Reflected Text Analysis beyond Linguistics

DGfS-CL fall school

Nils Reiter,
nils.reiter@ims.uni-stuttgart.de

Sept. 9-13, 2019

Part I

Introduction and Preliminaries

Experiments in Text Analysis

Agenda

…beyond Linguistics

About this class

- Overview class
  - Putting stuff together so that it makes sense
  - Not exhaustive: Many aspects not in depth
    - Pointers to continue reading
- Heterogeneous audience
  - Hopefully, everyone recognises things they know already, and learns something new
- Practical exercises
- Course page: `https://nilsreiter.de/refl2019`

About Me

Figure: Nils (right)

Nils Reiter

- Master("Diplom") in Computational Linguistics (Saarland University)
- PhD in Computational Linguistics (Heidelberg University, 2007-2013)
- Postdoc at the IMS (Stuttgart University, 2014-2019)
- Now: Interim professor for Linguistic Information Processing / Digital Humanities (Cologne University)

- `https://nilsreiter.de`

About Me

Projects

- PhD: Extracting narrative structures from ritual descriptions (w/ classical indology)
- CRETA: Center for Reflected Text Analytics (w/ literary studies, linguistics, philosophy, social sciences, visualisation)
- QuaDramA: Quantitative Drama Analytics (w/ literary studies, M. Willand)
- SANTA: Shared task for developing annotation guidelines for narrative phenomena (w/ lit. studies, E. Gius & M. Willand)

About Me

Research Interests

- Artistic/non-standard use of language (e.g., humor, art, metaphors, literature), why do we express things in a certain (individual!) way?
- Operationalisation of complex research questions and tasks
- Integration of quantitative/statistical research methods/results into hermeneutic research (e.g., interpretable machine learning)

$\rightarrow$ 'Digital Humanities'

About Me

Research Interests

- Artistic/non-standard use of language (e.g., humor, art, metaphors, literature), why do we express things in a certain (individual!) way?
- Operationalisation of complex research questions and tasks
- Integration of quantitative/statistical research methods/results into hermeneutic research (e.g., interpretable machine learning)

$\rightarrow$ 'Digital Humanities'

- …also, I just like coding and team work

Section 1

Experiments in Text Analysis

Experiments

- Reproducibility
- Hypotheses about the operationalisation of language/text phenomena

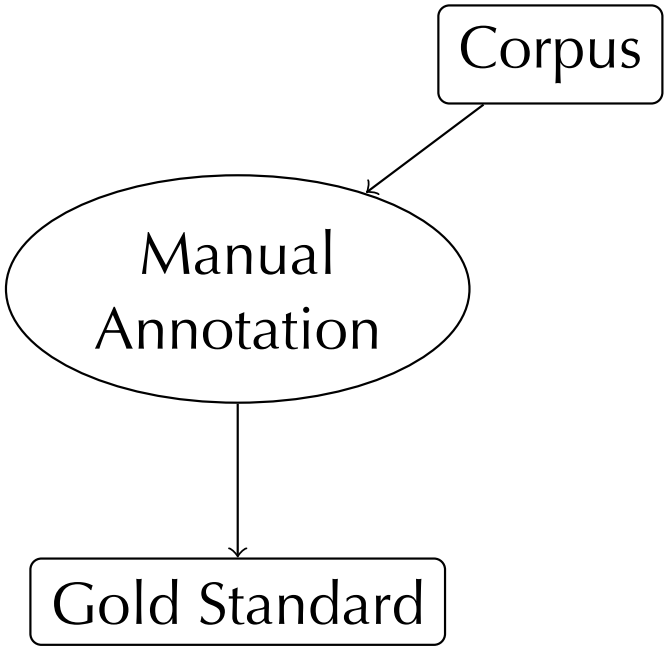
Example

- Position within a sentence is indicative for the part of speech
- Meaning of a word depends on its context
- The protagonist of a play is the character who talks the most

