Computational Argumentation Quality Assessment in Natural Language

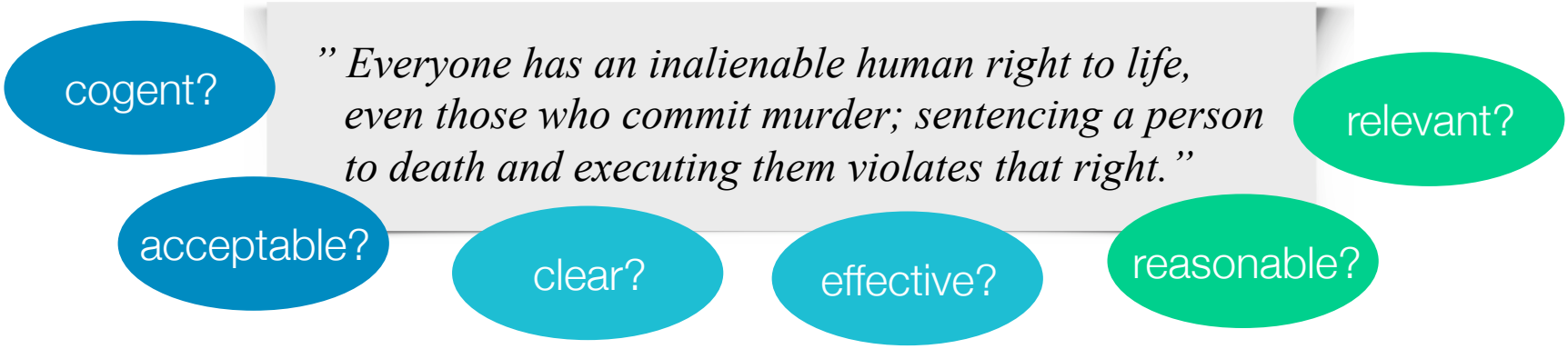
Henning Wachsmuth Bauhaus-Universität Weimar
Nona Naderi University of Toronto
Yonatan Bilu IBM Research – Haifa
Yufang Hou IBM Research – Ireland
Vinodkumar Prabhakaran Stanford University
Tim Alberdingk Thijm University of Toronto
Graeme Hirst University of Toronto
Benno Stein Bauhaus-Universität Weimar

Bauhaus-Universität Weimar

www.webis.de
henning.wachsmuth@uni-weimar.de

Motivation

- **Argument mining**
  - Identifies arguments in natural language text
  - Does not assess quality
- **Argumentation quality assessment**
  - Critical for any application built upon argument mining



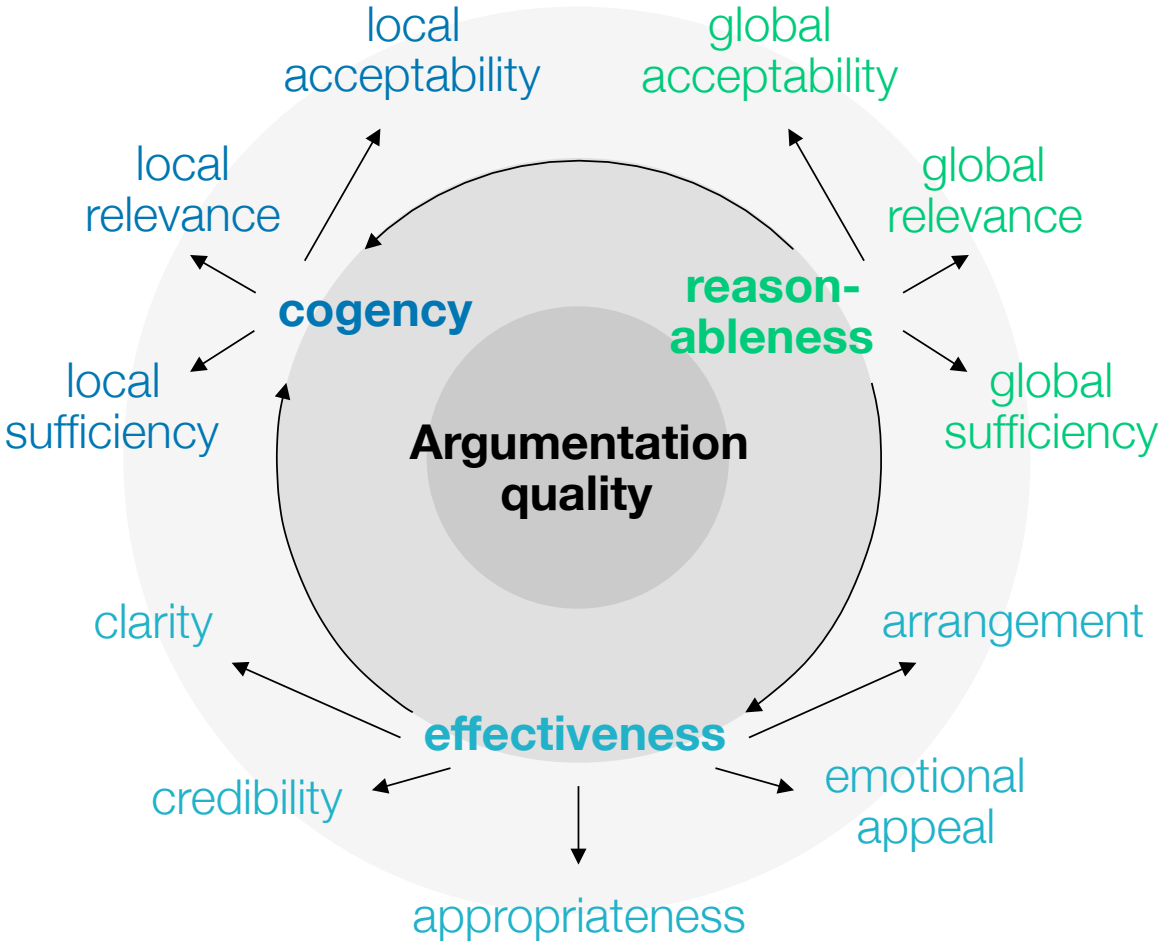
- **Challenges**
  - Several quality dimensions at different granularities
  - Some highly subjective
  - How *should* we argue vs. how *do* we argue

