.**
- **Response: A link between Callout and Target that occurs when a subsequent action refers back to (is a response to) a prior action.**

User1

But when we first tried the iPhone it felt natural immediately, we didn't have to 'unlearn' old habits from our antiquated Nokias & Blackberrys. That happened because the iPhone is a truly great design.

Target

User2

That's very true. With the iPhone, the sweet goodness part of The UI is immediately apparent. After a minute or two, you're feeling empowered and comfortable.

It's the weaknesses that take several days or we o really understanding and get frustrated by.

Callout

User3

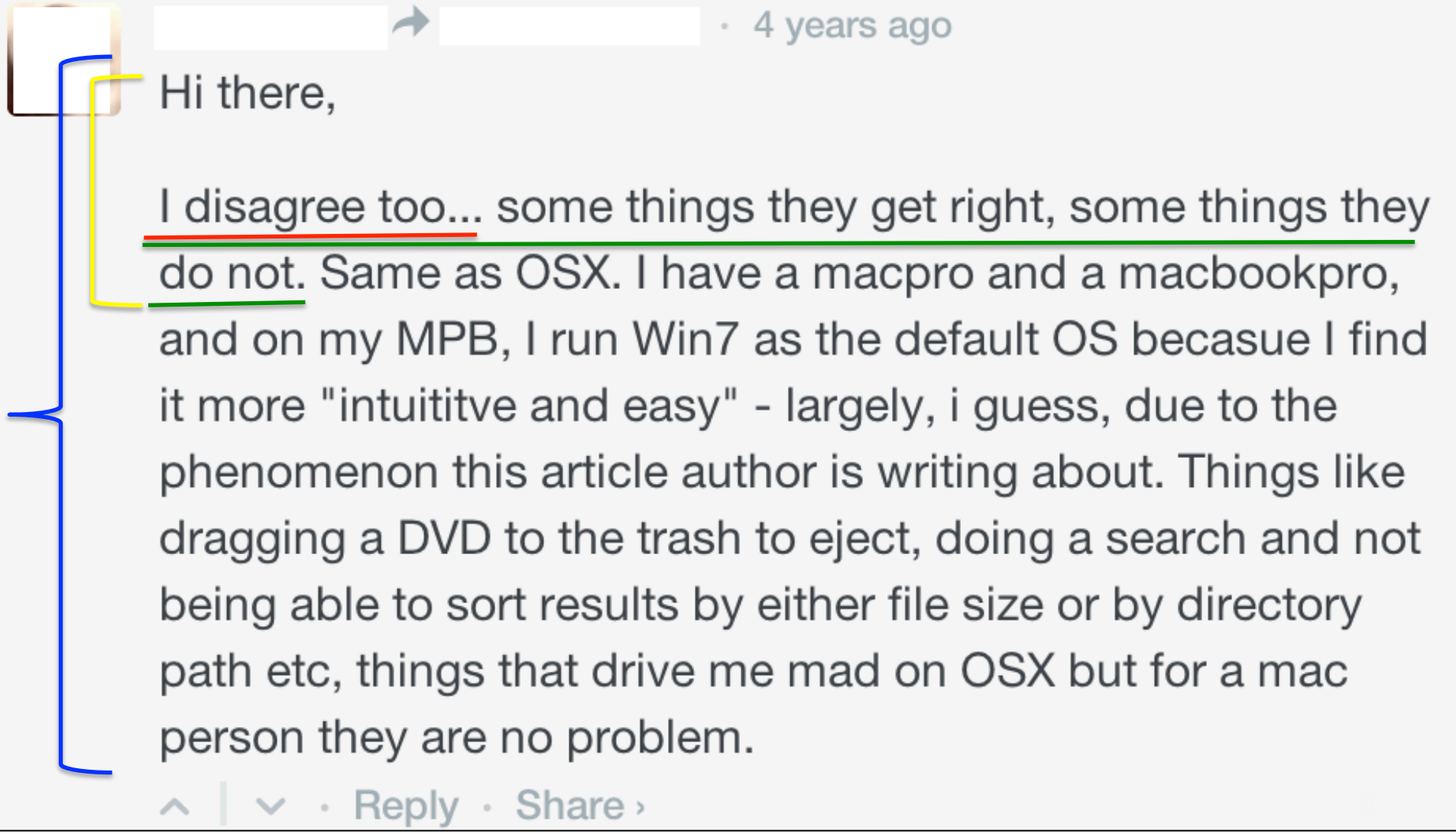
I disagree that the iPhone just "felt natural immediately"… in my Opinion it feels restrictive and over simplified, sometimes to the point of frustration.