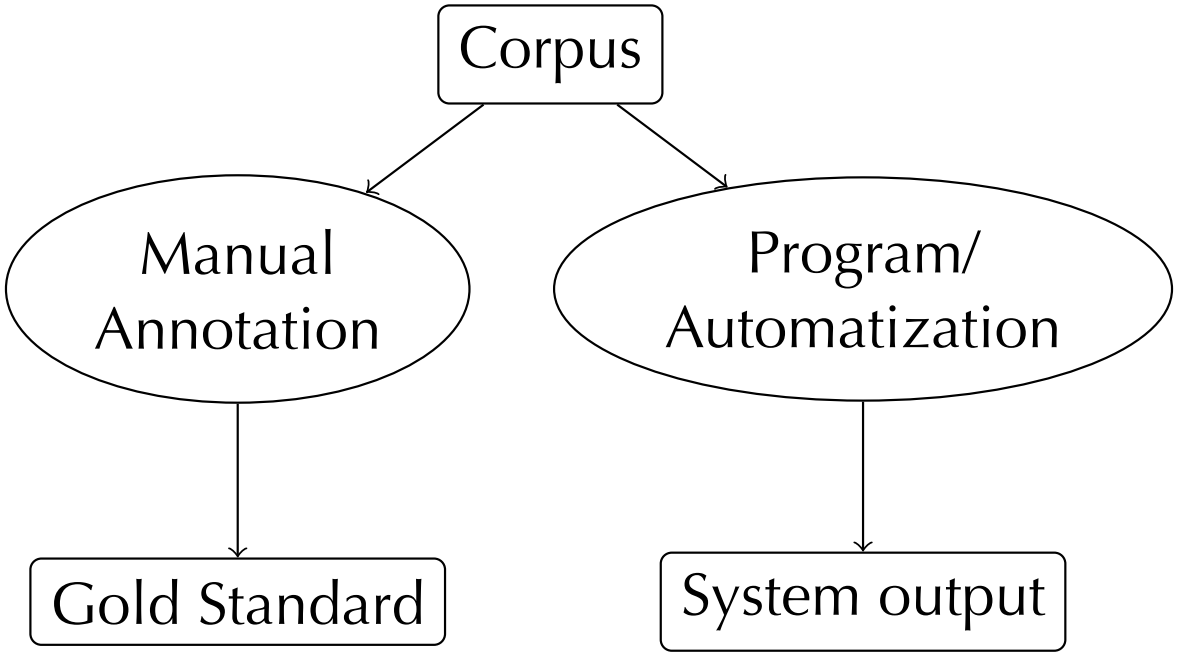
Experiments

Corpus

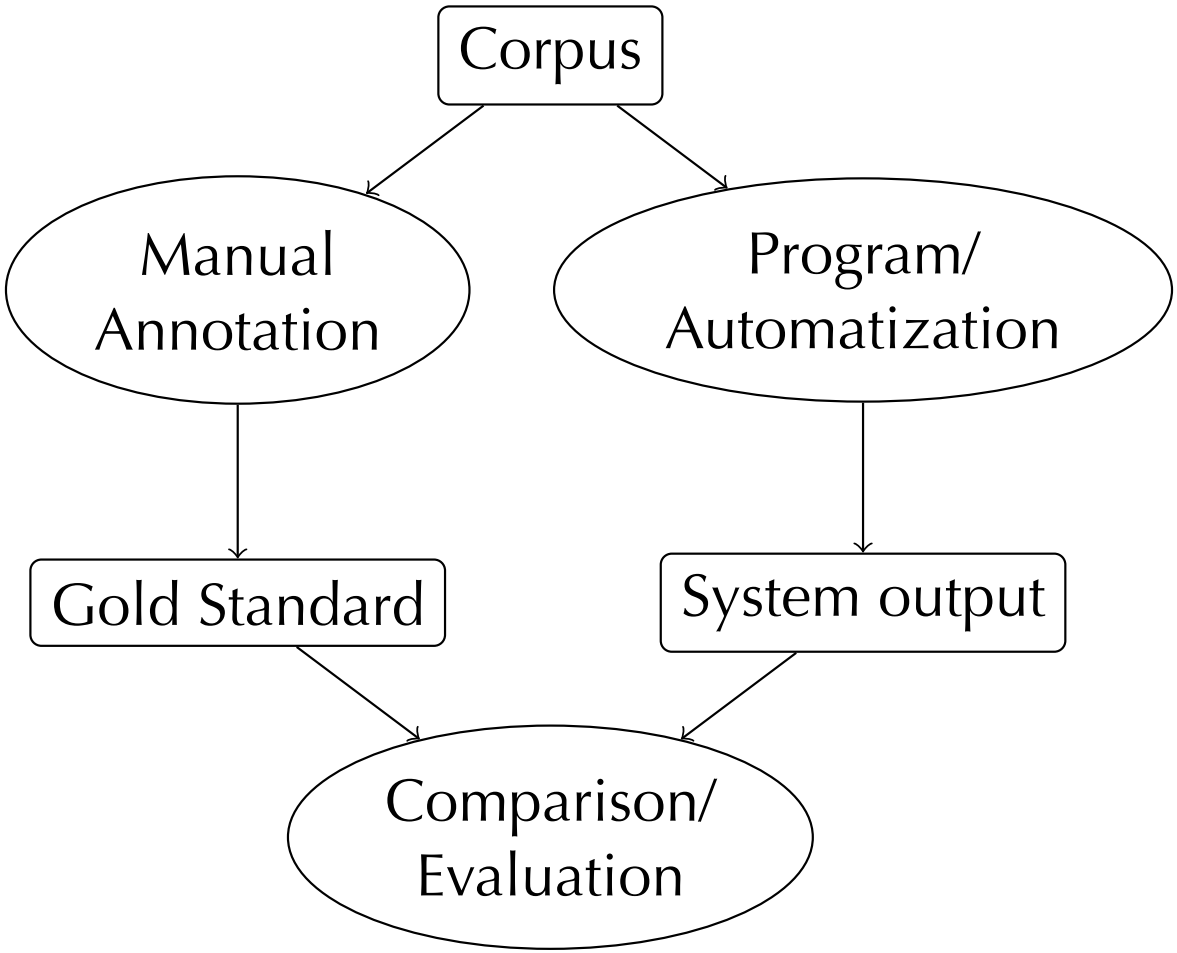
Experiments



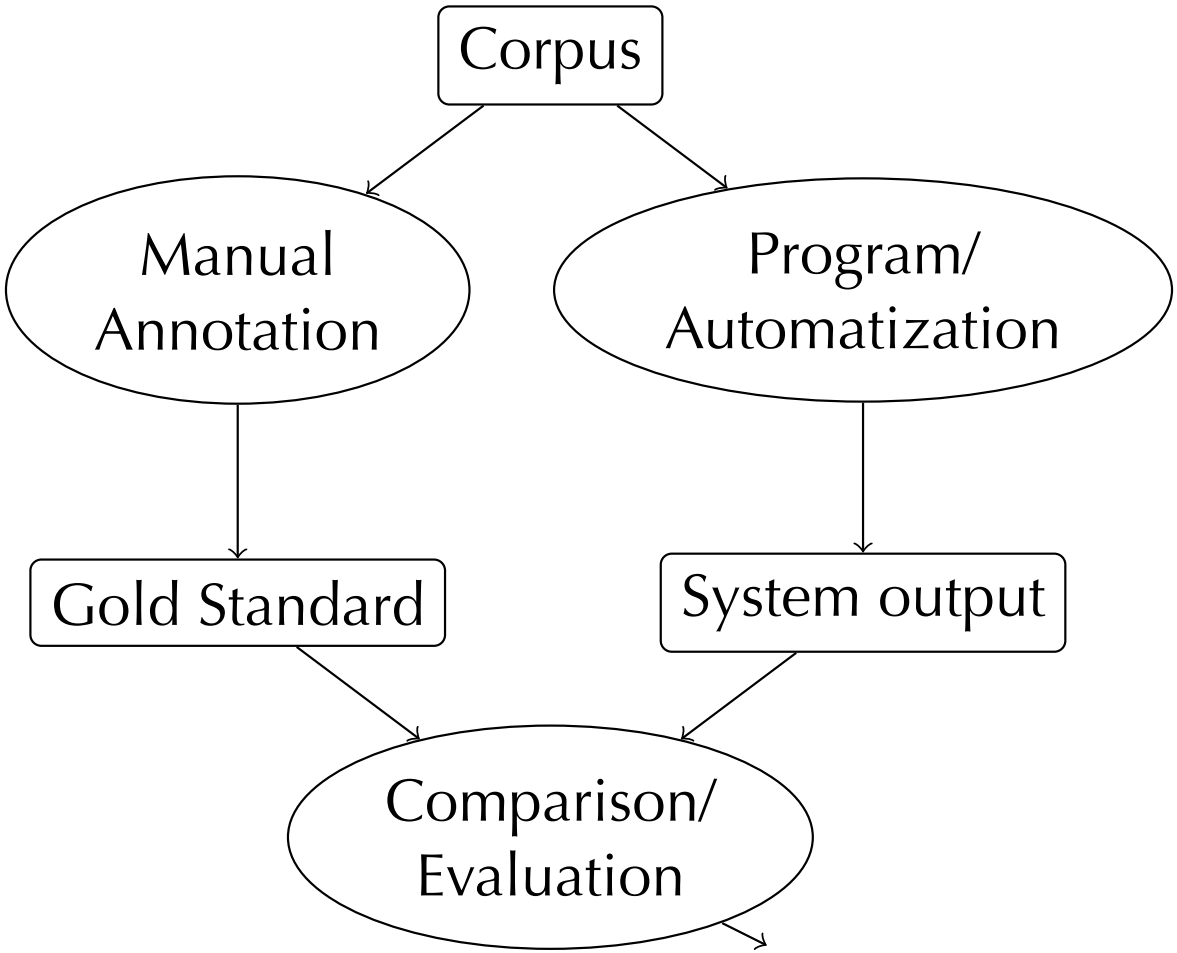
Experiments



Experiments

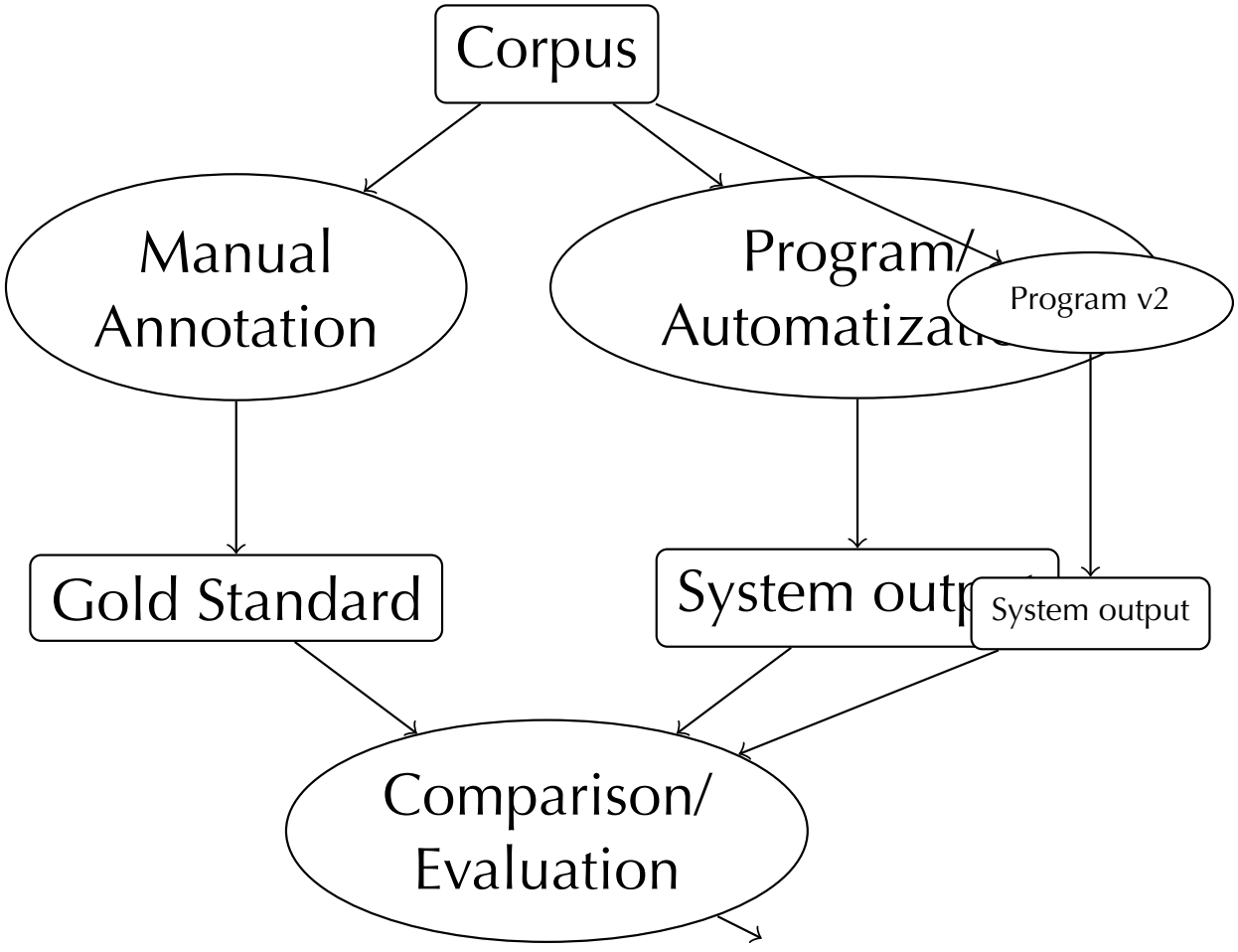


Experiments