Background

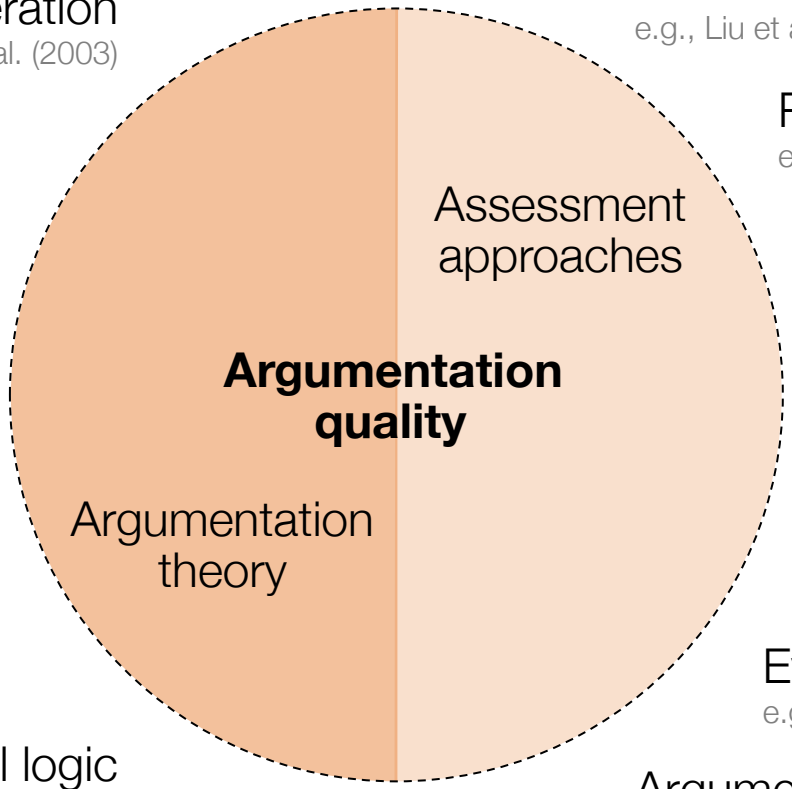
- **Debating Technologies at Dagstuhl**
  - Common understanding of argumentation quality missing
  - Working group to coordinate research

- **Research questions**
  - Can we unify the different views of quality?
  - Can we provide a common ground for quality assessment?