Callout

Argument Mining (Peldszus and Stede, 2013)

1. Segmentation
2. Segment Classification
3. Relation Identification

Callouts: Variation in assigning boundaries



· 4 years ago

Hi there,

I disagree too... some things they get right, some things they do not. Same as OSX. I have a macpro and a macbookpro, and on my MPB, I run Win7 as the default OS becasue I find it more "intuititve and easy" - largely, i guess, due to the phenomenon this article author is writing about. Things like dragging a DVD to the trash to eject, doing a search and not being able to sort results by either file size or by directory path etc, things that drive me mad on OSX but for a mac person they are no problem.

· Reply · Share ›

Core: Overlapping text identified by multiple judges

Annotator	Selected text string, in characters
A1	----------CORE----------
A2	--------CORE
A3	----------CORE
Char	0---------10---------20---------30---------40---------50

CORE

Outline

- Argumentative Discourse Units in Pragmatic Argumentation Theory (PAT))
- **Annotation study**
- Limited usefulness of IAA for understanding annotator behavior
- Our approach: Identify ADUs that are harder/ easier to annotate
- Conclusion

Annotation Study Overview

- Corpus: Five threads, each consisting of a blog posting about technology from technorati.com and the first 100 comments
- Judges: Five trained annotators (experts)
- Guidelines: Carefully developed and pre-tested
- Boundary assignment: Annotators free to choose any text segment to represent an ADU

Preliminary analysis of annotated text:

	CALLOUTS PER THREAD				
THREAD	A1	A2	A3	A4	A5
Android	73	99	97	118	110
Ban	46	73	66	86	83
iPad	68	86	85	109	118
Layoffs	71	83	74	109	117
Twitter	76	102	70	113	119
Avg.	66.8	88.6	78.4	107	109.4

Preliminary analysis of annotated text: Number of callouts per thread

	ANNOTATORS				
THREAD	A1	A2	A3	A4	A5
Android	73	99	97	118	110
Ban	46	73	66	86	83
iPad	68	86	85	109	118
Layoffs	71	83	74	109	117
Twitter	76	102	70	113	119
Avg.	66.8	88.6	78.4	107	109.4

Outline of the rest of the paper

- Argumentative Discourse Units in Pragmatic Argumentation Theory (PAT) (Van Eemeren et al. 1993)
- Report on annotation study
- **Limited usefulness of state-of-the-art IAA for understanding fuzzy boundaries and variation in annotator behavior**
- Our approach: Identify ADUs that are harder/ easier to annotate based on overlapping annotations
- Conclusion

Outline

- Callouts and Targets: Argumentative Discourse Units in Pragmatic Argumentation Theory (PAT) (Van Eemeren et al. 1993)
- Annotation study
- ***Challenges in assessing IAA***
- Our approach: Identify ADUs that are harder/ easier to annotate
- Conclusion

State of the art techniques for calculating IAA for text with fuzzy boundaries

- P/R/F1 based IAA (Wiebe et al., 2005)
  - exact match (EM)
  - overlap match (OM)

- Krippendorff's α (Krippendorff, 2004)

Inter Annotator Agreement (IAA)

Thread	F1_EM	F1_OM	Krippendorff's α
Android	54.4	87.8	0.64
Ban	42.5	85.3	0.75
iPad	51.2	86.0	0.73
Layoffs	51.9	87.5	0.87
Twitter	53.8	88.5	0.82

