Computational Argumentation — Part XI

Assessment of the Quality of Argumentation

Henning Wachsmuth
henningw@upb.de

June 19, 2019

PADERBORN UNIVERSITY

Outline

I. Introduction to computational argumentation
II. Basics of natural language processing
III. Basics of argumentation
IV. Applications of computational argumentation
V. Resources for computational argumentation

VI. Mining of argumentative units
VII. Mining of supporting and objecting units
VIII. Mining of argumentative structure
IX. Assessment of the structure of argumentation
X. Assessment of the reasoning of argumentation
XI. Assessment of the quality of argumentation
XII. Generation of argumentation
XIII. Development of an argument search engine

XIV. Conclusion

- Introduction
- A quality taxonomy
- Absolute rating
- Relative comparison
- Objective assessment
- Inclusion of subjectivity
- Conclusion

Learning goals

- **Concepts**
  - Get to know various quality dimensions of argumentation.
  - Learn about differences between quality in theory and in practice.
  - Understand the subjective nature of quality.

- **Methods**
  - Learn how to assess quality with supervised learning.
  - Learn how to assess quality through graph analyses.

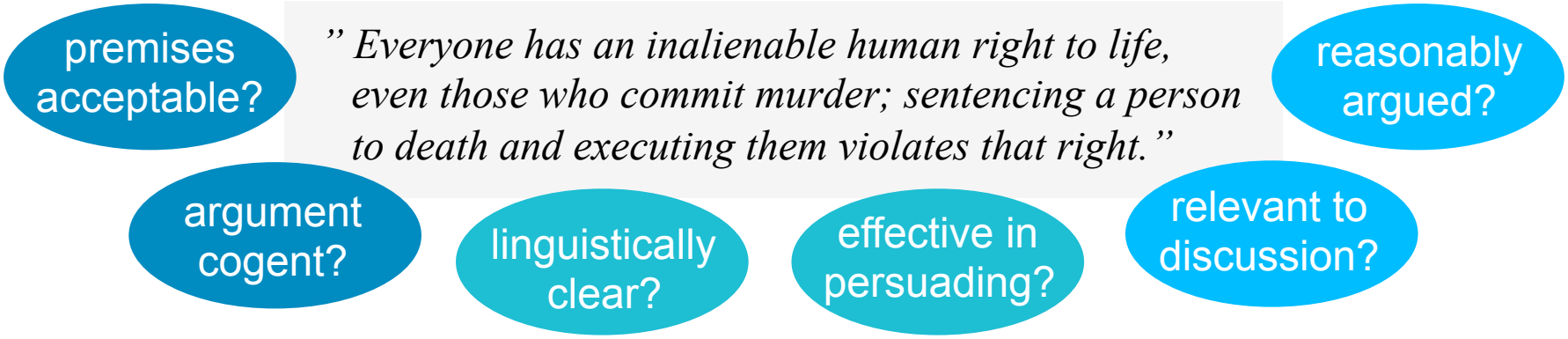
- **Associated research fields**
  - Argumentation theory and rhetoric
  - Computational linguistics

- **Within this course**
  - Understand how to distinguish good from bad arguments.
  - See to what extent computational assessment is doable currently.

Introduction

What is argumentation quality?

- **Argumentation quality**
  - Natural language argumentation is rarely logically *correct* or *complete.*
  - Need to measure how *good* an argument unit, argument, or argumentation is.



- **Observations**
  - Granularity. Quality may be addressed at different levels of text granularity.
  - Dimensions. Several dimensions of quality may be considered.
  - Goal orientation. What is important, depends on the goal of argumentation.

- **Notice**
  - The study of logical quality in terms of fallacies is beyond the score here.

Granularity levels of argumentation (recap)

Alice. *Some people say refugees threaten peace, as many of them were criminals.*
In fact, Spiegel Online just reported results from a study of the federal police about numbers of refugees and crimes:
Overall, there is no correlation at all!
Rather, the police confirmed that the main reason for committing crime is poverty.
So, if you believe the police then you shouldn't believe those people.
Syrians are even involved less in crimes than Germans according to the study.
So, the more Syrians come to Germany, the more peaceful it gets there, right?

Bob. *The question is here why should I believe the police!? Argument failed :P*

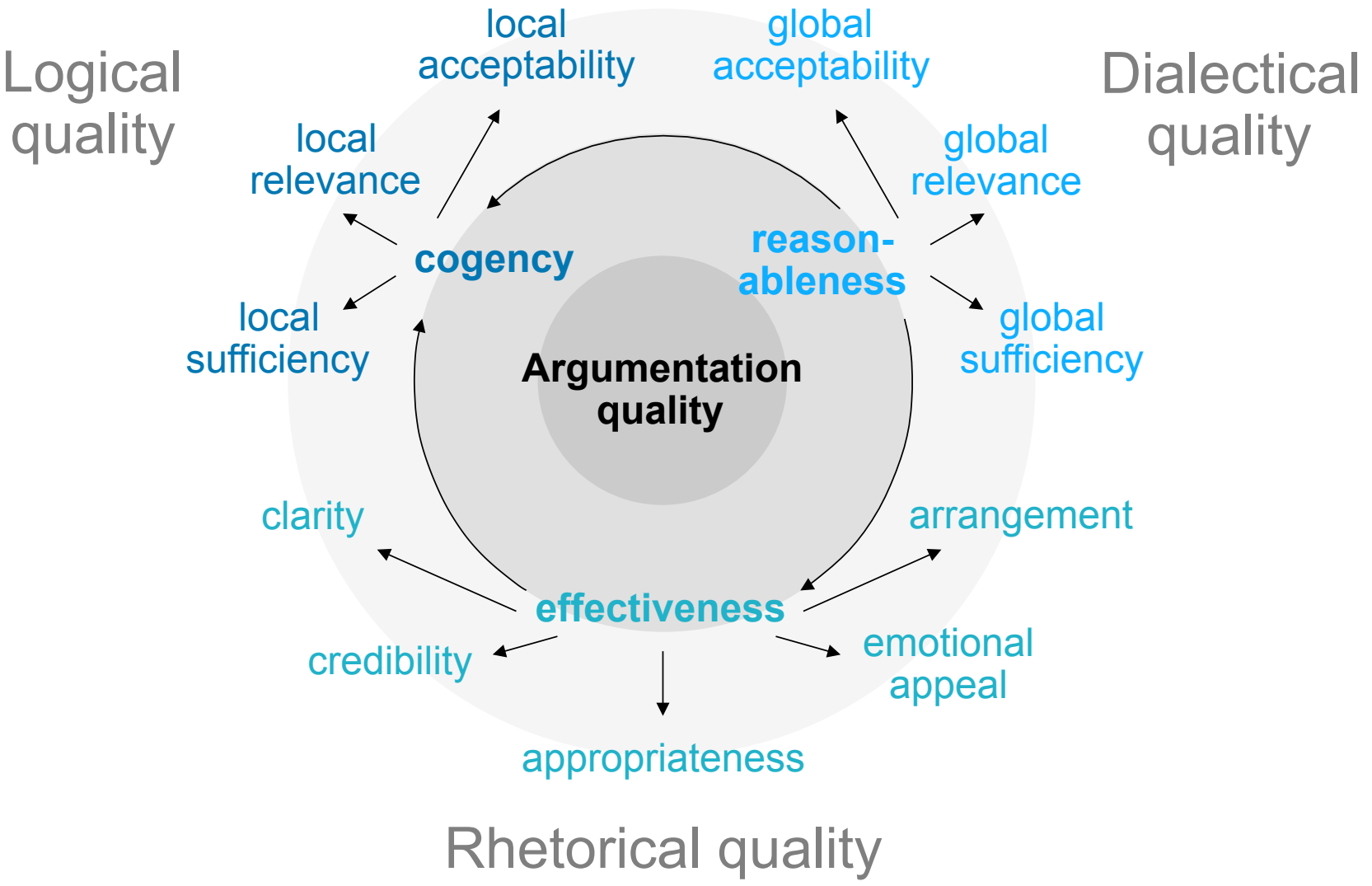
Argumentative discourse unit

Argument

Argumentation (monological)

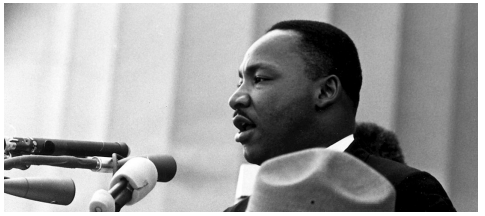
Debate (dialogical argumentation)

Argumentation quality dimensions (Wachsmuth et al., 2017b)



Goals of argumentation (recap) based on Tindale (2007)

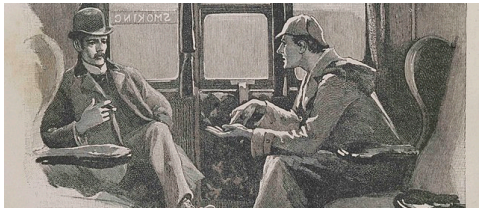
- **Persuasion**
  - Changing or reinforcing the stance of an audience towards an issue.
- **Agreement**
  - Resolving a dispute between multiple parties or achieving a settlement in a negotiation.
- **Justification**
  - Giving reasons or explanations for an attitude or action that might be controversial.
- **Recommendation**
  - Suggesting a decision to make, an action to take, a product to buy, or similar.
- **Deliberation**
  - Deepening one‘s own understanding of an issue.


https://de.wikipedia.org


https://commons.wikimedia.org


https://pixabay.com


https://pixabay.com



https://de.m.wikipedia.org

What is argumentation quality assessment?

- **Argumentation quality assessment**
  - Identification of indisputable flaws or requirements of argumentation.
  - Judgment about a specific quality dimension.
  - Determination whether argumentation successfully achieves its goal.

- **Observations**
  - Choice of comparison. Dimensions can be assessed *absolutely* or *relatively*.
  - Subjectivity. Perceived quality depends on the view of the reader/audience.
    (and maybe also on the author/speaker)

- **How to approach quality assessment**
  - Input. Argumentative text, metadata (e.g., author), external knowledge, ...
  - Techniques. Supervised classification/regression, graph-based analyses, ...
    Several example approaches discussed in this lecture.

Absolute vs. relative assessment

- **Two ways of assessing a quality dimension**
  - Absolute rating. Assignment of a score from a predefined scale.
    Typical scales: Integers (possibly with half-points): 1–3, 1–4, 1–5, 1–10, -2–2, ... Real valued: [0,1], [-1,1]
  - Relative comparison. Given two instances, which of them is better.

> *"If you wanna hear my view I think that the death penalty should be abolished. It legitimizes an irreversible act of violence. As long as human justice remains fallible, the risk of executing the innocent can never be eliminated."*

4 / 5

better than

> *"Human beings never act freely and thus should not be punished for even the most horrific crimes."*

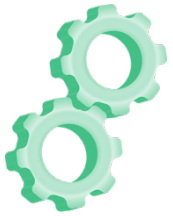
- **Observations**
  - Both allow for ranking the assessed instances.
  - Absolute ratings entail relative comparisons.
  - Absolute ratings imply a maximum and minimum.
- **Absolute vs. relative assessment**
  - A relative assessment is often much easier.
  - Still, absolute ratings are widely spread and often work well.