Core result


local acceptability
global acceptability
local relevance
global relevance
cogency
reason-ableness
local sufficiency
global sufficiency
Argumentation quality
clarity
arrangement
effectiveness
credibility
emotional appeal
appropriateness

Starting point



Deliberative quality
e.g., Gold et al. (2015)

Review helpfulness
e.g., Liu et al. (2008)

Review deception
e.g., Ott et al. (2011)

Wikipedia quality flaws
e.g., Anderka et al. (2012)

Readability
e.g., Pitler and Nenkova (2008)

Essay overall quality
e.g., Burstein et al. (1998)

Evidence types
e.g. Rinott et al. (2015)

Argumentation schemes
e.g. Feng and Hirst (2011)

Argument mining
e.g., Duthie et al. (2016)

Formal logic
e.g., Dung (1995)

Probabilistic logic
e.g., Pfeifer (2013)

Critical thinking
e.g. Freeley and Steinberg (2009)

Persuasion research
e.g., Zhao et al. (2011)

Deliberation
e.g., Steenbergen et al. (2003)

Survey of existing research

Toulmin (1958)

Walton et al. (2008)

Cabrio and Villata (2012)

van Eemeren and Grootendorst (2004)

Braunstain et al. (2016)

Hamblin (1970)

Walton (2006)

Boltužić and Šnajder (2015)

Tindale (2007)

Logic

Damer (2009)

Dialectic

Cohen (2011)

Rahimi et al. (2014)

Johnson and Blair (2006)

Wachsmuth et al. (2017)

Stab and Gurevych (2017)

Argumentation quality

Mercier and Sperber (2011)

Govier (2010)

Blair (2012)

van Eemeren (2015)

Freeman (2011)

Persing and Ng (2015)

Rahimi et al. (2015)

Persing and Ng (2013)

Perelman and Olbrecht-Tyteca (1969)

Persing et al. (2010)

Feng et al. (2014)

Rhetoric

Tan et al. (2016)

Wei et al. (2016)

Hoeken (2001)

O'Keefe and Jackson (1995)

Persing and Ng (2014)

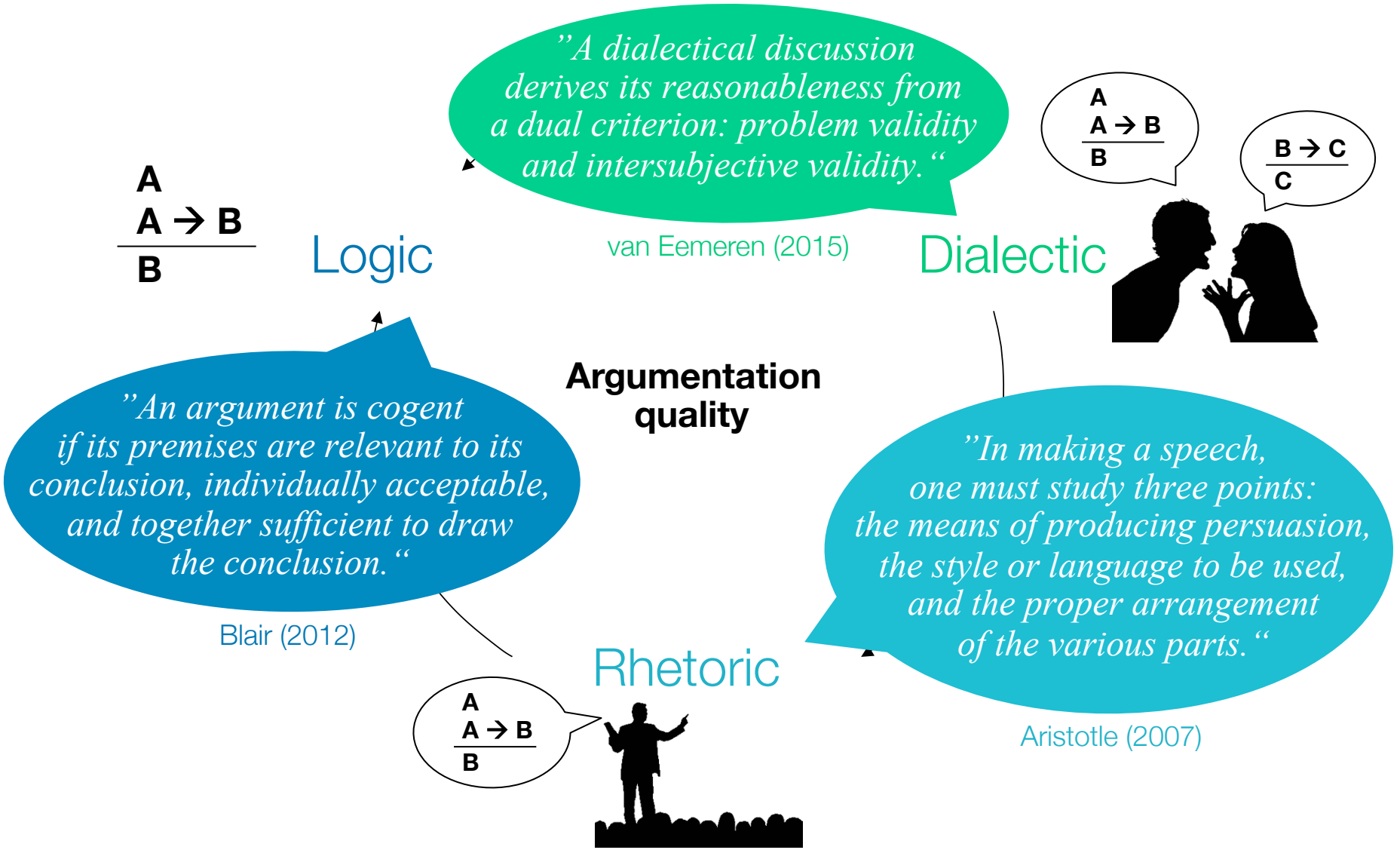
Zhang et al. (2016)