Preliminary conclusion: IAA provides limited understanding of annotator behavior

- IAA gives only a single rating for an entire document/ conversation
- Difficult to infer from IAA what segments of text are easier or harder to annotate
- More information needed about:
  - Characteristics of phenomenon being studied
  - Properties of text being annotated
  - Annotator behavior

Outline of the rest of the paper

- Argumentative Discourse Units in Pragmatic Argumentation Theory (PAT) (Van Eemeren et al. 1993)
- Report on annotation study
- Limited usefulness of state-of-the-art IAA for understanding fuzzy boundaries and variation in annotator behavior
- **Our approach: Identify ADUs that are harder/ easier to annotate based on overlapping annotations**
- Conclusion

Core: Overlapping text identified by multiple judges

Annotator	Selected text string, in characters
A1	-----------------------------------
A2	--------
A3	----------------------
Char	0---------10---------20---------30---------40---------50

CORE

Hierarchical clustering algorithm (Hastie 2009)

- Each ADU starts in its own cluster
- Overlapping text indicates that two ADUs belong in the same cluster
- Merge pairs of clusters with overlapping text as they move up in the hierarchy
- Stop when no more clusters can be merged

Clusters where all five annotators agree on Callout

Annotators	Callout
A1, A2, A3, A4, A5	I disagree too. some things they get right, some things they do not.
A1, A2, A3, A4, A5	I remember Apple telling people give the UI and the keyboard a month and you'll get use to it. Plus all of the commercials showing the interface. So no, you didn't just pick up the iPhone and know how to use it. It was pounded into you.

Cluster with *core* supported by all five judges

Anno-tators	Callout
A1	I'm going to agree that my experience required a bit of getting used to. . .
A2, A3, A4	I'm going to agree that my experience required a bit of getting used to. . . I had arrived to the newly minted 2G Gmail and browsing.
A5	I'm going to agree that my experience required a bit of getting used to. . . I had arrived to the newly minted 2G Gmail and browsing. Greatbrowser on the iPhone but . . . Opera Mini can work wonders

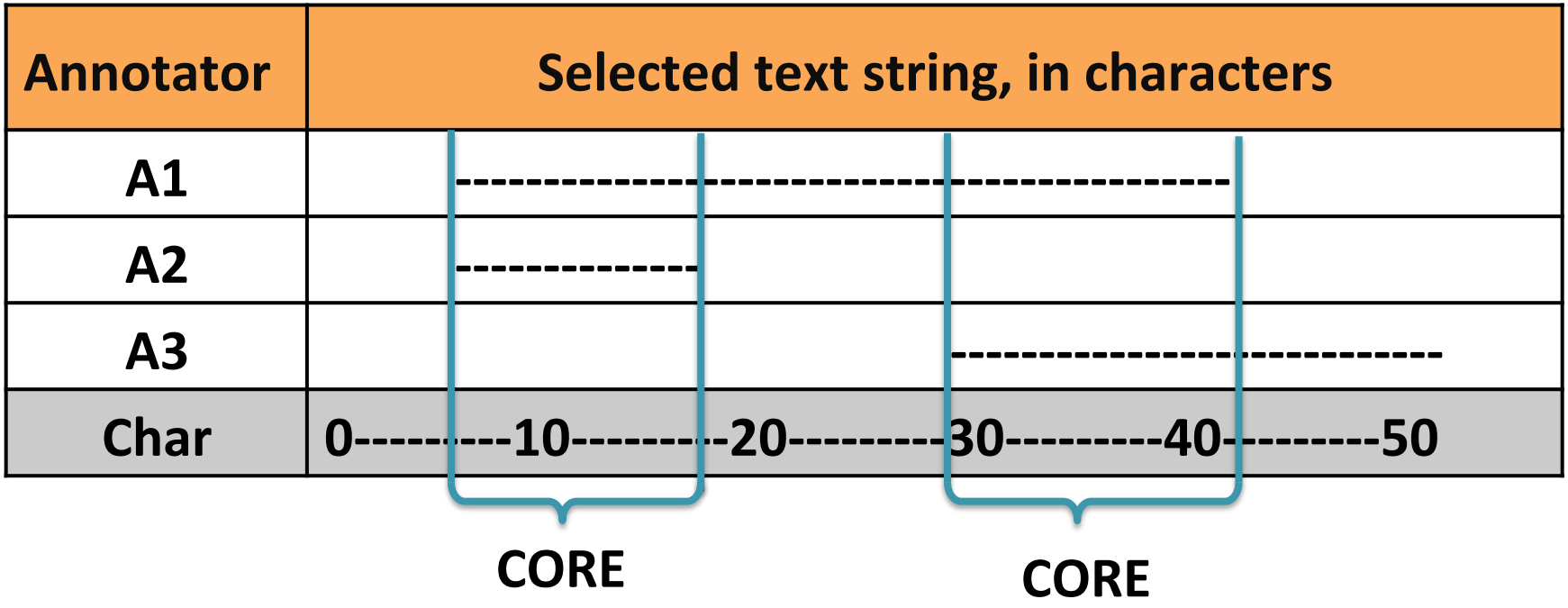
Clusters with cores selected by fewer annotators