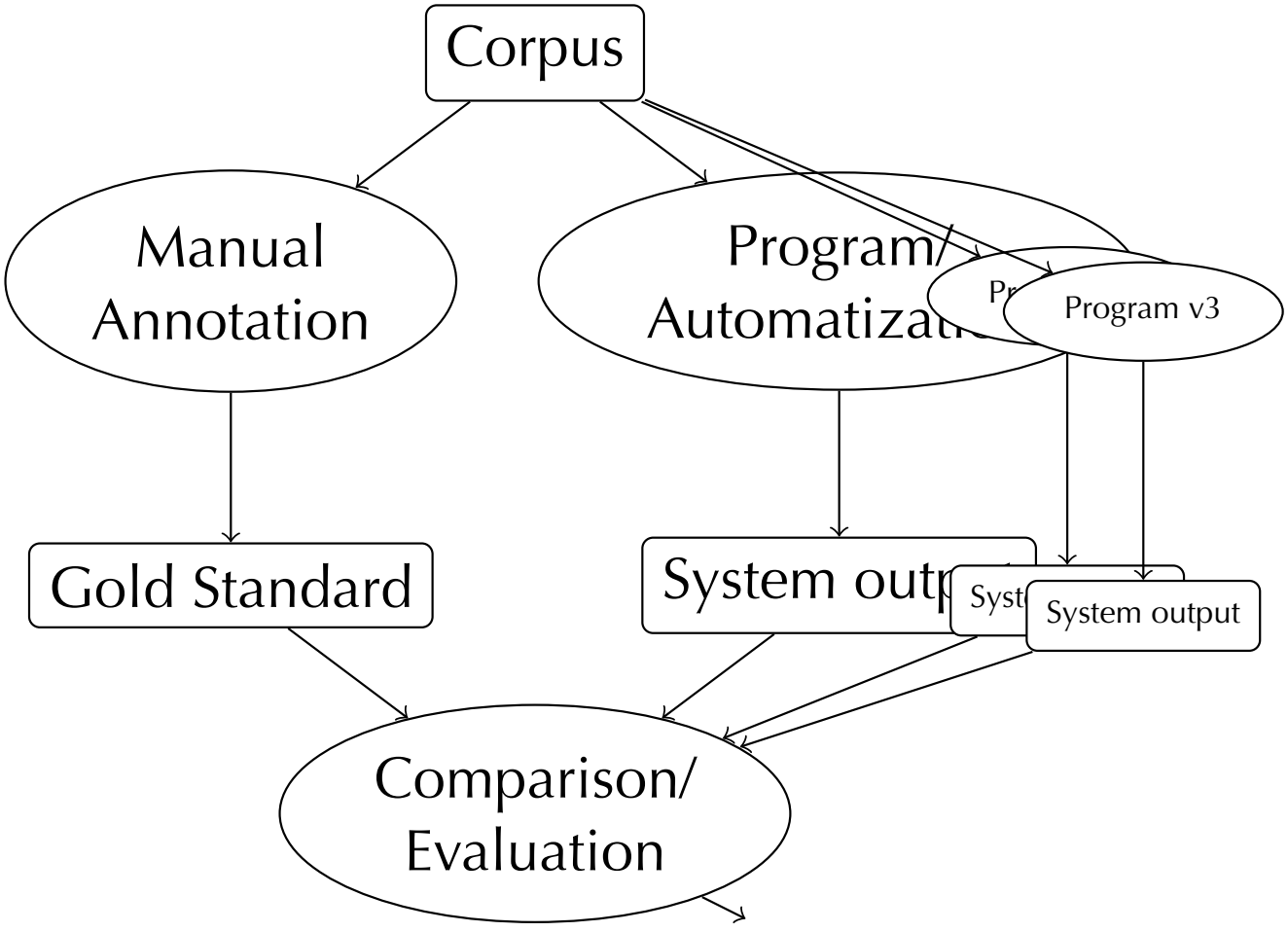
	P	R	F
Programm	5.7	9.2	…

Experiments



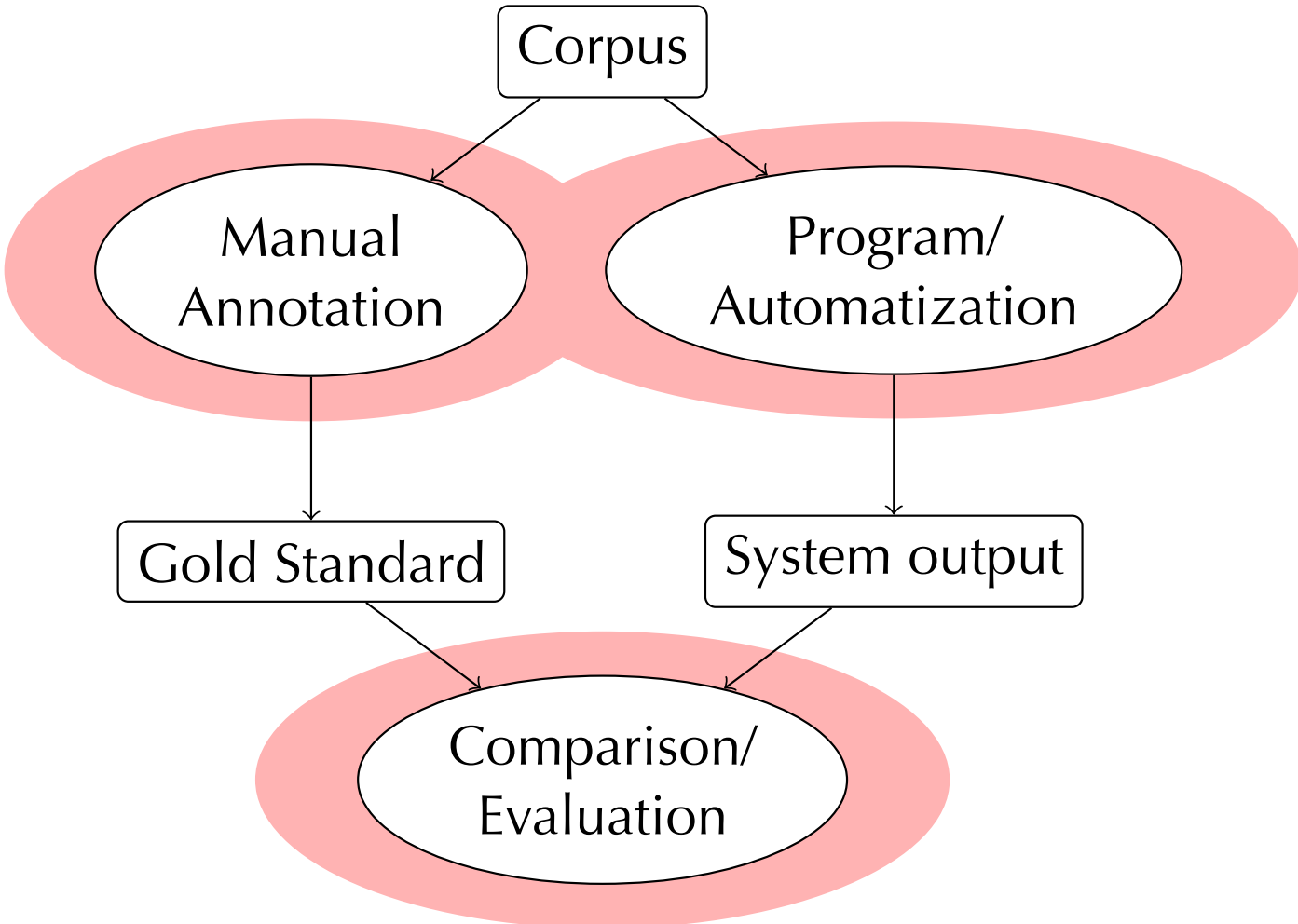
	P	R	F
Programm	5.7	9.2	…
v2	9.9	16.7	…

Experiments



	P	R	F
Programm	5.7	9.2	…
v2	9.9	16.7	…
v3	15.3	21.8	…

Experiments



What do we need?