Argumentation quality in theory and in practice

- **Quality in theory**
  - The normative view of quality in terms of cogency, reasonableness, or similar.
  - Suggests to use absolute quality ratings.

https://commons.wikimedia.org

- **Quality in practice**
  - Quality is decided by the effectiveness on (some type of) people.
  - Relative comparisons are often more suitable.



https://de.wikipedia.org

> *” Is a strong argument an effective argument which gains the adherence of the audience, or is it a valid argument, which ought to gain it?“*

(Perelman and Olbrechts-Tyteca, 1969)

- **Unresolved questions**
  - Should quality be aligned with how we *should* or how with we *do* argue?
  - Is this actually so different? → more on this below

The role of participants in argumentation (recap)

- **Author (or speaker)**
  - Argumentation is connected to the person who argues.
  - The same argument is perceived differently depending on the author.
- **Reader (or audience)**
  - Argumentation often targets a particular audience.
  - Different arguments and ways of arguing work for different readers.

”University education must be free. That is the only way to achieve equal opportunities for everyone.“

”According to the study of XYZ found online, avoiding tuition fees is beneficial in the long run, both socially and economically.“

https://pixabay.com

https://commons.wikimedia.org

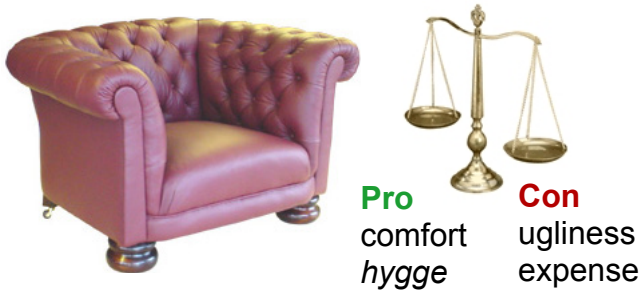
https://pixabay.com

- **Questions**
  - May the assessment ignore the author/speaker? And the reader/audience?
    The author/speaker is unknown in some application scenarios, but rarely the reader/audience is.

Subjective (and objective) assessment

- **Subjectiveness of quality assessment**
  - Many dimensions are inherently subjective.
  - Quality depends on the subjective weighting of different aspects of an issue.
  - Also depends on preconceived opinions.

”Should we buy a Chesterfield armchair?”



(credit to Christian Kock for this example)

example and figures from https://argmining2017.files.wordpress.com/2016/12/argmining2017-invited-talk-christian-kock.pptx

- **Example: Which argument is more relevant?**

” The death penalty legitimizes an irreversible act of violence. As long as human justice remains fallible, the risk of executing the innocent can never be eliminated.”

” The death penalty doesn’t deter people from committing serious violent crimes. The thing that deters is the likelihood of being caught and punished.”

- **Two ways to approach this problem** (both will be detailed below)
  - Either, focus on properties that can be assessed ”objectively“.
  - Or, include a model of the reader/audience in the quality assessment.

Importance of quality assessment

- **Why assessing argumentation quality?**
  - Mining arguments and understanding the reasoning is not enough in practice.
  - For successful argumentation, we need to choose the ”best“ arguments.
  - Critical for any application of computational argumentation.

> *”In some, sense the question about the quality of an argument is the ‘ultimate’ one for argumentation mining.“*

(Stede and Schneider, 2018)

- **Example applications**
  - Argument search. What argument to rank highest?
  - Writing support. How good is an argumentative text, what flaws does it have?
  - Automatic decision making. Which arguments outweigh which others?

A quality taxonomy

based on Wachsmuth et al. (2017b)

Survey of existing research

argumentation theory

assessment approaches

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. (2017a)

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)

Hoeken (2001)

Rhetoric

Tan et al. (2016)

Wei et al. (2016)

Persing and Ng (2014)

O'Keefe and Jackson (1995)

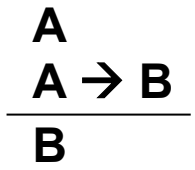
Zhang et al. (2016)

Park et al. (2015)

Aristotle (2007)

Habernal and Gurevych (2016)

Three main quality aspects

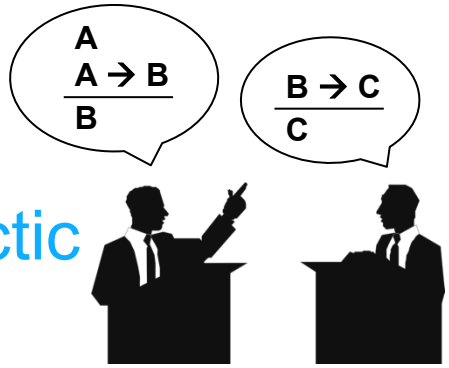


Logic

*"A dialectical discussion derives its **reasonableness** from a dual criterion: **problem validity** and **intersubjective validity**."*

van Eemeren (2015)

Dialectic



https://de.wikipedia.org

Argumentation quality

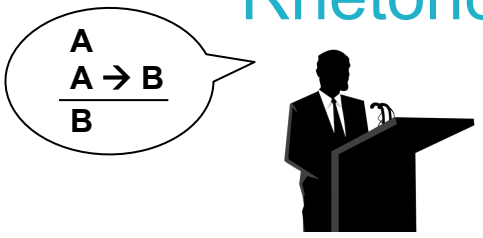
*"An argument is **cogent** if its premises are **relevant** to its conclusion, individually **acceptable**, and together **sufficient** to draw the conclusion."*

Blair (2012)

*"In making a speech, one must study three points: the **means of producing persuasion**, the **style** or language to be used, and the proper **arrangement** of the various parts."*

Aristotle (2007)

Rhetoric

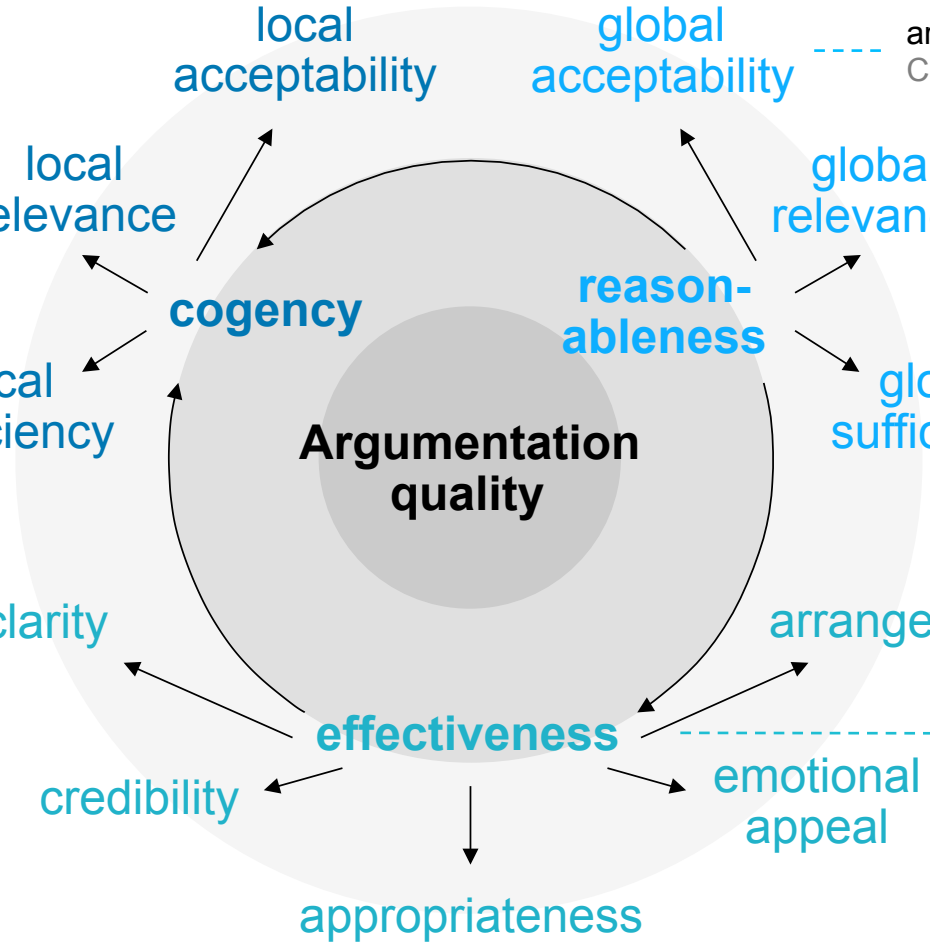


https://commons.wikimedia.org

Unification of views

validity
premise acceptability
intersubjective acceptability
argument acceptability
soundness
level of support
local/probative relevance
fallaciousness
global/dialectical relevance
argument relevance
Logic
reason-ableness
Dialectic
amount of evidence
cogency
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 argumentation quality



argument acceptability
Cabrio and Villata (2012)

level of support
Braunstain et al. (2016)

argument prominence
Boltužić and Šnajder (2015)

argument relevance
Wachsmuth et al. (2017a)

amount of evidence
Rahimi et al. (2014)

sufficiency
Stab and Gurevych (2017)

thesis clarity
Persing and Ng (2013)

prompt adherence
Persing and Ng (2014)

global coherence
Feng et al. (2014)

evaluability
Park et al. (2015)

organization
Persing et al. (2010)
Rahimi et al. (2015)

argument strength
Persing and Ng (2015)

persuasiveness
Tan et al. (2016)

winning side
Wang et al. (2016)
Zhang et al. (2016)

convincingness
Habernal and Gurevych (2016)

Quality dimensions in the taxonomy

- **A cogent argument.** Has acceptable, relevant, and sufficient premises.
  - Local acceptability. The premises are worthy being believed as true.
  - Local relevance. The premises are relevant to the conclusion.
  - Local sufficiency. The premises are sufficient to draw the conclusion.

Logic

- **Effective argumentation.** Persuades the target audience.
  - Credibility. Makes the authors worthy of credence.
  - Emotional appeal. Makes the audience open to be persuaded.
  - Clarity. Is linguistically clear and as simple as possible.
  - Appropriateness. Linguistically matches the audience and issue.
  - Arrangement. Presents content in the right order.

Rhetoric

- **Reasonable argumentation.** Is acceptable, relevant, and sufficient.
  - Global acceptability. Worthy to be considered in the way stated.
  - Global relevance. Contributes to resolution of issue.
  - Global sufficiency. Adequately rebuts potential counterarguments.

Dialectic

Notice: cogency also adds to effectiveness, and cogency and effectiveness also add to reasonableness.