Annotators	Callout
A4, A5	These iPhone Clones are playing catchup. Good luck with that.
A4	Do you know why the Pre was able to get big games companies to come onboard when it is a wasteland in the android world? Did they pay them off? Is it various handset/builds/resolution issues?

ADUs that are harder/easier to annotate

Thread	Number of Clusters	Support for core				
		5	4	3	2	1
Android	91	52	16	11	7	5
Ban	89	25	18	12	20	14
Ipad	88	41	17	7	13	10
Layoffs	86	41	18	11	6	10
Twitter	84	44	17	14	4	5
DIFFICULTY		EASIEST ⟷ HARDEST				

Clusters with more than one core



Hierarchical clustering algorithm revised (Hastie 2009)

- Each ADU starts in its own cluster
- Overlapping text indicates that two ADUs belong in the same cluster
- Using hierarchical clustering, merge pairs of clusters as they move up in the hierarchy
- Stop when no more clusters can be merged
- ***Split clusters with multiple cores into separate clusters and merge with other clusters with overlapping text***

ADUs that are harder/easier to annotate – after splitting

Thread	Number of Clusters (change after clustering)	Number of annotators per ciuster who identify the core				
		5	4	3	2	1
Android	93 (91 + 2)	51	15	14	8	5
Ban	91 (89 + 2)	25	19	12	21	14
Ipad	89 (88 + 1)	41	16	9	13	10
Layoffs	89 (86 + 3)	40	17	14	8	10
Twitter	87 (84 + 3)	43	15	20	4	5
SCALE OF DIFFICULTY		EASIEST ⟷ HARDEST				

Thread	Number of Clusters	Number of identical annotations per cluster				
		5	4	3	2	1
Android	93 (91 + 2)	51 (52)	15 (16)	14 (11)	8 (7)	5 (5)
Ban	91 (89 + 2)	25 (25)	19 (18)	12 (12)	21 (20)	14 (14)
Ipad	89 (88 + 1)	41 (41)	16 (17)	9 (7)	13 (13)	10 (10)
Layoffs	89 (86 + 3)	40 (41)	17 (18)	14 (11)	8 (6)	10 (10)
Twitter	87 (84 + 3)	43 (44)	15 (17)	20 (14)	4 (4)	5 (5)
SCALE OF DIFFICULTY		EASIEST ⟷ HARDEST				

Outline

- Callouts and Targets: Argumentative Discourse Units in Pragmatic Argumentation Theory (PAT) (Van Eemeren et al. 1993)
- Annotation study
- Challenges in assessing IAA
- Our approach: Identify ADUs that are harder/ easier to annotate
- ***Conclusion***