- Gold standard
  - Formal, machine-readable truth
- Program (rule-based, machine learning)
  - Encodes our hypotheses
- Evaluation metric
  - Formalised comparison of annotations

What do we learn?

- Directly
  - Prediction quality of the program on this corpus
- Indirectly
  - Insights, why the program works well (or not)
  - Estimation of the quality on other corpora
- Long term
  - Iterative improvement of the programs (e.g., in shared tasks)

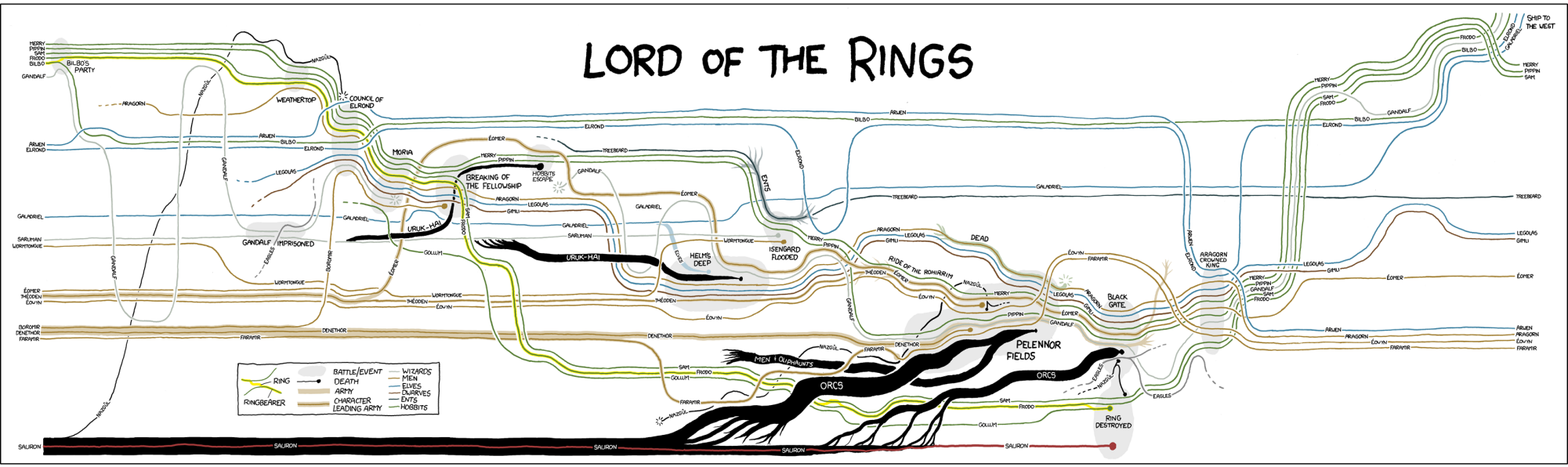
Agenda for this week

Day	14:00-15:30	16:00-17:30
Monday	Introduction, overview, annotation	Annotation exercise, inter-annotator agreement
Tuesday	Machine learning overview and evaluation, algorithms	Algorithms
Wednesday	Introduction into shared task, hands on session	Hands on session
Thursday	Excursion to the German Literature Archive, Marbach	
Friday	Hands on session, shared task evaluation	What to do next, closing discussion

(It's a plan. It will change.)

Section 3

...beyond Linguistics


LORD OF THE RINGS
BILBO'S PARTY
WEATHERTOP
COUNCIL OF ELROND
MORIA
BREAKING OF THE FELLOWSHIP
HOBBITS ESCAPE
URUK-HAI
GANDALF IMPRISONED
HELM'S DEEP
ISENGARD FLOODED
RIDE OF THE ROHIRRIM
DEAD
BLACK GATE
PELENNOR FIELDS
MEN + OLIPHAUNTS
ORCS
ORCS
EAGLES
RING DESTROYED
ARAGORN CROWNED KING
SHIP TO THE WEST
BATTLE/EVENT
DEATH
ARMY
CHARACTER LEADING ARMY
RING
RINGBEARER
WIZARDS
MEN
ELVES
DWARVES
ENTS
HOBBITS
SAURON

Reconstructing Plot

- Linguistic analyses
  - Parts of speech, syntax, semantic roles, coreference
    - Resolving “shadow in the east” requires geographical knowledge in fictional world

Reconstructing Plot

- Linguistic analyses
  - Parts of speech, syntax, semantic roles, coreference
    - Resolving "shadow in the east" requires geographical knowledge in fictional world
- Narrative analysis
  - Time: Simultaneous events
  - Sub narratives: Characters describing events
    - E.g., Bilbo describes events taking place in *The Hobbit*
  - Misleading cues
    - E.g., Pippin and Merry presumed dead

Reconstructing Plot

- Linguistic analyses
  - Parts of speech, syntax, semantic roles, coreference
    - Resolving “shadow in the east” requires geographical knowledge in fictional world
- Narrative analysis
  - Time: Simultaneous events
  - Sub narratives: Characters describing events
    - E.g., Bilbo describes events taking place in *The Hobbit*
  - Misleading cues
    - E.g., Pippin and Merry presumed dead
- “Computational Literary Studies” $\subset$ Digital Humanities
  - Interesting: Methodological impact of digital stuff

Digital Humanities for Computational Linguists

Digital Humanities: Umbrella term

- Public history, citizen science, science communication in social media, interim field ('Brückentechnologie'), digital editions, new publication forms and venues, open science/source, (3D) visualisation, virtual reality, ...
- `http://whatisdigitalhumanities.com`

Digital Humanities for Computational Linguists

Digital Humanities: Umbrella term

- Public history, citizen science, science communication in social media, interim field ('Brückentechnologie'), digital editions, new publication forms and venues, open science/source, (3D) visualisation, virtual reality, ...
- `http://whatisdigitalhumanities.com`
  - Two areas for NLPers to contribute
    - Text processing tools: from pos-tagging to semantic role labeling
    - Methodology: How to analyse texts properly

Digital Humanities for Computational Linguists

Digital Humanities: Umbrella term

- Public history, citizen science, science communication in social media, interim field ('Brückentechnologie'), digital editions, new publication forms and venues, open science/source, (3D) visualisation, virtual reality, ...
- `http://whatisdigitalhumanities.com`
  - Two areas for NLPers to contribute
    - Text processing tools: from pos-tagging to semantic role labeling
    - Methodology: How to analyse texts properly

→ Wednesday: Evening lecture by Fotis Jannidis

Digital Humanities for Computational Linguists

Challenges

- Limited amount of data
  - Shakespeare just didn't write anything new in the past years
  - Only a limited amount of text from the middle ages is still available
    - …and new texts will not be written
- Definitions and concepts
  - Humanities don't have formal definitions
  - Often: Highly context-dependent, selecting the 'right' context is part of the research
    - Context usually text-external
  - Tasks need to be defined (so that we may solve them)
  - Concepts need to be operationalised, so that literary scholars trust you
- Non-technical users
  - Users may not be able to interpret table-like results properly
  - Depend on visualisations
  - Responsibility