The Dagstuhl-15512 ArgQuality corpus

- **Corpus based on the taxonomy**
  - 320 debate portal arguments (Habernal and Gurevych, 2016a)
  - 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%**

Absolute rating

Absolute quality rating: Overview

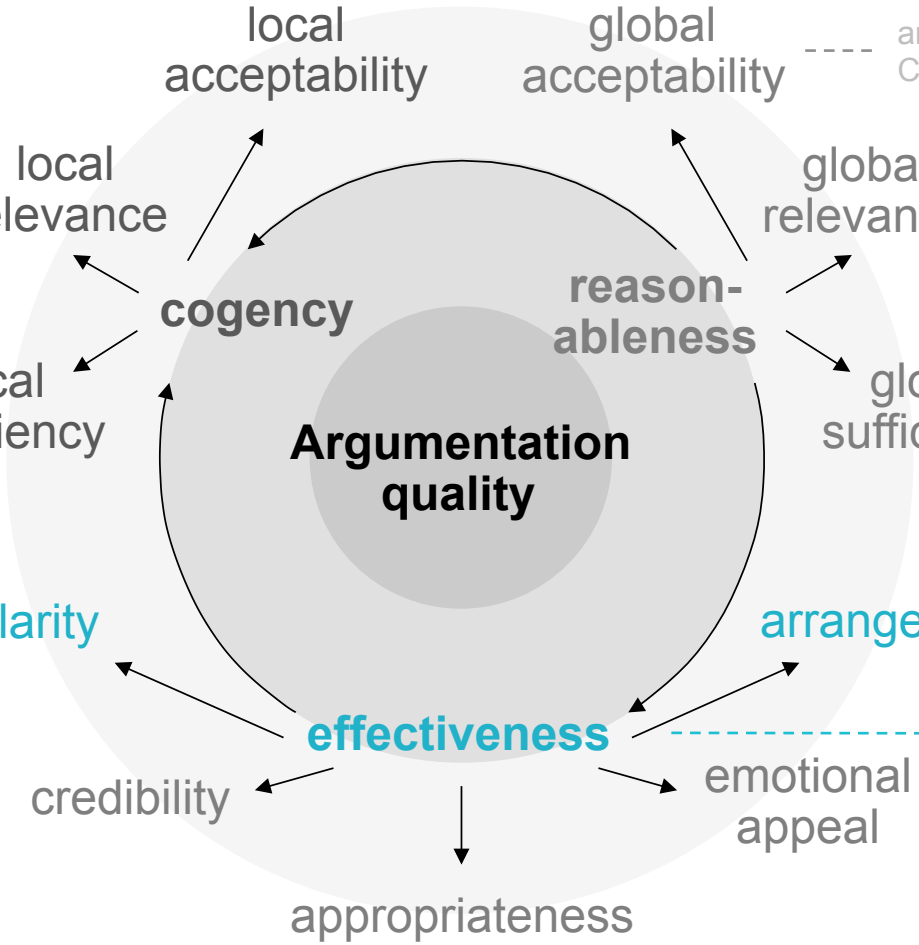
- **Problem**
  - Can we predict *whether* an argument(ation) is good (cogent, effective, ...)?
  - Can we rate *how* good it is?

- **Main idea**
  - See quality assessment as a standard classification or regression task.
  - Learn what linguistic feature or metadata speaks for quality?

Conclusion / Premises

- **Existing approaches**
  - Persuasiveness. Prediction based on interaction of participants. (Tan et al., 2016)
  - Organization. Assessment based on tuned features. (Persing et al., 2010)
    Analog approaches for thesis clarity, prompt adherence, and argument strength (Persing and Ng, 2013–2015).
  - Amount of evidence. Count of evidence supporting conclusion. (Rahimi et al., 2014)
  - Sufficiency. Prediction using convolutional neural networks (Stab and Gurevych, 2017).
    ... among other approaches

Absolute rating: Covered dimensions



argument acceptability
Cabrio and Villata (2012)

level of support
Braunstain et al. (2016)

argument prominence
Boltužić and Šnajder (2015)

argument relevance
Wachsmuth et al. (2017a)

amount of evidence
Rahimi et al. (2014)

sufficiency
Stab and Gurevych (2017)

organization
Persing et al. (2010)
Rahimi et al. (2015)

thesis clarity
Persing and Ng (2013)

prompt adherence
Persing and Ng (2014)

global coherence
Feng et al. (2014)

evaluability
Park et al. (2015)

argument strength
Persing and Ng (2015)

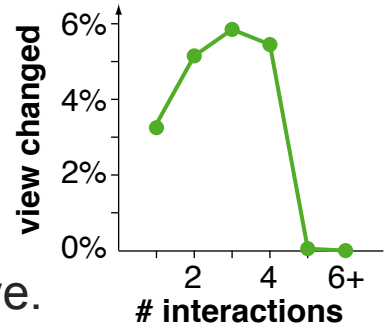
persuasiveness
Tan et al. (2016)

winning side
Wang et al. (2016)
Zhang et al. (2016)

convincingness
Habernal and Gurevych (2016)

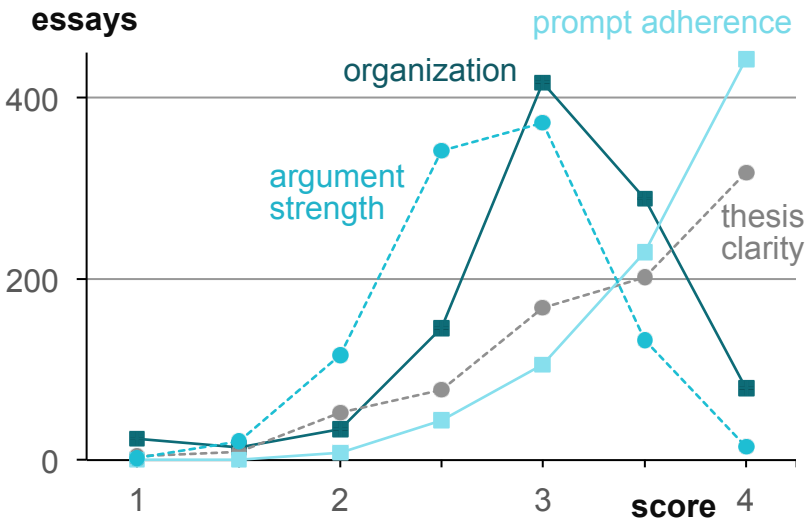
Absolute rating of effectiveness (Tan et al., 2016)

- **Task**
  - In a discussion, what will persuade someone open to be persuaded?
- **Approach**
  - Analysis of correlations between linguistic, interaction, and meta-discussion features with persuasion.
  - Prediction based on features as to whether persuasion will happen.
- **Data**
  - 20k+ discussions from Reddit ChangeMyView.
  - Discussion. An opinion poster (OP) states a view, others argue against, OP gives Δ to convincing arguments.
- **Selected results**
  - Accuracy. 69% in balanced setting.
  - Insights. Some interactions and many participants help; appropriate style, not to similar to OP's style most persuasive.

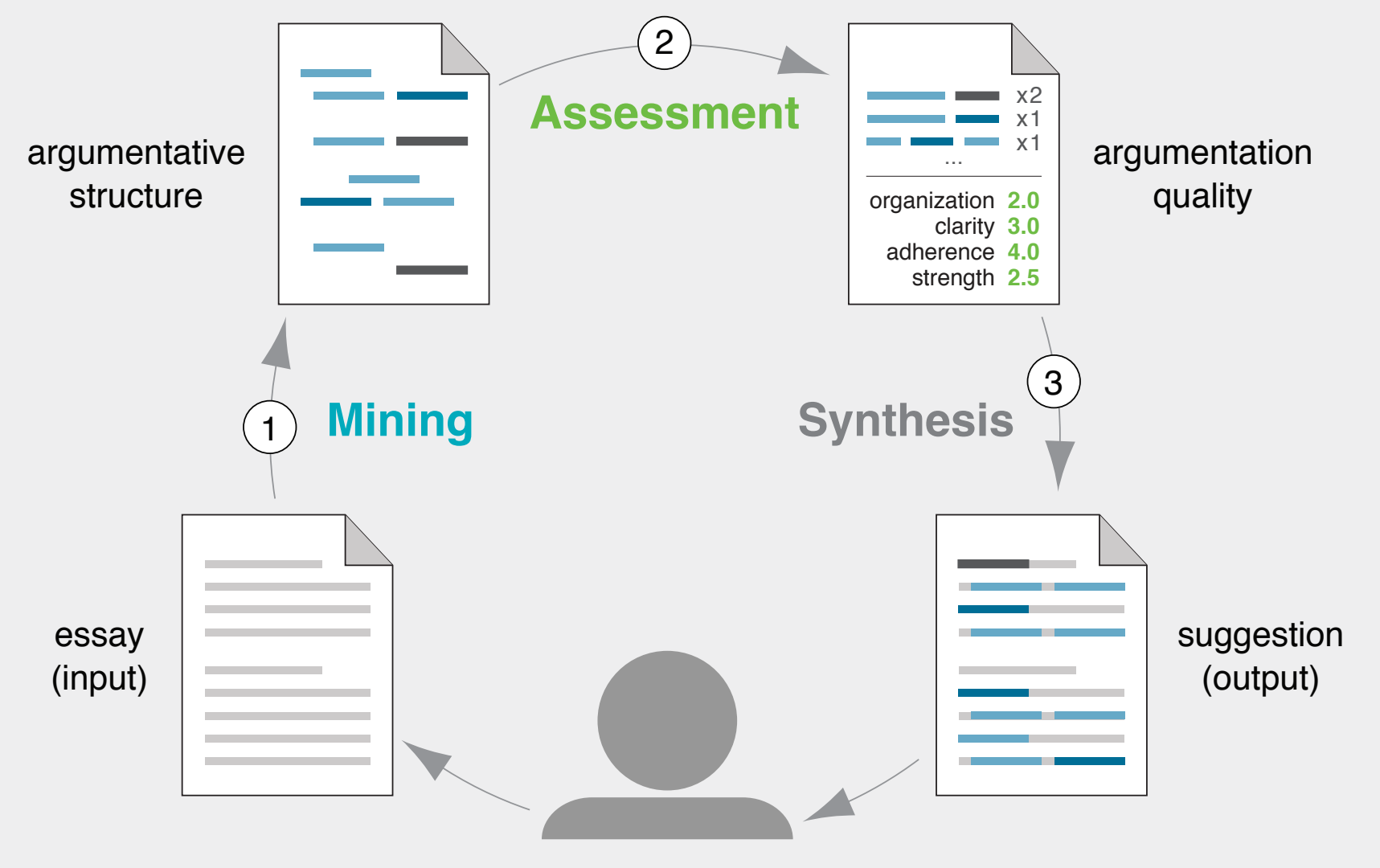


Absolute rating of four rhetorical dimensions (Wachsmuth et al., 2016)

- **Task**
  - Given a persuasive essay, rate argumentation-related quality dimensions.
- **Dimensions**
  - Organization. How well is the essay‘s argumentation arranged?
  - Thesis clarity. How easy to understand is the essay‘s thesis?
  - Prompt adherence. How close does the essay stay to the prompt?
  - Argument strength. How strong is the argument made for the thesis?
- **Research question**
  - Can we leverage argument mining to assess the argumentation quality of persuasive essays?
- **Data** (Persing et al., 2010; Persing and Ng, 2013–2015)
  - 800–1003 essays with scores from [1,4] annotated for each dimension



Motivation: Argumentative writing support (Wachsmuth et al., 2016)



Shallow mining of argumentative structure (Wachsmuth et al., 2016)

- **Mining of argument units**
  - Task. Classify sentence-level units as thesis, conclusion, premise, or none.
  - Approach. Support vector machine (SVM) with different standard features.
  - Data. AAE corpus (Stab and Gurevych, 2014)
  - Results. Comparable to state of the art.