Research contributions

- New approach to analyzing characteristics of ADUs, text and judges
- Use hierarchical clustering (Hastie 2009) to characterize ADUs in terms of how readily annotators judge them:
  - Easier: Identified by more annotators
  - Harder: identified by fewer annotators
- Advance our understanding of variation in annotation of ADUs

Continuing and Future Work

- Qualitative analysis of Callouts and Targets in this study
- Quantify extent of agreement among judges
- Annotation of ADUs in different settings(e.g. healthcare forums) or on different topics
- Develop ontology of ADUs and relationships between ADUs (Aakhus et al. 2013)
- Use of crowd sourcing to do finer grained annotation (Ghosh et al. 2014)
- Build learning models for Callouts and Targets

Resource Sharing

Annotation guidelines, raw text, annotated text and hierarchical clustering algorithm are available at

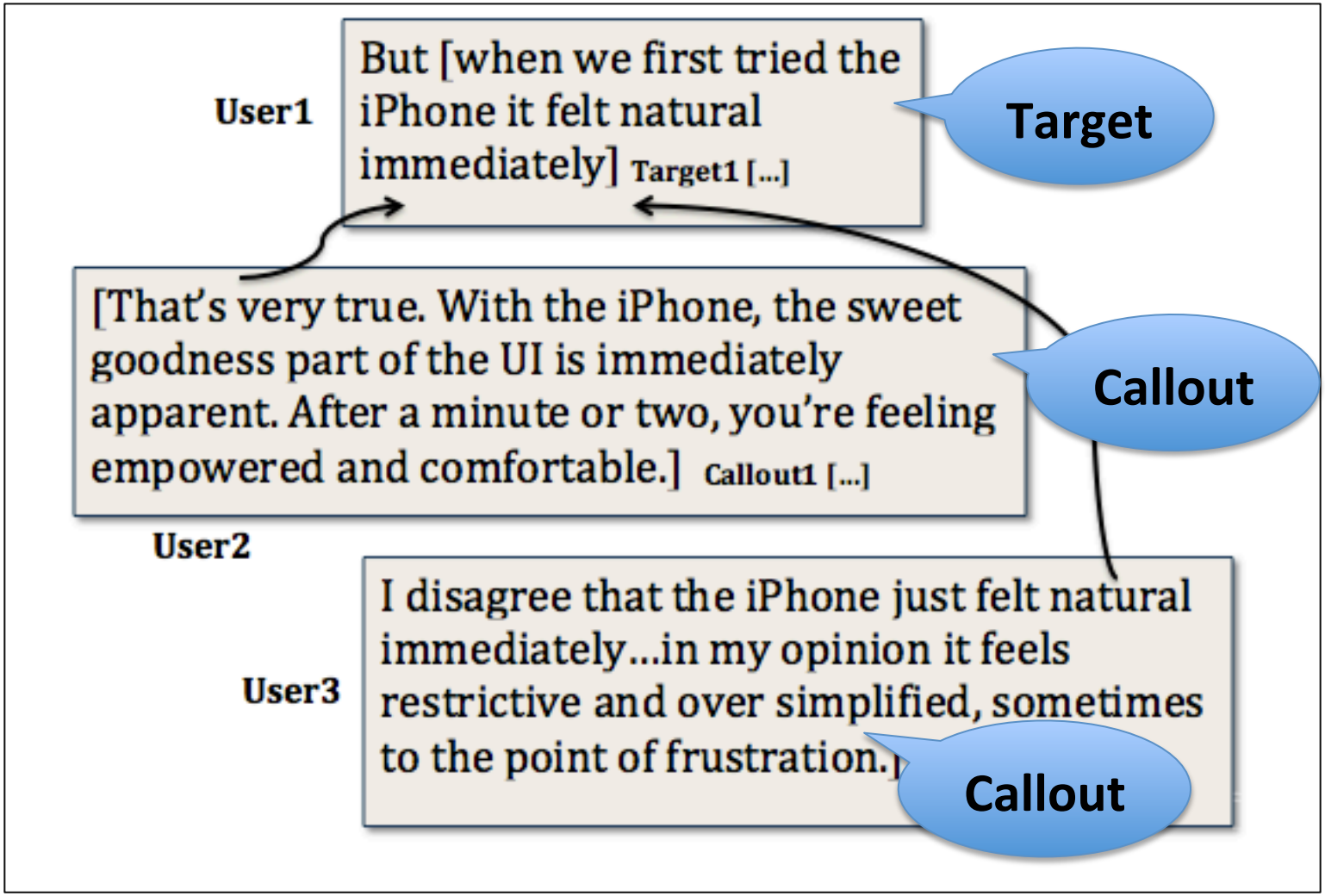
http://wp.comminfo.rutgers.edu/salts/?p=80