Digital Humanities for Computational Linguists

Text processing

- Many Humanities disciplines are text-oriented
- Automatically analysing specific texts (or corpora) with standard NLP tools can lead to interesting findings

Digital Humanities for Computational Linguists

Text processing

- Many Humanities disciplines are text-oriented
- Automatically analysing specific texts (or corpora) with standard NLP tools can lead to interesting findings
  - American short stories Siewert and Reiter (2018)
    - 1820-1915: Industrialisation and increase of social mobility; representation of minorities in literature
    - Increase of proper names (via NER)
    - Increase of vernacular direct speech (via spell checker)

Digital Humanities for Computational Linguists

Text processing

- Many Humanities disciplines are text-oriented
- Automatically analysing specific texts (or corpora) with standard NLP tools can lead to interesting findings
  - American short stories Siewert and Reiter (2018)
    - 1820-1915: Industrialisation and increase of social mobility; representation of minorities in literature
    - Increase of proper names (via NER)
    - Increase of vernacular direct speech (via spell checker)
  - User-generated book reviews/fan fiction Willand et al. (2018)
    - Perspective on perception of literature
    - People associate based on character properties: Sherlock Holmes → Mr. Spock; The Three Investigators secretly gay

Digital Humanities for Computational Linguists

Text processing – issues

- Domain issues
  - NLP tools are typically trained on news data
  - Humanities texts deviate from news in various ways
    - News rarely contains first person perspective or imperatives
    - Literary texts may refer to fictional places – or no named entities at all
- Text structure
  - NLP tools often ignore text structure: “Text” is a sequence of tokens
    - Prose texts have chapters, sometimes character lists or chapter summaries
    - Dramatic texts contain acts, scenes, character lists and character speech
    - Lyrics texts contain stanzas which form a metrical structure (rhyming)
  - Sometimes NLP development task, sometimes engineering task

Digital Humanities for Computational Linguists

Methodology

- NLP/CL has 50+ years experience in working with the challenges posed by natural language
- Know-how can be applied to new problems
- Concept development through annotation
  - Today's linguistic concepts have been tested, discussed and strengthened by annotation projects
- "Annotatability" is a core requirement of NLP concepts
  - Annotated corpora come with some measures of inter-annotator agreement
  - How to measure inter-annotator agreement is an active research area
    Fournier and Inkpen (2012) and Mathet et al. (2015)

Digital Humanities for Computational Linguists

Methodology

- Annotation of literary studies concepts using NLP annotation workflow just beginning
  - Proppian folktale event types — Finlayson (2015)
  - Time-related narrative structure — Bögel et al. (2015)

Digital Humanities for Computational Linguists

Methodology

- Annotation of literary studies concepts using NLP annotation workflow just beginning
  - Proppian folktale event types — Finlayson (2015)
  - Time-related narrative structure — Bögel et al. (2015)
- Experimental setup for method development
  - Shared tasks to foster tool development
    - STs for annotation schema development — Reiter, Gius, et al. (2017)
- "Virtual tasks" like RTE as a unifying abstraction layer to solve typical problems in a stylised way

Part II

Annotation

Background Reading

Eduard Hovy and Julia Lavid. "Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics". In: *International Journal of Translation Studies* 22.1 (Jan. 2010). URL: https://www.cs.cmu.edu/~hovy/papers/10KNS-annotation-Hovy-Lavid.pdf

Nancy Ide and James Pustejovsky, eds. *Handbook of Linguistic Annotation*. Springer, 2017. URL: https://www.springer.com/de/book/9789402408799

Janis Pagel et al. "A Unified Text Annotation Workflow for Diverse Goals". In: *Proceedings of the Workshop on Annotation in Digital Humanities, co-located with ESSLLI 2018*. Ed. by Sandra Kübler and Heike Zinsmeister. Sofia, Bulgaria, Aug. 2018. URL: http://ceur-ws.org/Vol-2155/pagel.pdf

Nils Reiter, Marcus Willand, et al. "A Shared Task for the Digital Humanities: Introduction to Annotation, Narrative Levels and Shared Tasks". In: *Cultural Analytics* (to appear)

What are Annotations?

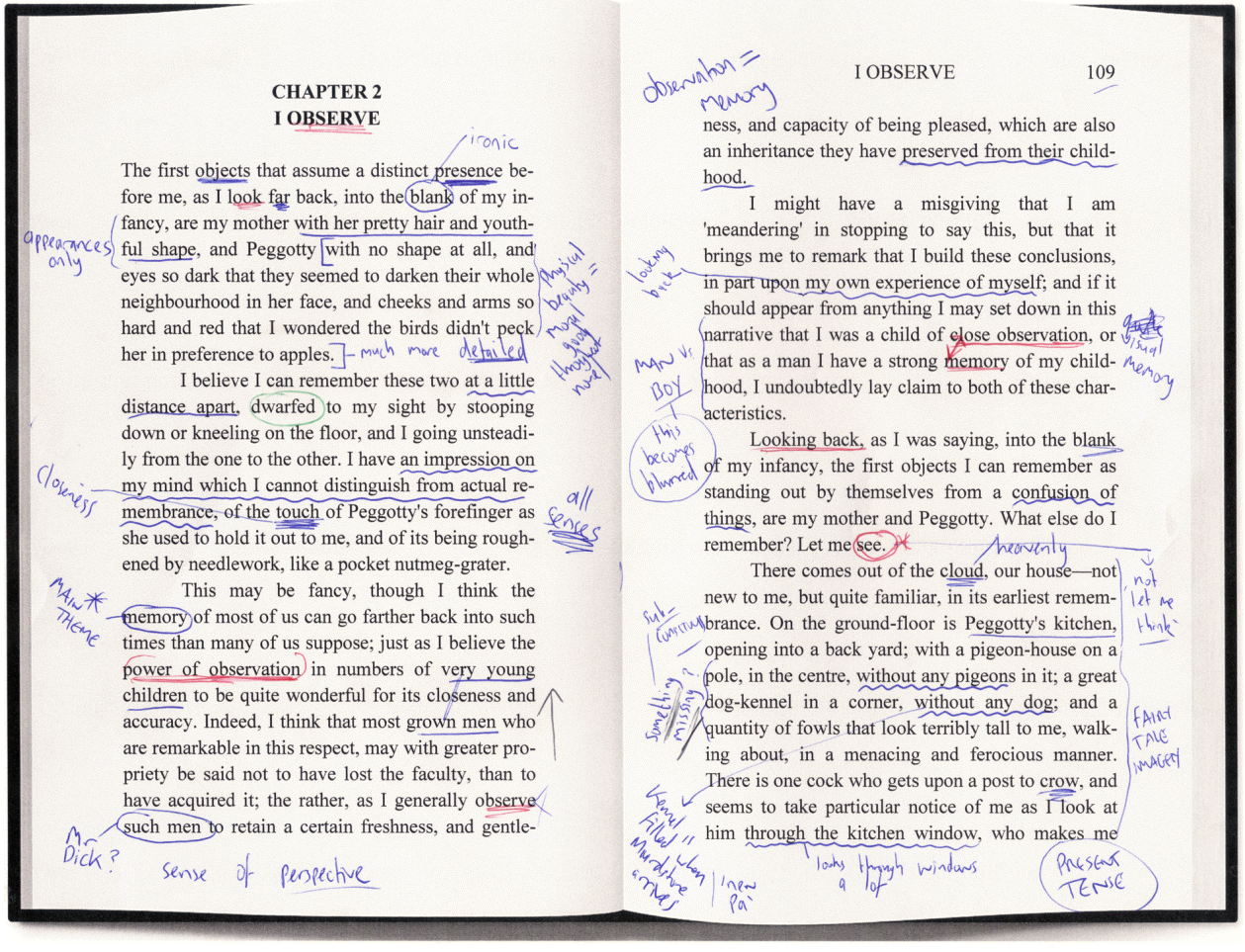
An annotation is a metadatum (e.g. a post, explanation, markup) attached to location or other data.

WP: Annotation, Version 880441583

What are Annotations?

An annotation is a metadatum (e.g. a post, explanation, markup) attached to location or other data.