Approach	Acc.	F_1
Majority baseline	52.5	36.1
State of the art	**77.3**	72.6
Our classifier	74.5	**74.5**

- **Analysis of mined argumentative structure**
  - Task. Mine and analyze common unit type flows (consider changes only).
  - Data. All paragraphs of full ICLE corpus (6085 student essays). (Granger et al., 2009)
  - Insights. Some flows very common, 1st and last flow in text differ entirely.

Unit type flows	Average	First	Last
Conclusion, Premises	25.1%	–	13.1%
Conclusion, Premises, Conclusion	17.0%	–	27.2%
None, thesis	3.4%	25.9%	–
Premises, Conclusion	2.9%	–	2.7%

Example essay with mined structure (Wachsmuth et al., 2016)

- **Prompt**

"Some people say that in our modern world, dominated by science and technology and industrialisation, there is no longer a place for dreaming and imagination. What is your opinion?"

Organization	3.0
Thesis clarity	2.0
Prompt adherence	4.0
Argument strength	2.0

- **Essay**

None

"If we take a look back in time we are in a position to see man dreaming, philosophizing and using his imagination of whatever comes his way. We see man transcending his ego I a way and thus becoming a God - like figure. And by putting down these sacred words, what is taking shape in my mind is the fact that using his imagination Man is no longer this organic and material substance like his contemporary counterpart who is putting his trump card on science, technology and industrialization but Man is a way transcends himself through his imagination.

Conclusion

For instance, if we take into account the Renaissance or Romantic periods of mankind and close our eyes we could see Shakespeare applying his imagination in the fancy world of his comedies: elf and nymphs circling the stage making it a dream that will lost forever in our minds. We could even hear their high-pitched weird chuckle piercing with a gentle touch our ears, but "open those eyes that must eclipse the day" and you'll wee the high-tech wiping out every trace of the human elevated spirit that have dominated over the previous centuries. What we see now is "deux aux machina" or the fake "God from the machine" who with the touch of a button could unleash Armageddon.

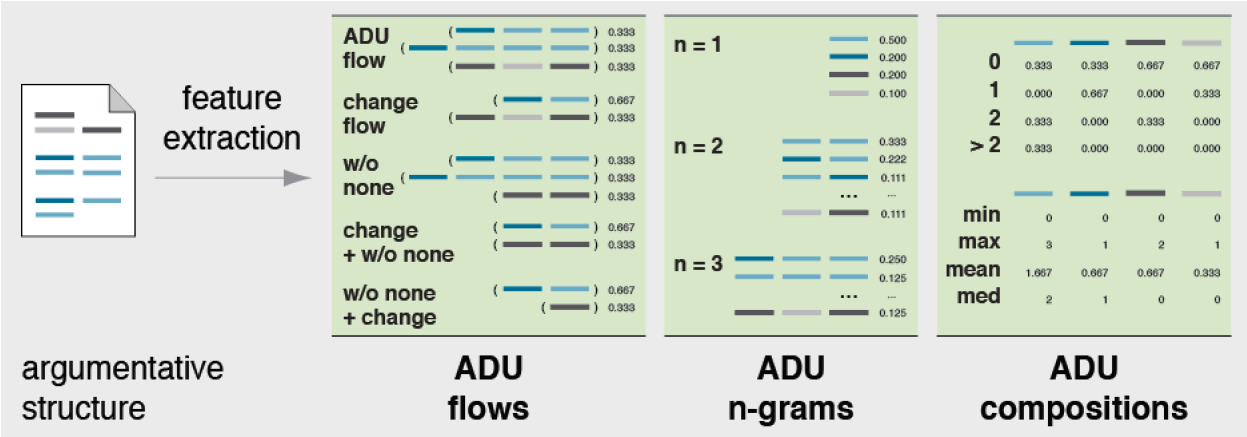
Premise

For poets and literate people of yore it was a common idea to transcend reality or to go beyond it by using their imagination not by using reason as we the homosapiens of our time do. For example, if we indulge in entertaining the idea of the film "The matrix" it has a lot to do with the period of Romanticism. But the difference is that a poet from that time could transcend reality, become one with Nature, and cruise wherever he wants using his imagination. Whereas now in the 21st century and in "The matrix" in particular the scientific type of Man thinks that at last he has succeeded in making travelling without boundaries via the virtual reality of his PC.

As a logical conclusion to my essay I would like to put only one thing. 'Wouldn't it be better if imagination makes the world go round'. If I was to answer this question, the answer would be positive, but given the aquisitive or consumer society conditions we live in let's make a match between imagination and science. It would be somewhat more realistic."

Assessment of argumentation quality (Wachsmuth et al., 2016)

- **Quality assessment based on structure**
  - Approach. SVM based on standard and argument-specific features.



- **Evaluation**
  - Results. Lowest mean squared error for the structure-related dimensions.
  - Insights. Best feature type captures composition of argument units.

Approach	Organization	Clarity	Adherence	Strength
Average baseline	0.349	0.469	0.291	0.266
Previous state of the art	0.175	**0.369**	**0.197**	0.244
Our approach	**0.164**	0.425	0.216	**0.226**

Relative comparison

Relative quality comparison: Overview

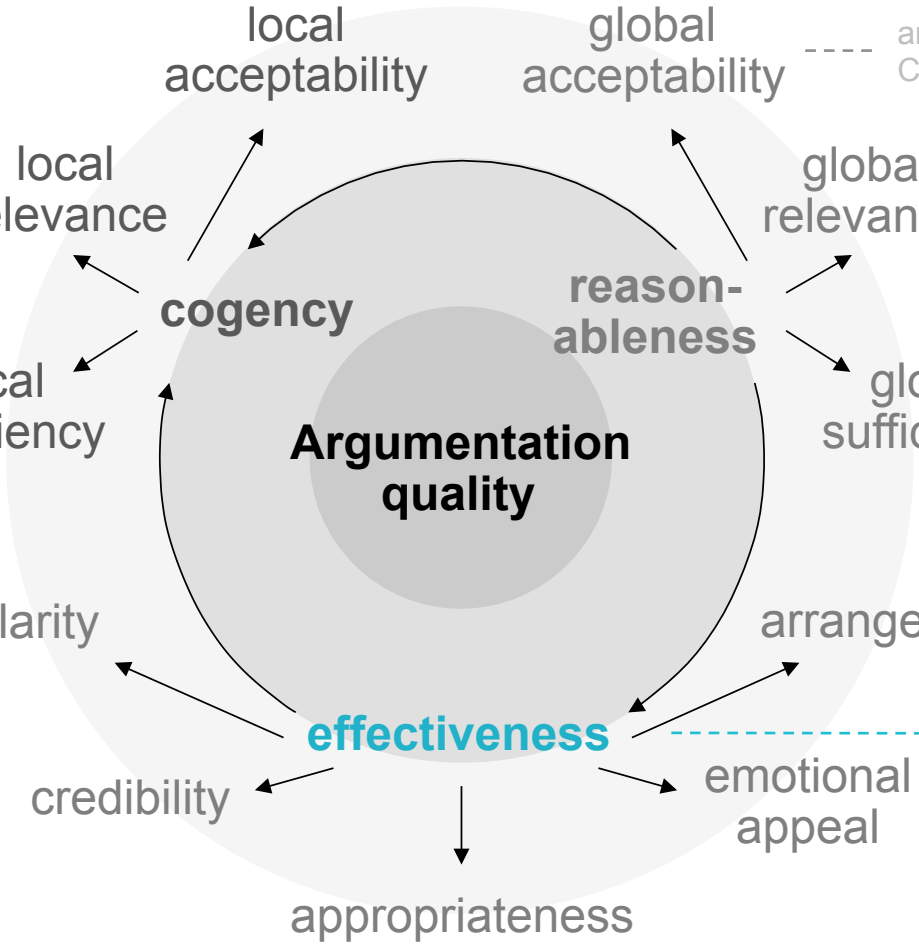
- **Problem**
  - Rating the quality of an argument in isolation may be hard or even doubtful.
  - Is there an easier or more realistic way to assess quality?

- **Main idea**
  - Often, we are only interested in the best available argument.
  - It‘s enough to compare the quality of an argument to others.
  - Dilemma. Unclear in the end whether the best argument is good.

$$\frac{\text{Conclusion}}{\text{Premises}} \quad \textbf{vs} \quad \frac{\text{Conclusion}}{\text{Premises}}$$

- **Existing approaches**
  - Winning side. Prediction of the debate winner from debate flow. (Zhang et al., 2016)
  - Winning side. Prediction of the winner from content and style (Wang et al., 2016)
  - Convincingness. Argument comparison with standard supervised learning. (Habernal and Gurevych, 2016a)
  - Level of support. Ranking of arguments by support of claim. (Braunstain et al., 2016)

Relative quality comparison: Covered dimensions



argument acceptability
Cabrio and Villata (2012)

level of support
Braunstain et al. (2016)

argument prominence
Boltužić and Šnajder (2015)

argument relevance
Wachsmuth et al. (2017a)

amount of evidence
Rahimi et al. (2014)

sufficiency
Stab and Gurevych (2017)

organization
Persing et al. (2010)
Rahimi et al. (2015)

thesis clarity
Persing and Ng (2013)

prompt adherence
Persing and Ng (2014)

global coherence
Feng et al. (2014)

evaluability
Park et al. (2015)

argument strength
Persing and Ng (2015)

persuasiveness
Tan et al. (2016)

winning side
Wang et al. (2016)
Zhang et al. (2016)

convincingness
Habernal and Gurevych (2016)

Relative comparison of effectiveness (Zhang et al. 2016)

- **Task**
  - Given a full Oxford-style debate, which opponent wins?

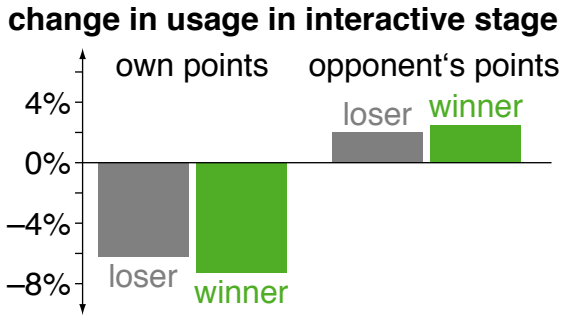
- **Approach**
  - Mining of supporting points each side.
  - Modeling of the ”conversational flow“: When does a side puts forward own points, when does it attack opponent points.
  - Logistic regression classifier with features capturing the flow.