Backup Slides

Argument Mining (Peldszus and Stede 2013)

- Identify the boundaries of Argumentative Discourse Units (ADUs)
- Classify ADUs
- Relate ADUs to each other
  - Which comments respond to which?
  - What is the nature of the response?

ADUs: Callout and Target



Reliability *and* Validity (Krippendorff 2004)

- **Reliability "In Kaplan and Goldsen's (1965 words", reliable data, by definition, are data that remain constant throughout variations in the measuring process.'(pp.83-84). Accordingly, a research procedure is reliable when it responds to the same phenomena in the same way regardless of the circumstances of its implementation. This is the measurement theory of conception of reliability" (p.211)**

- **"Validity is that quality of research results that leads us to accept them as true, as speaking about the real world of people, phenomena, events, experiences and actions. … A content analysis is valid if the inferences drawn from the available texts withstand the test of independently available evidence, of new observations, of competing theories or interpretations, or of being able to inform successful actions." (p.313)**

Annotation study design

- Guidelines developed following recommendations of Krippendorff (2004a)
  - Clear, carefully tested instructions
  - No communication among annotators about study during study
  - No discussion among judges to decide 'best' annotation

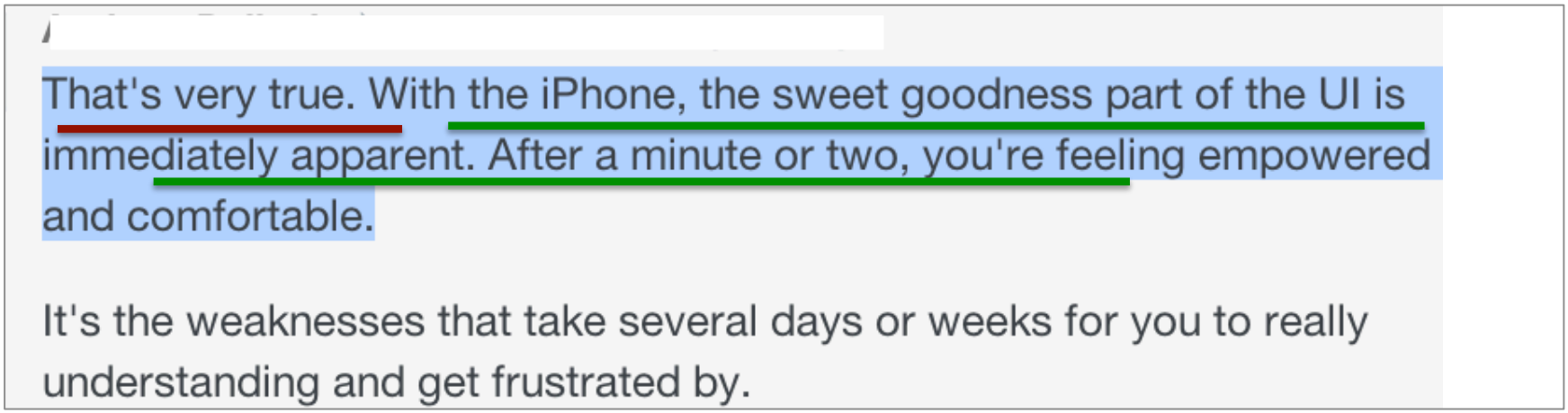
Cluster where all annotators agree on Callout but not Target