WP: Annotation, Version 880441583



CHAPTER 2
I OBSERVE

The first objects that assume a distinct presence before me, as I look far back, into the blank of my infancy, are my mother with her pretty hair and youthful shape, and Peggotty with no shape at all, and eyes so dark that they seemed to darken their whole neighbourhood in her face, and cheeks and arms so hard and red that I wondered the birds didn't peck her in preference to apples.

I believe I can remember these two at a little distance apart, dwarfed to my sight by stooping down or kneeling on the floor, and I going unsteadily from the one to the other. I have an impression on my mind which I cannot distinguish from actual remembrance, of the touch of Peggotty's forefinger as she used to hold it out to me, and of its being roughened by needlework, like a pocket nutmeg-grater.

This may be fancy, though I think the memory of most of us can go farther back into such times than many of us suppose; just as I believe the power of observation in numbers of very young children to be quite wonderful for its closeness and accuracy. Indeed, I think that most grown men who are remarkable in this respect, may with greater propriety be said not to have lost the faculty, than to have acquired it; the rather, as I generally observe such men to retain a certain freshness, and gentle-

ness, and capacity of being pleased, which are also an inheritance they have preserved from their childhood.

I might have a misgiving that I am 'meandering' in stopping to say this, but that it brings me to remark that I build these conclusions, in part upon my own experience of myself; and if it should appear from anything I may set down in this narrative that I was a child of close observation, or that as a man I have a strong memory of my childhood, I undoubtedly lay claim to both of these characteristics.

Looking back, as I was saying, into the blank of my infancy, the first objects I can remember as standing out by themselves from a confusion of things, are my mother and Peggotty. What else do I remember? Let me see.

There comes out of the cloud, our house—not new to me, but quite familiar, in its earliest remembrance. On the ground-floor is Peggotty's kitchen, opening into a back yard; with a pigeon-house on a pole, in the centre, without any pigeons in it; a great dog-kennel in a corner, without any dog; and a quantity of fowls that look terribly tall to me, walking about, in a menacing and ferocious manner. There is one cock who gets upon a post to crow, and seems to take particular notice of me as I look at him through the kitchen window, who makes me

Figure: Manual annotation on paper

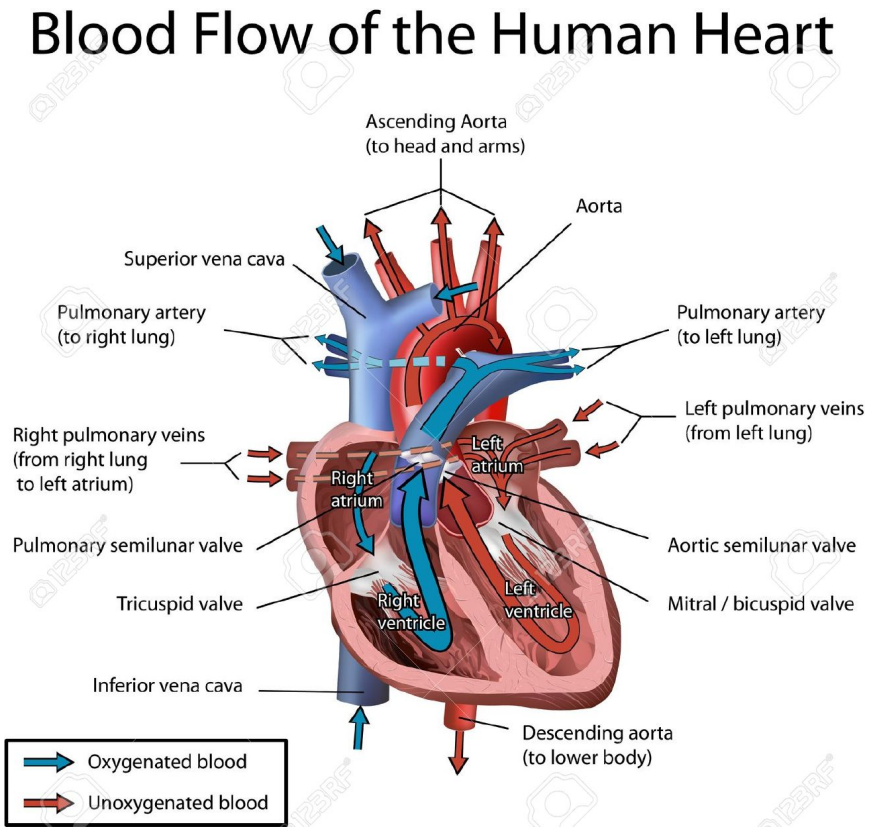


Figure: Image annotation

Figure: Computer-based annotation

Figure: Digital annotations of parts of speech

Annotation

Process and result

- Specific positions in an artefact
- Different types
  - Text vs. image artefact
  - on paper vs. on screen (digital)
  - Automatic vs. manual
  - free vs. fixed categories
  - subjective vs. objective

Annotation

Process and result

- Specific positions in an artefact
- Different types
  - Text vs. image artefact
  - on paper vs. on screen (digital)
  - Automatic vs. manual
  - free vs. fixed categories
  - subjective vs. objective

CL and DH

- Digital (sometimes manual)
- Annotation of text
- Manual and automatic
- Fixed categories and free

Section 5

Why Annotation?

Why Annotation?

- Concept Sharpening
- Data creation data for developing automatic tools
  - Training
  - Test

Concept Sharpening

- Theories make statements about categories
  - “Narration speed varies in narrative texts”
  - “Determiner and noun form a nominal phrase”
- Annotation: Application of the theory on text(s)

Concept Sharpening

- Theories make statements about categories
  - “Narration speed varies in narrative texts”
  - “Determiner and noun form a nominal phrase”
- Annotation: Application of the theory on text(s)
- Sharpening
  - Annotators notice item classes that are not covered by the theory
    - Annotator: “Some words cannot be annotated”
  - Systematically confused categories are likely ill defined, or their differences need be clearer

Concept Sharpening

- Theories make statements about categories
  - “Narration speed varies in narrative texts”
  - “Determiner and noun form a nominal phrase”
- Annotation: Application of the theory on text(s)
- Sharpening
  - Annotators notice item classes that are not covered by the theory
    - Annotator: “Some words cannot be annotated”
  - Systematically confused categories are likely ill defined, or their differences need be clearer
    - The Duke was pretty last night.
    - The Duchess was entertaining last night.
    - Adjektive (JJ) oder verb (gerund, VBG)?

Concept Sharpening

- Theories make statements about categories
  - “Narration speed varies in narrative texts”
  - “Determiner and noun form a nominal phrase”
- Annotation: Application of the theory on text(s)
- Sharpening
  - Annotators notice item classes that are not covered by the theory
    - Annotator: “Some words cannot be annotated”
  - Systematically confused categories are likely ill defined, or their differences need be clearer
    - The Duke was pretty last night.
    - The Duchess was entertaining last night.
    - Adjektive (JJ) oder verb (gerund, VBG)?
- Annotations allow quantitative statements about categories
  - “x% of the words are verbs”
  - “Narration speed is slowest in the middle of a narration”

Concept Sharpening

- Annotation as a means towards an end
  - Theories need to be adapted in order to be used quantitatively
  - Adaptation: Formalisation, restriction, generalisation
  - Annotation can be a tool to ensure that
    - Forces exactness
    - Allows comparing different interpretations of a theory
- Parts of speech: More or less solved (STTS, Penn Treebank)
- Narrative structures: Still in early stages Bögel et al. (2015)