”Millennials don’t stand a chance“

debt college boomer **pro** reality

economy volunteer home **con** engage

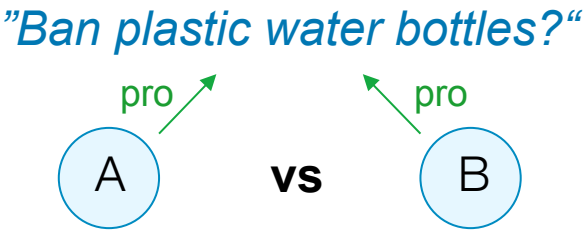
- **Data**
  - 108 Intelligence^2 debates (117 turns on average).
  - Winning side and audience feedback given.



- **Results**
  - Accuracy. Approach (0.65) beats audience feedback (0.6).
  - Insights. Attacking the opponent’s points better than focus on own points.

Relative comparison of effectiveness (Habernal et al., 2016a)

- **Task**
  - Given two arguments with the same topic and stance, which one is more convincing?



- **Supervised learning approaches**
  - SVM. SVM with RBF kernel and a rich set of linguistic features.
  - BiLSTM. Bi-directional long short-term memory neural network using GloVe.

  Notice: The focus of the paper was not the approaches but the data construction.

- **Crowdsourced data**
  - 16,927 pairs of 1052 debate portal arguments for 32 topic-stance pairs.
  - Each annotated 5 times for convincingness (most reliable annotation taken).

  Reliability can be estimated with MACE (Hovy et al., 2013). Annotators also had to give reasons.

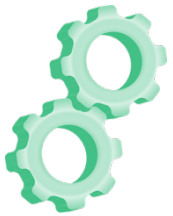
- **Results in 32-fold cross-validation**
  - Accuracy. SVM (0.78) beats BiLSTM (0.76). Human performance 0.93.
  - Insights. Surface features like capitalization easy, ”inverted“ sentiment hard.

Absolute vs. relative assessment ~ Theory vs. practice

- **Data representing theory**
  (Wachsmuth et al., 2017b)
  - Absolute expert ratings
  - Normative guidelines
  - 15 predefined quality dimensions

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

- **Data representing practice**
  (Habernal and Gurevych, 2016b)
  - Relative lay comparisons
  - No guidelines
  - 17+1 resulting reason labels



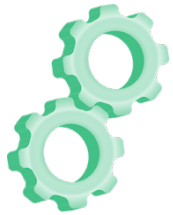
off-topic
no credible evidence
only opinion
language/grammar issues
attacking/abusive
unclear/hard to follow
insufficient reasoning
irrelevant reasons
makes you think
convincing
close to topic
non-sense/confusing
crisp / well-written
generally weak/vague
credible / confident
objective/two-sided
well thought through
details/facts/examples

- **Empirical comparison of theory and practice**
  (Wachsmuth et al., 2017d)
  - 736 argument pairs are available with ratings *and* labels.
  - Compute Kendall's τ correlations of all dimensions and reasons.

How different is assessment in theory and in practice?

- **Selected insights**
  - Convincing correlates most with **overall quality** (0.64).
  - Generally high ”correlations“ between 0.3 and 1.0.
  - Perfect: Global acceptability + attacking/abusive (1.0).
  - Mostly very intuitive, such as clarity + unclear (0.91).

vs



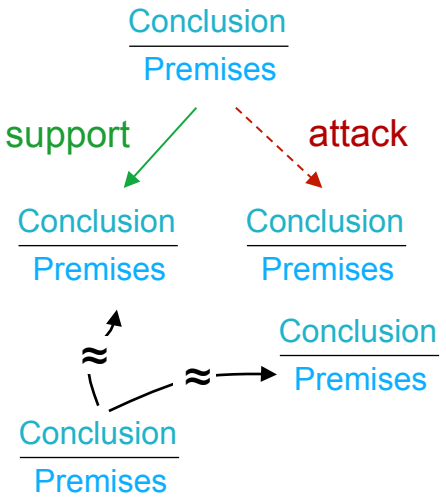
  - Top **overall quality** for well thought through (mean score 1.8 of 3).
  - Lowest **overall quality** for off-topic (mean score 1.1 of 3).
  - Few unintuitive results, e.g., ”only“ 0.52 for credibility + no credible evidence.
  - Local sufficiency + global sufficiency hard to separate.
- **Conclusions**
  - Theory and practice match more than expected.
  - Theory can guide quality assessment in practice.
  - Practice indicates what to focus on to simplify theory.

Objective assessment

Objective quality assessment: Overview

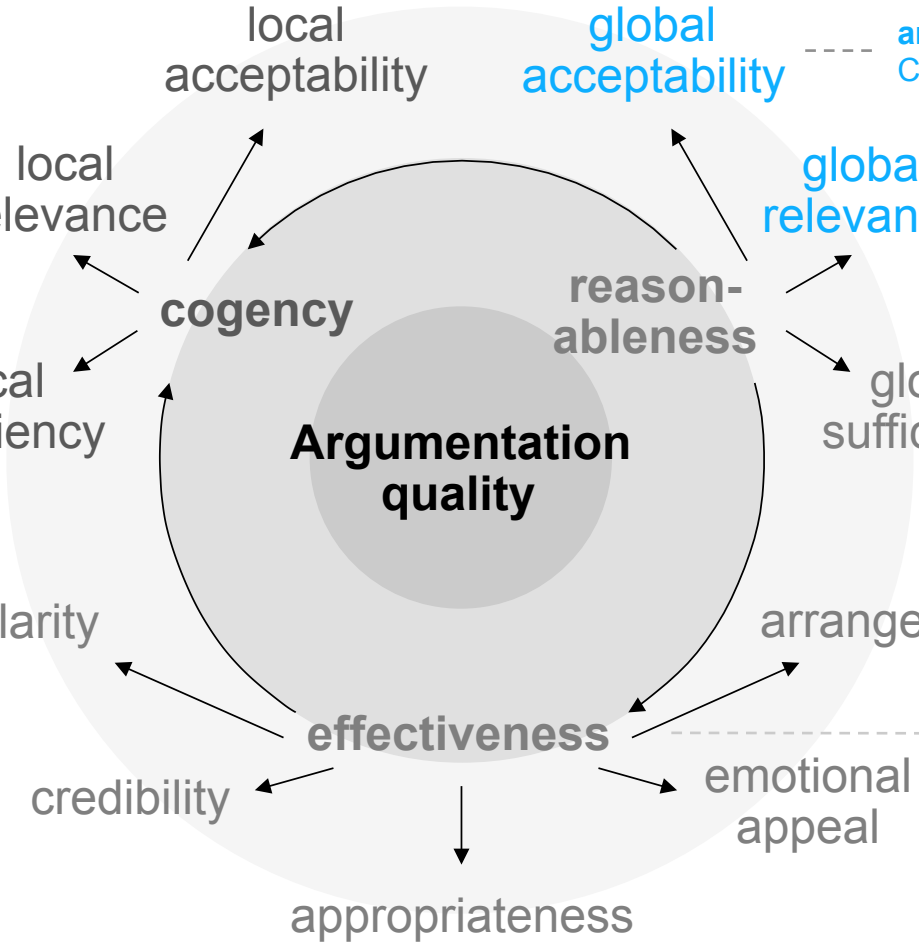
- **Problem**
  - How to assess quality without learning from subjective annotations?
  - What are objective argumentation quality indicators?

- **Main idea**
  - Assess quality based on the structure induced by the set of all arguments.
  - Works for both for absolute and relative assessment.
  - Dilemma. Evaluation on subjective annotations?
    A solution is to rely on majority assessments of many annotators.



- **Existing approaches**
  - Acceptability. Assessment based on the attack relations. (Cabrio and Villata, 2012)
  - Relevance. Assessment based on reuse of argument units. (Wachsmuth et al., 2017a)
  - Prominence. Assessment based on argument frequency. (Boltužic and Šnajder, 2015)

Objective quality assessment: Covered dimensions



argument acceptability
Cabrio and Villata (2012)

level of support
Braunstain et al. (2016)

argument prominence
Boltužić and Šnajder (2015)

argument relevance
Wachsmuth et al. (2017a)

amount of evidence
Rahimi et al. (2014)

sufficiency
Stab and Gurevych (2017)

organization
Persing et al. (2010)
Rahimi et al. (2015)

thesis clarity
Persing and Ng (2013)

prompt adherence
Persing and Ng (2014)

global coherence
Feng et al. (2014)

evaluability
Park et al. (2015)

argument strength
Persing and Ng (2015)

persuasiveness
Tan et al. (2016)

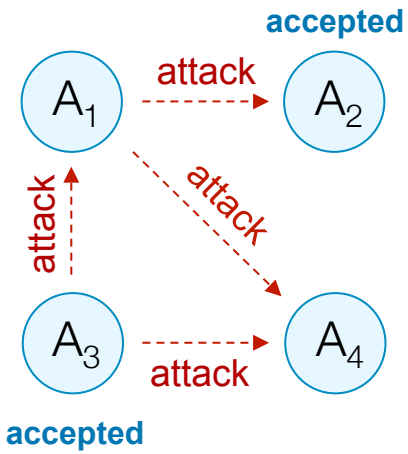
winning side
Wang et al. (2016)
Zhang et al. (2016)

convincingness
Habernal and Gurevych (2016)

Objective assessment of global acceptability

- **Background: Abstract argumentation framework** (Dung, 1995)
  - A directed graph where nodes represent arguments and edges attack relations between arguments.
  - Graph analysis reveals whether to accept an argument.
  - Accepted. If all arguments attacking it are rejected.
  - Not accepted. If an accepted argument attacks it.

  Extensions with weightings and with support+attack exist.



- **Approach** (Cabrio and Villata, 2012)
  - Given a set of arguments, use textual entailment algorithm to classify attacks.
  - Assess acceptability of arguments following Dung‘s framework.

- **Evaluation**
  - Tested on 100 argument pairs from idebate.org, 45 attacking each other.
  - Attack classification. Accuracy 0.67
  - Acceptability assessment. Accuracy 0.75

Objective assessment of global relevance (Wachsmuth et al., 2017a)

- **Task**
  - Given a set of arguments, which one is most relevant to some issue?
  - Problem. Relevance is highly subjective.

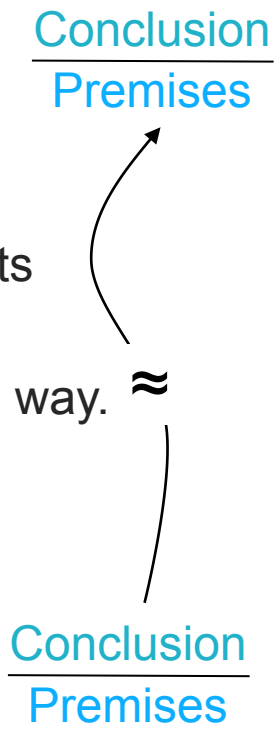
> *" The death penalty legitimizes an irreversible act of violence. As long as human justice remains fallible, the risk of executing the innocent can never be eliminated."*

> *" The death penalty doesn't deter people from committing serious violent crimes. The thing that deters is the likelihood of being caught and punished."*

- **Research question**
  - Can we develop an "objective" measure of relevance?