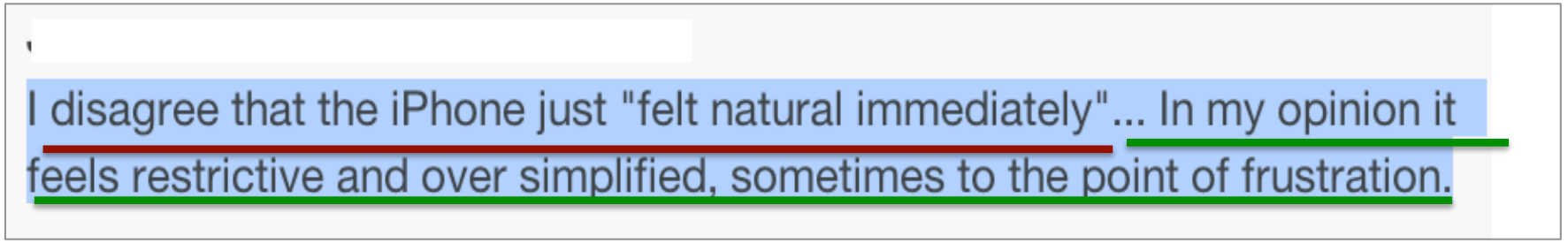
# of Anno	Callout	Target
5	I disagree too. some things they get right, some things they do not.	the iPhone is a truly great design.
		That happened because the iPhone is a truly great design.

Examples of Clusters with Less Than Full Support

# of Anns	Callout	Target
2	These iPhone Clones are playing catchup. Good luck with that.	griping about issues that will only affect them once in a blue moon
1	Do you know why the Pre ...various hand-set/builds/resolution issues?	Except for games?? iPhone is clearly dominant there.

Stance and rationale





Stance

Rationale

Variation in P/R/F depending on which annotator is selected as gold standard

CALLOUTS – EXACT MATCHING					
Ann	**Avg P**	**Avg R**	**Avg F1**	**Max F1**	**Min F1**
A1	**40.7**	**57.7**	**47.8**	**60**	**36.7**
A2	**51.7**	**51.2**	**51.4**	**58.3**	**43**
A3	**54.2**	**57.8**	**55.9**	**61.4**	**47.9**
A4	**59.7**	**49.1**	**53.9**	**61.4**	**47.3**
A5	**55**	**45.6**	**49.9**	**58.3**	**36.7**

Variation in P/R/F1

Copy previous slide for target to show variation depends on type of phenomenon

Technical difficulty with Krippendorf's α

- Identification of Targets is dependent on (temporally secondary to) identification of Callouts.
- In multiple instances an annotator links multiple Callouts to two or more overlapping Targets. Depending on the Callout, the same unit (i.e., text segment) can represent an annotation (a Target) or a gap between two Targets. Computation of is based on the overlapping characters of the annotations and the gaps

between the annotations. Naturally, if a single text string is assigned different labels (i.e. annotation

or a gap between annotations) in different annotations, does not produce meaningful results. The

inapplicability of Krippendorff's to Targets is a significant limitation for its use in discourse annotation

(To save space we only show results for Callouts in subsequent tables.)

The examples in Section 3 show a fundamental limitation of both P/R/

Thread rank by IAA (descending)

	Thread rank by IAA (descending)		
Rank	**F1_EM**	**F1_OM**	**Krippendorff's α**
1	Android	Twitter	Layoffs
2	Twitter	Android	Twitter
3	Layoffs	Layoffs	Ban
4	iPad	iPad	iPad
5	Ban	Ban	Android

Expert Annotators: Lumpers vs. Splitters

- Annotators were free to chose any text segment to represent an ADU