Aristotle (2007)

Park et al. (2015)

Habernal and Gurevych (2016)

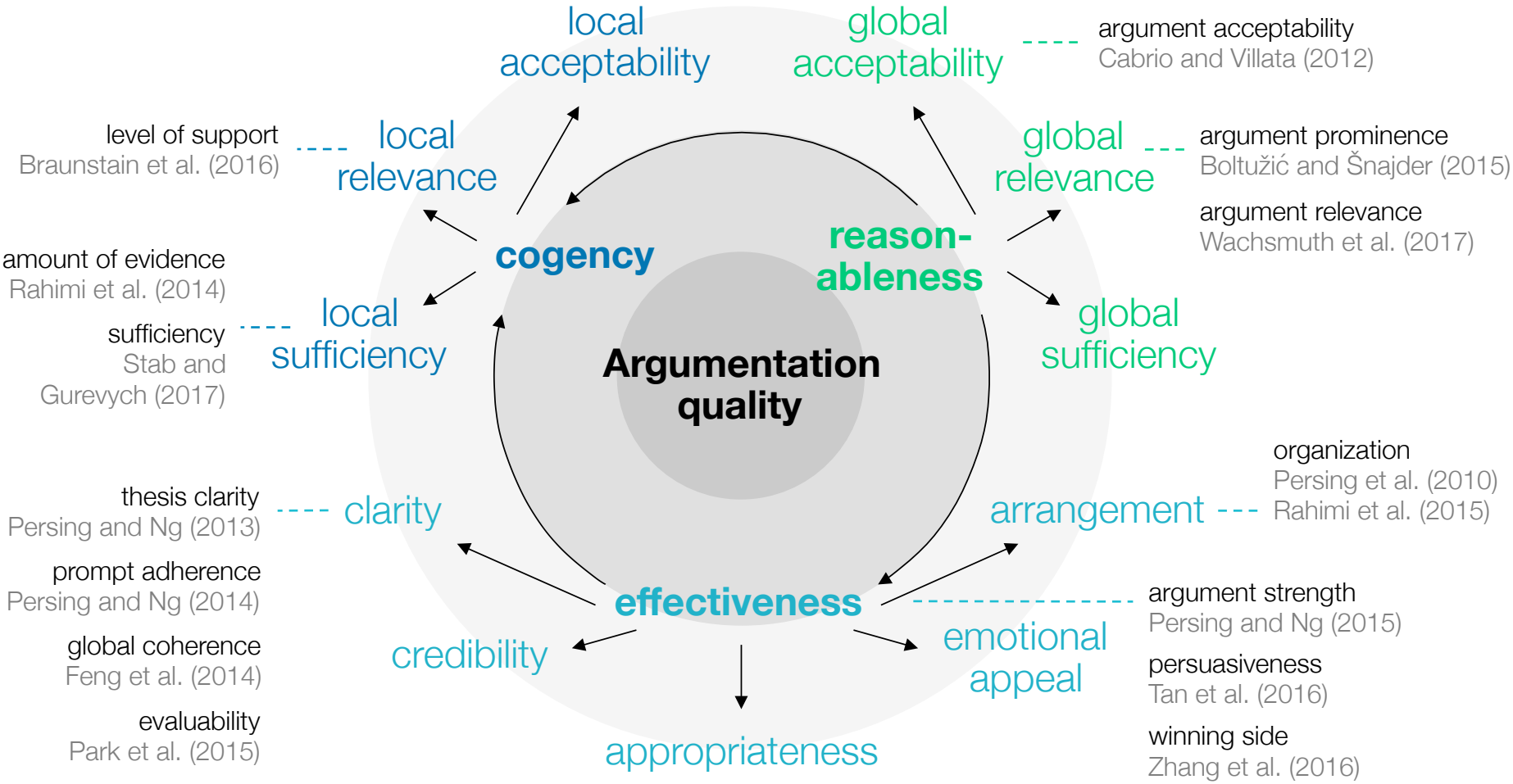
Three main quality aspects



Unification of views

validity
premise acceptability
intersubjective acceptability
argument acceptability
soundness
level of support
local/probative relevance
fallaciousness
global/dialectical relevance
argument relevance
Logic
amount of evidence
cogency
reason-ableness
Dialectic
prominence
premise sufficiency
amount of rebuttal
dialectical sufficiency
sufficiency
Argumentation quality
satisfac-toriness
well-formedness
strength
convincingness
global coherence
clarity of style
argument strength
winning side
arrangement
thesis clarity
organization
effectiveness
emotional appeal
prompt adherence
credibility
Rhetoric
evaluability
appropriateness of style
persuasiveness

A taxonomy of computational argumentation quality



The Dagstuhl-15512 ArgQuality corpus

available at www.arguana.com

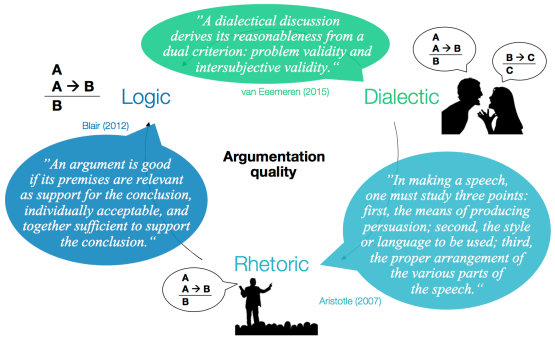
- **Corpus based on the taxonomy**
  - 320 debate portal arguments (Habernal and Gurevych, 2016)
  - 10 per issue/stance pair
  - 3 annotators per argument
  - Score from [1, 3] for all 15 dimensions
- **Agreement**
  - Krippendorff's alpha limited
  - Majority agreement very high
- **Correlations**
  - Overall quality correlates most with reasonableness (.86), cogency (.84), and effectiveness (.81)