- **Key hypothesis**
  - The relevance of a conclusion depends on what other arguments across the web use it as a premise.
  - Rationale. Author cannot control who "cites" a conclusion in this way.

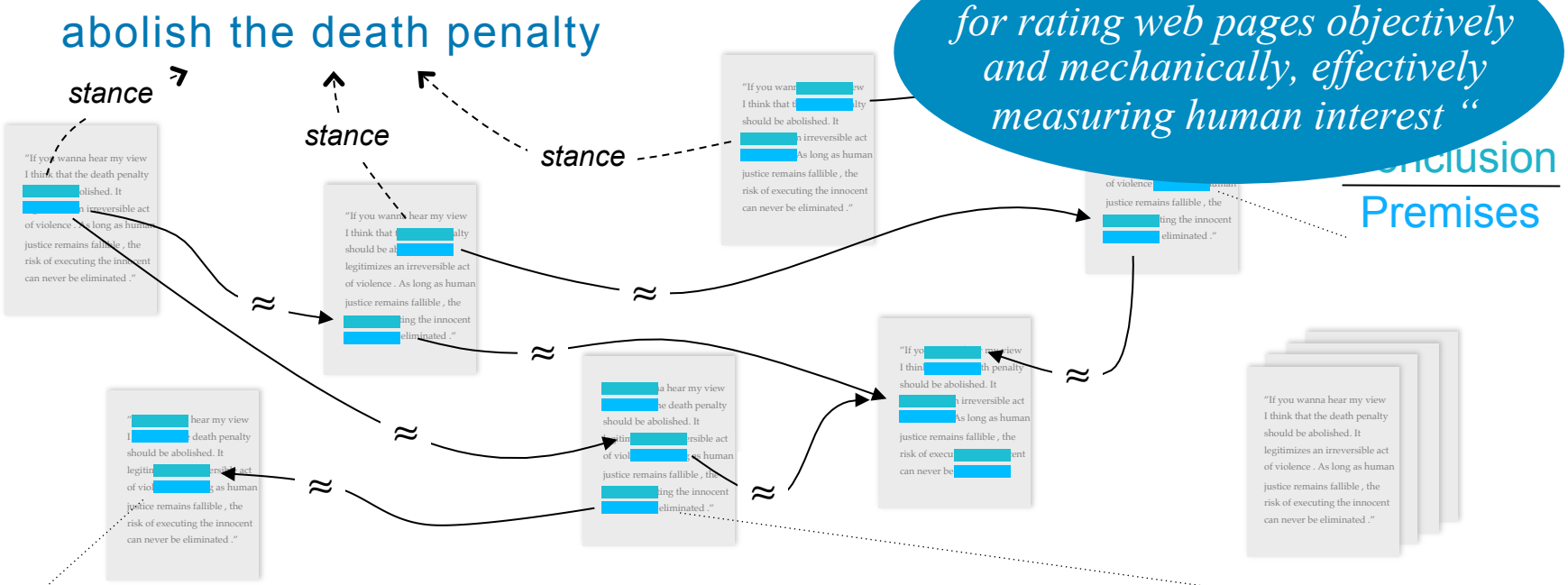
- **Approach**
  - Ignore content and inference of arguments (for now).
  - Derive relevance structurally from the reuse of conclusions at web scale.



Building an argument graph for the web

Page et al. (1999)

” PageRank, a method for rating web pages objectively and mechanically, effectively measuring human interest “



The death penalty doesn‘t deter people from committing serious violent crimes.

A survey of the UN on the relation between the death penalty and homicide rates gave no support to the deterrent hypothesis.

≈

The death penalty should be abolished.

It does not deter people from committing serious violent crimes.

Even if it did, is it acceptable to pay for predicted future crimes of others?

Approach: Adapt PageRank for argument relevance

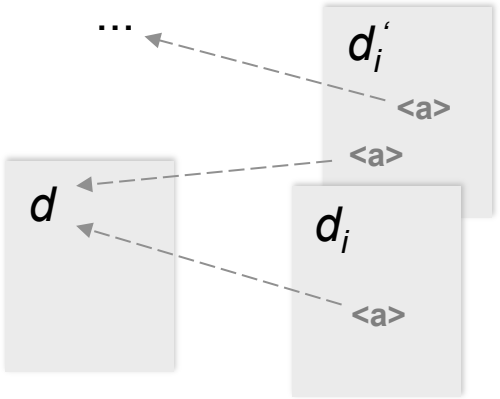
- **Original PageRank score** of a web page *d* (Page et al., 1999)

same score for each page

$$p(d) = (1-\alpha) \cdot \underbrace{\frac{1}{|D|}}_{\text{ground relevance}} + \alpha \cdot \underbrace{\sum_i \frac{p(d_i)}{|D_i|}}_{\text{recursive relevance}}$$

page d_i links to d

pages d_i links to



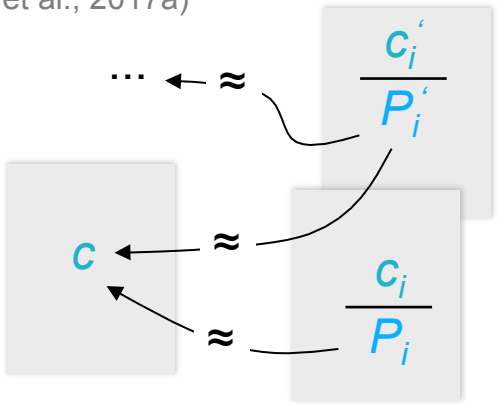
- **Adapted PageRank score** of an argument unit *c* (Wachsmuth et al., 2017a)

PageRank of page *d* containing *c*

$$\hat{p}(c) = (1-\alpha) \cdot \underbrace{\frac{p(d) \cdot |D|}{|A|}}_{\text{ground relevance}} + \alpha \cdot \underbrace{\sum_i \frac{\hat{p}(c_i)}{|P_i|}}_{\text{recursive relevance}}$$

conclusion c_i uses c as premise

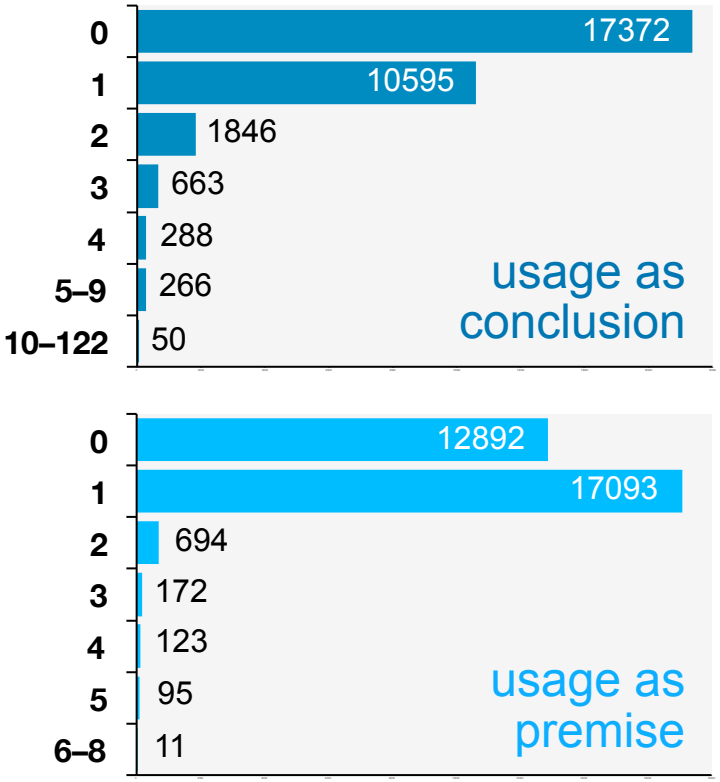
premises of c_i



- **Argument relevance** is aggregation of premise scores
  - Minimum, average, maximum, or sum

Data (Wachsmuth et al., 2017a)

- **No use of argument mining here**
  - Evaluation of PageRank without noise.
- **A ground-truth argument graph**
  - 57 argument corpora from www.aifdb.org.
  - Merged all arguments except for duplicates.
  - 17,877 arguments, 31,080 different units.
  - PageRank computed based on assumption that units match if they span the same text.
- **Benchmark rankings**
  - Since no objective relevance assessments exist, use average assessments a proxy.
  - 110 arguments for 32 general claims.
    2-6 arguments per claim.
  - Ranked by seven annotators (mean Kendall's τ = .36, highest τ = .59).



Evaluation of relevance assessment (Wachsmuth et al., 2017a)

- **Evaluation of unsupervised ranking approaches**

PageRank of premises $\hat{p}$ | Frequency of premises $\sum$ | Similarity of units $c \sim P$ | Sentiment of premises ☺ | Number of premises $|P|$ | Random ranking ⚄

each for minimum, average, maximum, and sum aggregation

- **Experiment on ground-truth graph**
  - Rank arguments with each approach.
  - Correlate with benchmark rankings.
- **Results**
  - PageRank best (with sum aggregation).
  - Notable correlation despite ignorance of content and inference.

best results for each ranking approach

#	Approach	Kendall's τ
1	**PageRank**	**0.28**
2	Number	0.19
3	Sentiment	0.12
4	Frequency	0.10
5	Similarity	0.02
6	Random	0.00

Examples of ”objective“ argument relevance

https://pixabay.com

” Strawberries are the best choice for your breakfast meal.”

#1 *” Berries are superfoods because they’re so high in antioxidants without being high in calories, says Giovinazzo MS, RD, a nutritionist at Clay health club and spa, in New York City.”*

#3 *” Strawberries are good for your ticker.”*

#2 *” One cup of strawberries, for instance, contains your full recommended daily intake of vitamin C, along with high quantities of folic acid and fiber.”*

https://de.wikipedia.org

” Technology has enhanced the daily life of humans.”

#3 *” The use of technology has revolutionized business.”*

#1 *” The internet has enabled us to widen our knowledge.”*

#2 *” Technology has given us a means of social interaction that wasn't possible before.”*

Inclusion of subjectivity

Inclusion of Subjectivity: Overview

- **Problem**
  - Ultimately, effective argumentation requires to consider the target audience.
  - Humans would barely argue without doing so.