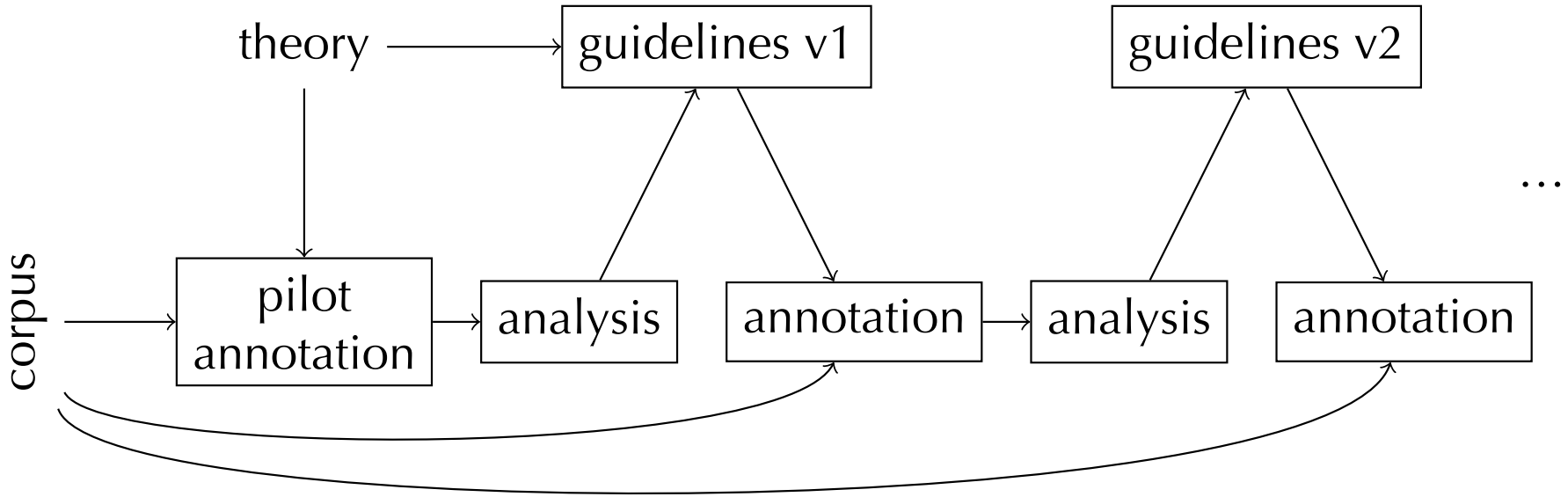
Data creation for developing automatic tools

- Test data for all kinds of automatic tools
  - All automatic tools should be tested systematically
  - Test data needed for rule-based and machine learning systems
- Training data for machine learning systems
  - Machine learning (tomorrow)
  - $\rightarrow$ systems usable on new, not yet annotated data

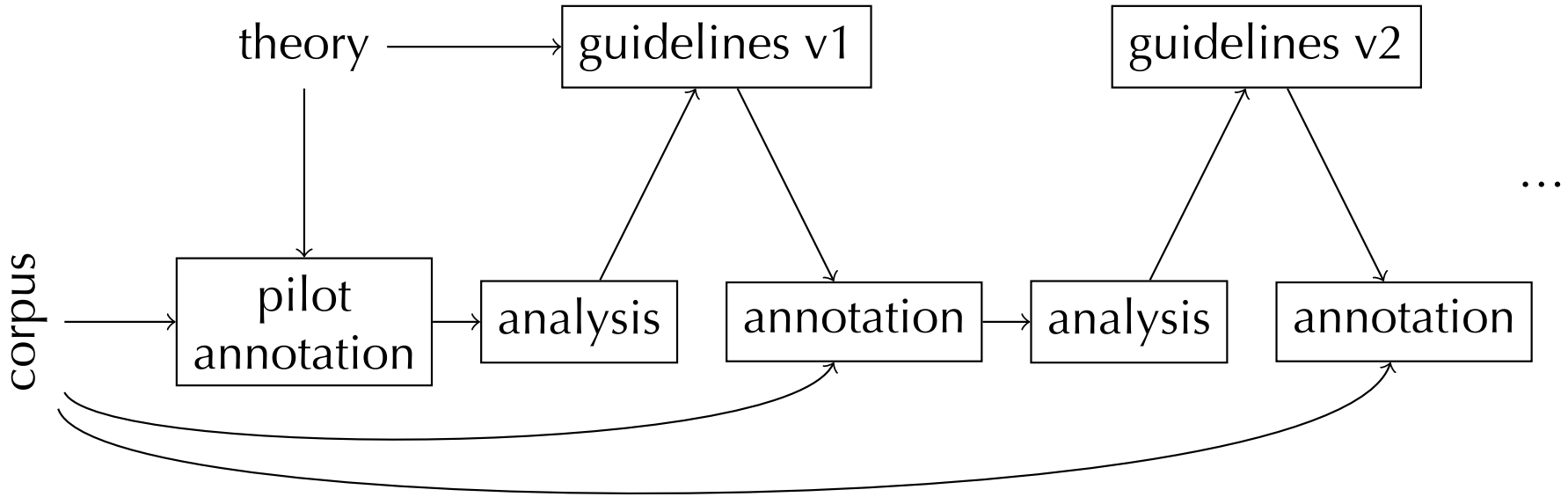
Section 6

How to Annotate

Annotation Workflow



Annotation Workflow



- Iterative process
- Annotating new texts tends to challenge your guidelines
- Perfection vs. 'good enough'

Parallel Annotation

- Annotation of the same text (parts) by multiple, independent annotators
- Allows comparison of the annotations
  - Increases reliability of the annotation process
  - Uncovers annotation guideline issues

Parallel Annotation

- Annotation of the same text (parts) by multiple, independent annotators
- Allows comparison of the annotations
  - Increases reliability of the annotation process
  - Uncovers annotation guideline issues

Who does the annotation?

Ideally: Independent persons, who don't have interests in specific outcomes

- Student assistants (paid)
- Crowd sourcing
- Students in class

Annotation Guidelines

- Instantiation of the (underlying) theory
- Objectivisation
- Annotators should annotate only on the basis of the guidelines (and their language understanding)

Annotation Guidelines

Examples

Stuttgart-Tübingen Tagset (STTS)

- Annotation guidelines for German parts of speech, used in large projects
- 11 top level categories:
  Nomen, Verben, Artikel, Adjektive, Pronomina, Kardinalzahlen, Adverbien, Konjunktionen, Adpositionen, Interjektionen, Partikeln
- → PDF on course page

Penn Treebank

- Guidelines for English parts of speech
- Not hierarchically organised, 36 tags in total
- → PDF on course page

Annotation Exercise

Exercise in student pairs

1. Annotate text 1 (individually)
2. Compare your annotations and discuss your disagreements
3. Make decision on how to avoid the the disagreements, take notes
4. Annotate text 2 (individually)

   back to 2.

Annotation Exercise

Exercise in student pairs

1. Annotate text 1 (individually)
2. Compare your annotations and discuss your disagreements
3. Make decision on how to avoid the the disagreements, take notes
4. Annotate text 2 (individually)

   back to 2.

The Task: Annotate entity references

- Specific objects that are distinguishable by naming in a real or fictional world
- Any category (characters, locations, objects, ...)
- Typically three linguistic forms: Pronouns, proper names, nominal phrases

Annotation Exercise

Exercise in student pairs

1. Annotate text 1 (individually)
2. Compare your annotations and discuss your disagreements
3. Make decision on how to avoid the the disagreements, take notes
4. Annotate text 2 (individually)

   back to 2.

The Task: Annotate entity references

- Specific objects that are distinguishable by naming in a real or fictional world
- Any category (characters, locations, objects, ...)
- Typically three linguistic forms: Pronouns, proper names, nominal phrases

Do it!

Section 8

Analysing Parallel Annotations

- Goal: High agreement
  - Based on the same (version of) guidelines, annotators should come to the same annotations
  - Achieved agreement used to measure guideline quality

Section 8

Analysing Parallel Annotations

- Goal: High agreement
  - Based on the same (version of) guidelines, annotators should come to the same annotations
  - Achieved agreement used to measure guideline quality
  - Is high agreement the only goal?

Inter-Annotator Agreement

- Why measuring?
  - To compare annotations across different configurations
    - (5 or 6 categories, 2 or 3 annotators, 10 or 20 instances)

Inter-Annotator Agreement

- Why measuring?
  - To compare annotations across different configurations
    - (5 or 6 categories, 2 or 3 annotators, 10 or 20 instances)
- How to measure agreement?
  - We don't know what's correct
    - IAA is a statement about the agreement, not about the correctness
  - Metric that works for arbitrary numbers of categories, annotators, instances

Annotation Analysis

Side note:
It doesn't hurt to actually talk to the annotators and ask them about their impressions!