  - Several other intuitive correlations

Dimension	Mean	Alpha	Maj.
cogency	**1.6**	**.44**	**92%**
local acceptability	1.9	.46	91%
local relevance	2.3	.47	92%
local sufficiency	1.5	.44	93%
effectiveness	**1.4**	**.45**	**94%**
credibility	1.7	.37	96%
emotional appeal	1.9	.26	94%
clarity	2.1	.35	90%
appropriateness	2.1	.36	88%
arrangement	1.8	.39	93%
reasonableness	**1.6**	**.50**	**96%**
global acceptability	1.9	.44	95%
global relevance	2.0	.42	90%
global sufficiency	1.2	.27	98%
overall quality	**1.6**	**.51**	**94%**

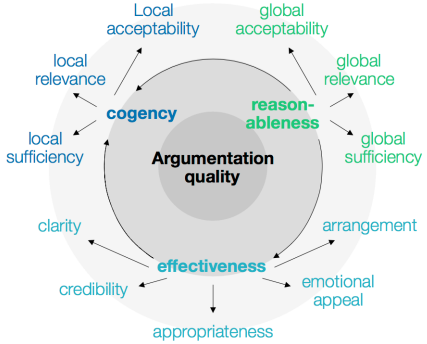
Contributions and outlook

- **Contributions**

Comprehensive survey



Unifying taxonomy



Annotated corpus

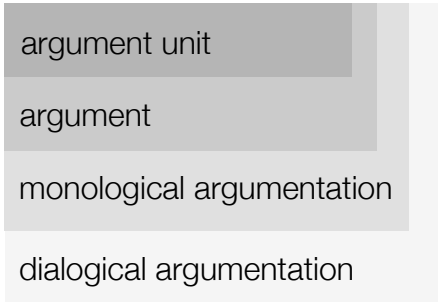
Dimension	ø	α	Maj.
cogency	**1.6**	**.44**	**92%**
local acceptability	1.9	.46	91%
local relevance	2.3	.47	92%
local sufficiency	1.5	.44	93%
effectiveness	**1.4**	**.45**	**94%**
credibility	1.7	.37	96%
emotional appeal	1.9	.26	94%
clarity	2.1	.35	90%
appropriateness	2.1	.36	88%
arrangement	1.8	.39	93%
reasonableness	**1.6**	**.50**	**96%**
global acceptability	1.9	.44	95%
global relevance	2.0	.42	90%
global sufficiency	1.2	.27	98%
overall quality	**1.6**	**.51**	**94%**

- **Outlook**

Reliable assessment

Target audience

Granularity levels



Theory vs. practice