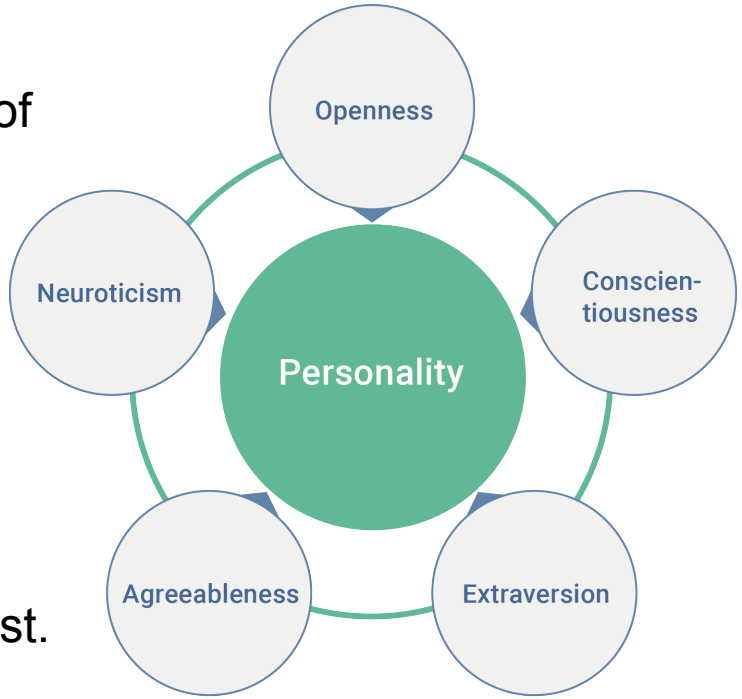
- **Main idea**
  - Model the target audience within quality assessment.
  - This also includes to have audience-specific ground-truth annotations.

- **Missing approaches**
  - Audience model have rarely been included explicitly so far.
  - Implicitly, some annotated corpora may actually represent specific audiences.
  - Recent studies analyze the quality perception of different audiences.

- **Studies**
  - Different personalities. Effectiveness of emotional vs. rational arguments. (Lukin et al., 2017)
  - Different ideologies. Effectiveness of news editorials. (El Baff et al., 2018)

Studying effectiveness based on personality (Lukin et al., 2017)

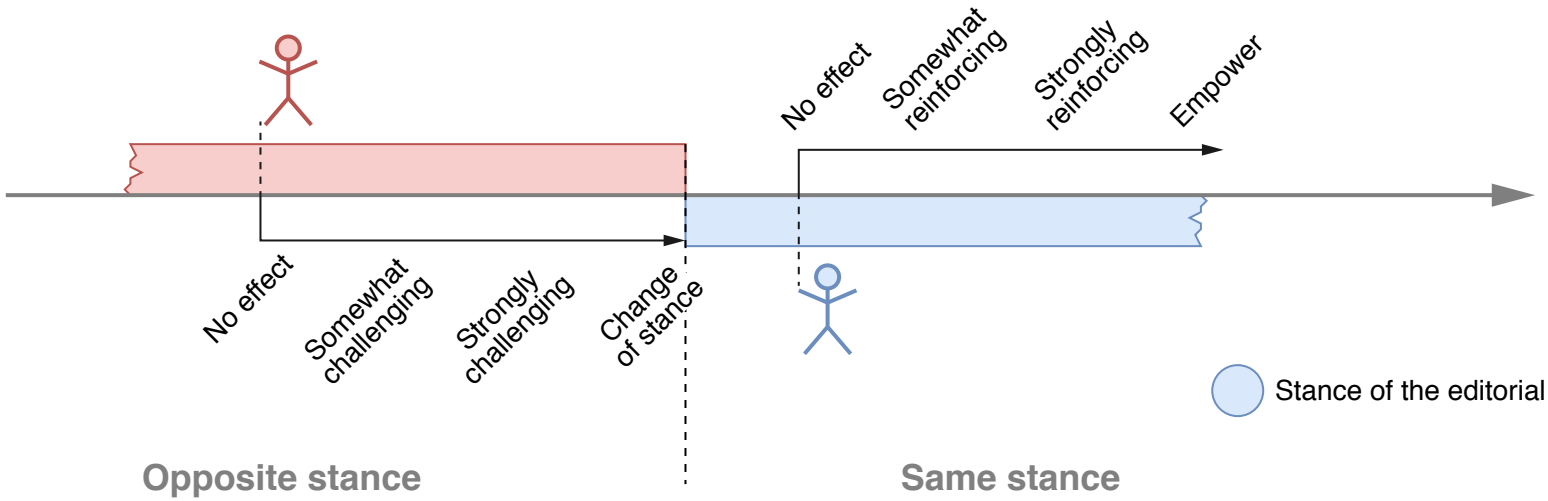
- **Hypothesis**
  - People with different personalities are open to different types of arguments.
- **Study**
  - Impact of personality on the effectiveness of emotional and factual arguments.
  - Personality. Here, the ”Big Five“.
- **Data**
  - 5185 arguments from online dialogs.
  - Quality. Each annotated for whether it changed the belief (to pro, to con, neither).
  - Personality. Each annotator did Big Five test.



- **Selected insights**
  - Agreeable people easiest to predict (F_1 ~.48), extroverted hardest (F_1 ~.44).
  - Factual arguments best for agreeable people, emotional best for open people.

Argumentation quality in news editorials (El Baff et al., 2018)

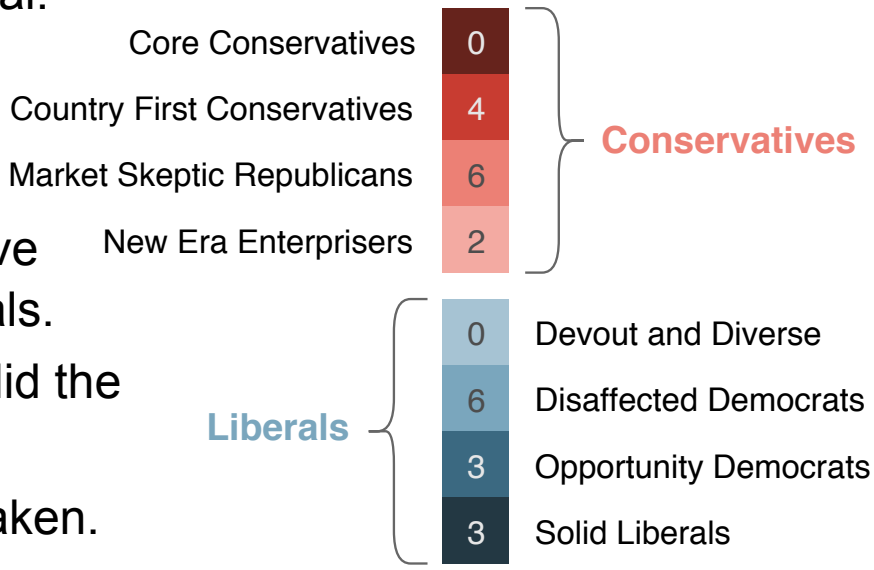
- **Effects of news editorials**
  - News editorials are said to shape public opinion, but they rarely *change* a reader's prior stance.
  - Rather, they challenge or reinforce stance — or neither.



- **Dialectical notion of argumentation quality**
  - A good editorial reinforces one side and challenges the other.
  - Or it challenges both sides.

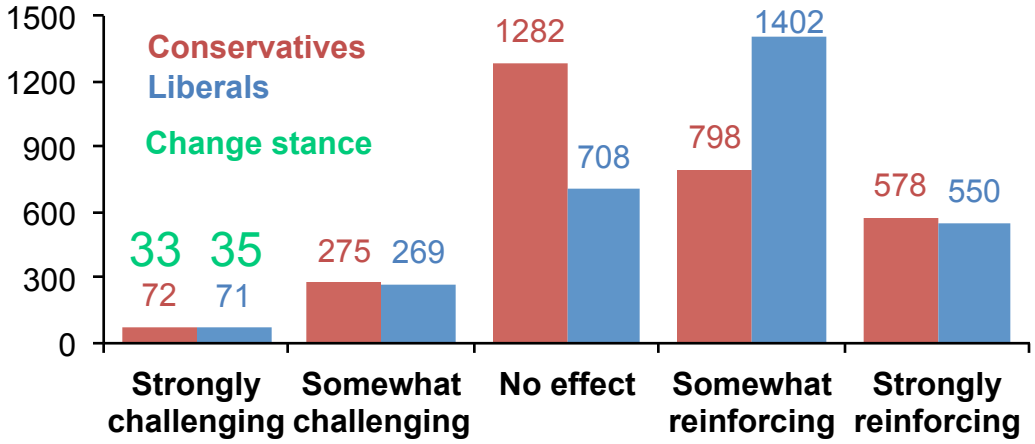
Studying effectiveness based on ideology

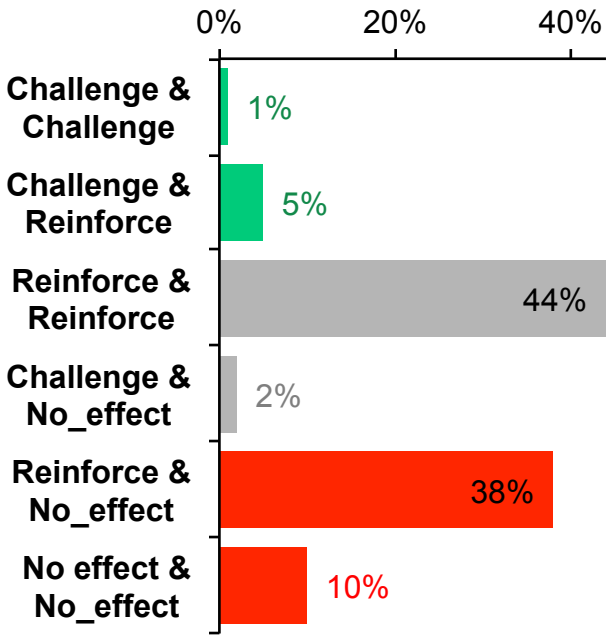
- **Hypothesis**
  - Prior stance depends on political ideology (and personality).
  - Ideology needs to be known to assess the effectiveness of news editorials.
- **Study**
  - Impact of ideology (and personality) on the effectiveness of news editorials.
  - Ideology. Here, conservative vs. liberal.
- **Data**
  - 1000 editorials from NY Times.
  - Quality. Each annotated for persuasive effect by 3 conservatives and 3 liberals.
  - Ideology. All 24 annotators (in total) did the Political Typology Quiz.
  - Personality. Also, Big Five test was taken.



Selected results of the ideology study (El Baff et al., 2018)

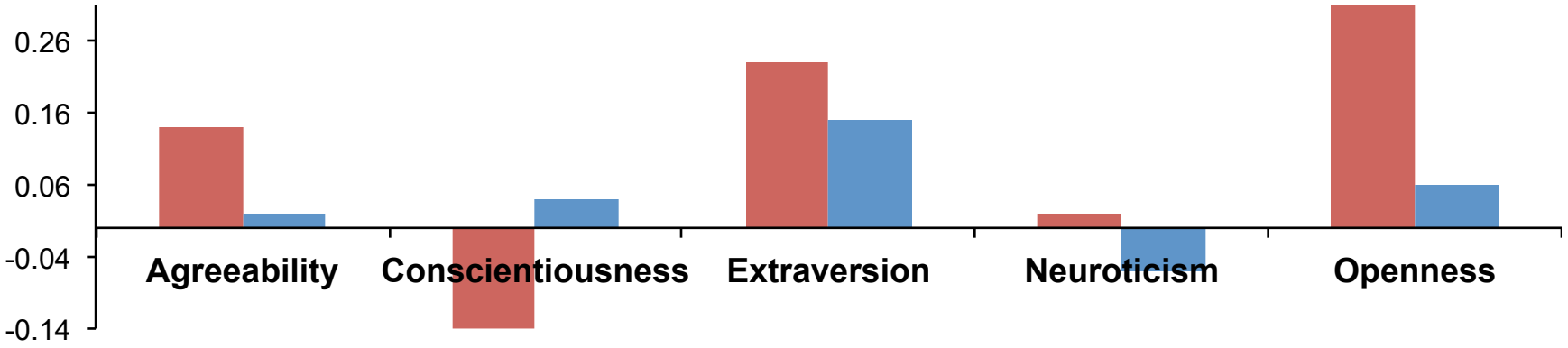
- **Majority effect distribution in the corpus**





- **Effect depending on ideology and personality**

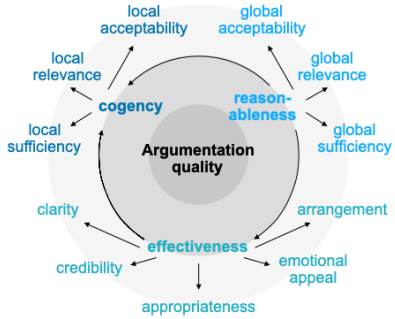
Kendall's τ correlation with challenge/reinforce



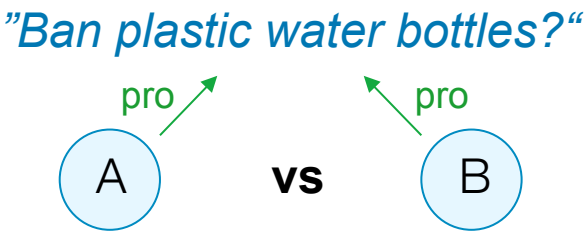
Conclusion

Conclusion

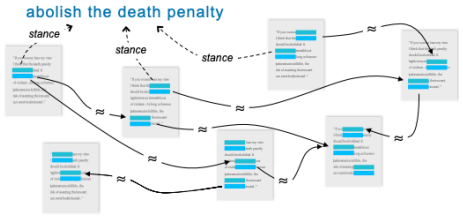
- **Argumentation quality**
  - Several quality dimensions at different granularity levels.
  - What dimension is important, depends on the goal.
  - Many dimensions are highly subjective.



- **Assessment of argumentation quality**
  - Either absolute rating or relative comparison.
  - Structural analyses help to counter subjectiveness.
  - Diverse approaches exist, often learning-based.



- **Selected assessment approaches**
  - Argument-specific features for rhetorical dimensions.
  - Modeling conversational flow to predict debate winners.
  - PageRank for ”objective“ argument relevance.



References

- **Aristotle (2007).** Aristotle (George A. Kennedy, Translator). On Rhetoric: A Theory of Civic Discourse. Clarendon Aristotle series. Oxford University Press, 2007.
- **Blair (2012).** J. Anthony Blair. Groundwork in the Theory of Argumentation. Springer Netherlands, 2012.
- **Boltužic and Šnajder (2015).** Filip Boltužic and Jan Šnajder. Identifying Prominent Arguments in Online Debates using Semantic Textual Similarity. In Proceedings of the 2nd Workshop on Argumentation Mining, pages 110–115, 2015.
- **Braunstain et al. (2016).** Liora Braunstain, Oren Kurland, David Carmel, Idan Szpektor, and Anna Shtok. Supporting Human Answers for Advice-seeking Questions in CQA Sites. In Proceedings of the 38th European Conference on IR Research, pages 129–141, 2016.
- **Cabrio and Villata (2012).** Elena Cabrio and Serena Villata. Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 208–212, 2012.
- **Cohen (2001).** Daniel H. Cohen. Evaluating Arguments and Making Meta-Arguments. Informal Logic, 21(2):73–84, 2001.
- **Damer (2009).** T. Edward Damer. Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments. Wadsworth, Cengage Learning, Belmont, CA, 6th edition, 2009.
- **Dung (1995):** Phan Minh Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. Artificial Intelligence, 77(2):321–357, 1995.
- **El Baff et al. (2018).** Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. Challenge or Empower: Revisiting Argumentation Quality in a News Editorial Corpus. In Proceedings of the 22nd Conference on Computational Natural Language Learning, pages 454–464, 2018.

References

- **Feng et al. (2014).** Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. The Impact of Deep Hierarchical Discourse Structures in the Evaluation of Text Coherence. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 940–949. Dublin City University and Association for Computational Linguistics, 2014.
- **Freeley and Steinberg (2009).** Austin J. Freeley and David L. Steinberg. Argumentation and Debate. Cengage Learning, 12th edition, 2008.
- **Freeman (2011).** Argument Structure: Representation and Theory. Springer, 2011.
- **Govier (2010).** Trudy Govier. A Practical Study of Argument. Wadsworth, Cengage Learning, Belmont, CA, 7th edition, 2010.
- **Granger et al. (2009).** Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. International Corpus of Learner English (version 2), 2009.
- **Habernal and Gurevych (2016a).** Ivan Habernal and Iryna Gurevych. 2016. Which Argument is More Convincing? Analyzing and Predicting Convincingness of Web Arguments using Bidirectional LSTM. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1589–1599.
- **Habernal and Gurevych (2016b).** Ivan Habernal and Iryna Gurevych. What makes a convincing argument? Empirical Analysis and Detecting Attributes of Convincingness in Web Argumentation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 1214–1223, 2016.
- **Hamblin (1970).** Charles L. Hamblin. Fallacies. Methuen, London, UK, 1970.
- **Hoeken (2001).** Hans Hoeken. Anecdotal, Statistical, and Causal evidence: Their Perceived and Actual Persuasiveness. Argumentation, 15(4):425–437, 2001.

References

- **Hovy et al. (2013).** Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning Whom to Trust with MACE. In *Proceedings of NAACL-HLT 2013*, pages 1120–1130.
- **Johnson and Blair (2006).** Ralph H. Johnson and J. Anthony Blair. 2006. Logical Self-defense. International Debate Education Association.
- **O'Keefe and Jackson (1995).** Daniel J. O'Keefe and Sally Jackson. Argument Quality and Persuasive Effects: A Review of Current Approaches. In Argumentation and Values: Proceedings of the Ninth Alta Conference on Argumentation, pages 88–92, 1995.
- **Mercier and Sperber (2011).** Hugo Mercier and Dan Sperber. 2011. Why Do Humans Reason? Arguments for an Argumentative Theory. Behavioral and Brain Sciences, 34:57–111.
- **Page et al. (1999).** Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120, 1999.
- **Park et al. (2015).** Joonsuk Park, Cheryl Blake, and Claire Cardie. Toward Machine-assisted Participation in eRulemaking: An Argumentation Model of Evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 206–210, 2015.
- **Perelman and Olbrecht-Tyteca (1969).** Chaïm Perelman and Lucie Olbrechts-Tyteca. 1969. The New Rhetoric: A Treatise on Argumentation (John Wilkinson and Purcell Weaver, translator). University of Notre Dame Press.
- **Persing and Ng (2013):** Isaac Persing and Vincent Ng. Modeling Thesis Clarity in Student Essays. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 260–269, 2013.
- **Persing and Ng (2014):** Isaac Persing and Vincent Ng. Modeling Prompt Adherence in Student Essays. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1534–1543, 2014.

References

- **Persing and Ng (2015):** Isaac Persing and V. Ng. Modeling Argument Strength in Student Essays. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pages 543–552, 2015.
- **Persing et al. (2010).** Isaac Persing, Alan Davis, and Vincent Ng. Modeling organization in student essays. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 229–239, 2010.
- **Rahimi et al. (2014).** Zahra Rahimi, Diane J. Litman, Richard Correnti, Lindsay Clare Matsumura, Elaine Wang, and Zahid Kisa. Automatic Scoring of an Analytical Response-to-Text Assessment. In Proceedings of the 12th International Conference on Intelligent Tutoring Systems, pages 601–610, 2014.
- **Rahimi et al. (2015).** Zahra Rahimi, Diane Litman, Elaine Wang, and Richard Correnti. Incorporating Coherence of Topics as a Criterion in Automatic Response-to-Text Assessment of the Organization of Writing. In Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 20–30, 2015.
- **Stab and Gurevych (2014).** Christian Stab and Iryna Gurevych. Annotating Argument Components and Relations in Persuasive Essays. In Proceedings of the 25th Conference on Computational Linguistics, pages 1501–1510, 2014.
- **Stab and Gurevych (2017).** Christian Stab and Iryna Gurevych. Recognizing Insufficiently Supported Arguments in Argumentative Essays. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 980–990, 2017.
- **Stede and Schneider (2018).** Manfred Stede and Jodi Schneider. Argumentation Mining. Synthesis Lectures on Human Language Technologies 40, Morgan & Claypool, 2018.
- **Tindale (2007).** Christopher W. Tindale. 2007. Fallacies and Argument Appraisal. Critical Reasoning and Argumentation. Cambridge University Press.
- **Toulmin (1958).** Stephen E. Toulmin. The Uses of Argument. Cambridge University Press, 1958.

References

- **van Eemeren (2015).** Frans H. van Eemeren. Reasonableness and Effectiveness in Argumentative Discourse: Fifty Contributions to the Development of Pragma-Dialectics. Argumentation Library. Springer International Publishing, 2015.
- **van Eemeren and Grootendoorst (2004).** Frans H. van Eemeren and Rob Grootendorst. 2004. A Systematic Theory of Argumentation: The Pragma-Dialectical Approach. Cambridge University Press, Cambridge, UK.
- **Wachsmuth et al. (2016).** Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. Using Argument Mining to Assess the Argumentation Quality of Essays. In: Proceedings of the 26th International Conference on Computational Linguistics, pages 1680–1692, 2016.
- **Wachsmuth et al. (2017a).** Henning Wachsmuth, Benno Stein, and Yamen Ajjour. ”PageRank“ for Argument Relevance. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 1116–1126, 2017.
- **Wachsmuth et al. (2017b).** Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational Argumentation Quality Assessment in Natural Language. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 176–187, 2017.
- **Wachsmuth et al. (2017d).** Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. Argumentation Quality Assessment: Theory vs. Practice. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Vancouver, Canada, pages 250–255, 2017.
- **Walton (2006).** Douglas Walton. Fundamentals of Critical Argumentation. Cambridge University Press, 2006.
- **Walton et al. (2008).** Douglas Walton, Christopher Reed, and Fabrizio Macagno. Argumentation Schemes. Cambridge University Press, 2008.

References

- **Wang et al. (2017).** Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. Winning on the Merits: The Joint Effects of Content and Style on Debate Outcomes. In: Transactions of the Association for Computational Linguistics 5, pages 219--232, 2017.
- **Wei et al. (2016).** Zhongyu Wei, Yang Liu, and Yi Li. Is this Post Persuasive? Ranking Argumentative Comments in Online Forum. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 195–200, 2016.
- **Zhang et al. (2016).** Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. Conversational Flow in Oxford-style Debates. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 136–141, 2016.