Inter-Annotator Agreement

- Balancing observed and expected agreement
- Fleiss' κ Fleiss (1971)
  - Applicable for all *classification* tasks

Different Metrics

- Not all annotation tasks are the same
  - PoS tagging: Assign each word to a category
    - Only classification
  - Sentence splitting: Mark sentence boundaries
    - Only unitising
  - Named entities: Select a span *and* assign it to a category
    - Unitising, classification
- Different metrics for different tasks!

Cohen 1960; Fleiss 1971; Fournier and Inkpen 2012; Mathet et al. 2015

Gamma

Section 8

Analysing Parallel Annotations

Metric γ has been published in this paper:
Yann Mathet et al. "The Unified and Holistic Method Gamma ( ) for Inter-Annotator Agreement Measure and Alignment". In: *Computational Linguistics* 41.3 (2015), pp. 437–479

Three Components

- Combination of expected and observed agreement
- Calculation of observed agreement
- Calculation of expected agreement

Combining Expected and Observed Agreement

Note: γ is defined based on disagreements!
Assuming we have calculated observed (δ_o) and expected (δ_e) disagreement

$$\gamma = 1 - \frac{\delta_o}{\delta_e} \tag{1}$$

Combining Expected and Observed Agreement

$$\gamma = 1 - \frac{\delta_o}{\delta_e}$$

Combining Expected and Observed Agreement

$$\gamma = 1 - \frac{\delta_o}{\delta_e}$$

δ_o	δ_e	γ	
0.99	0.01	0.98	(upper bound: 1)
0.01	0.99	−98	(lower bound: $-\infty$)
0.5	0.25	−1	
0.5	0.5	0	
0.5	0.75	0.33	
0.25	0.5	0.5	
0.5	0.5	0	
0.75	0.5	−0.5	

Table: γ scores for observed (δ_o) and expected (δ_e) disagreement

Calculating Observed Agreement

Basics

- Local level: Measuring dissimilarity between two annotations
- Global level: Create unitary alignments over all annotations by all annotators

Calculating Observed Agreement

Situations

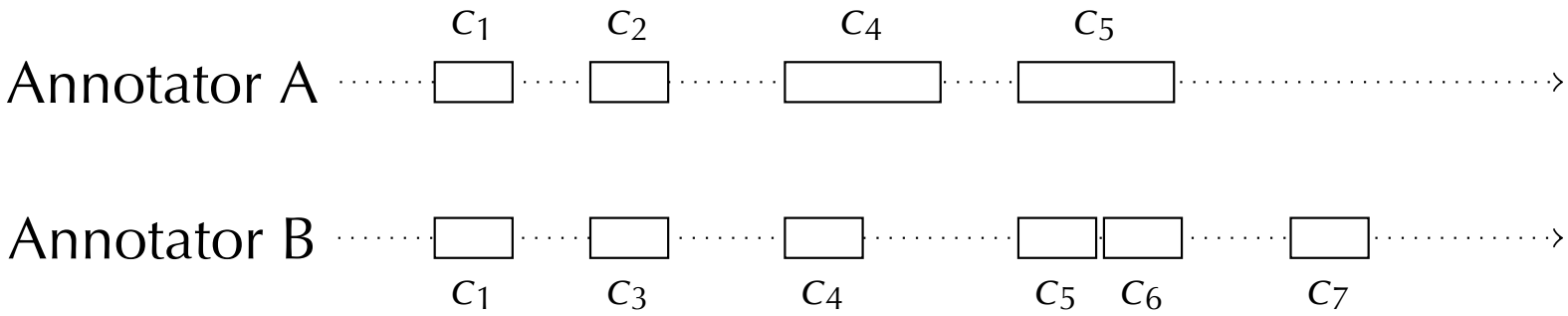


Figure: Two annotators and (some) possible situations

One Annotation is defined by

- begin/end
- feature values (including category)

If these are the same, we consider two annotations to be equal

Calculating Observed Agreement

Positional Dissimilarity

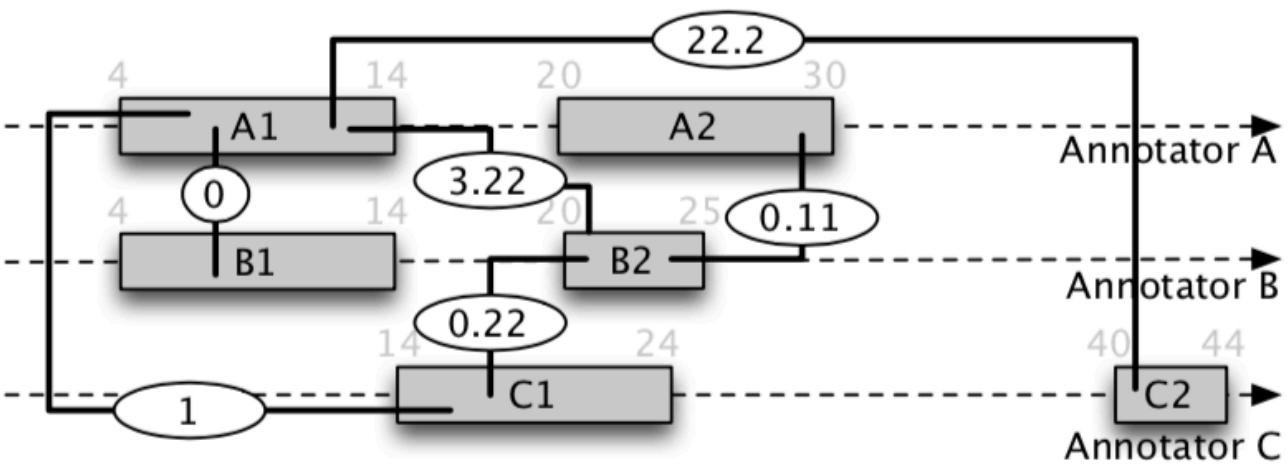
$$d_{pos}(u,v) = \left(\frac{\overbrace{|\mathsf{start}(u) - \mathsf{start}(v)|}^{\text{start difference}} + \overbrace{|\mathsf{end}(u) - \mathsf{end}(v)|}^{\text{end difference}}}{\underbrace{(\mathsf{end}(u) - \mathsf{start}(u))}_{\text{length of u}} + \underbrace{(\mathsf{end}(v) - \mathsf{start}(v))}_{\text{length of v}}} \right)^2$$

Calculating Observed Agreement

Positional Dissimilarity

$$d_{pos}(u,v) = \left(\frac{\overbrace{|\text{start}(u) - \text{start}(v)|}^{\text{start difference}} + \overbrace{|\text{end}(u) - \text{end}(v)|}^{\text{end difference}}}{\underbrace{(\text{end}(u) - \text{start}(u))}_{\text{length of u}} + \underbrace{(\text{end}(v) - \text{start}(v))}_{\text{length of v}}} \right)^2$$

Examples



Calculating Observed Agreement

Categorial Dissimilarity

Gamma

Define dissimilarity between categories in a matrix

	c_1	c_2	c_3
c_1	0	0.5	1
c_2	0.5	0	0.25
c_3	1	0.25	0

Calculating Observed Agreement

Combining Dissimilarity

$$d_{combi}(u, v) = \alpha d_{pos}(u, v) + \beta d_{cat}(u, v)$$

Calculating Observed Agreement

Combining Dissimilarity

$$d_{combi}(u, v) = \alpha d_{pos}(u, v) + \beta d_{cat}(u, v)$$

Intuitions and Remarks

- α and β can be used to express importance
  - $\alpha = \beta = 1$: Positional and categorial disagreement are equally important

Calculating Observed Agreement

Combining Dissimilarity

$$d_{combi}(u, v) = \alpha d_{pos}(u, v) + \beta d_{cat}(u, v)$$

Intuitions and Remarks

- α and β can be used to express importance
  - $\alpha = \beta = 1$: Positional and categorial disagreement are equally important
- Dissimilarity between two annotations is roughly between 0 (zero) and the squared length of the text (because of the positional dissimilarity)

Calculating Observed Agreement

Alignment

- Pairwise comparison of annotations ✓
- Which pairs do we compare?

Calculating Observed Agreement

Alignment

- Pairwise comparison of annotations ✓
- Which pairs do we compare?

Alignment

An alignment defines, which annotation of annotator 1 corresponds to which annotation of annotator 2 (if any)

Calculating Observed Agreement

Alignment

- Pairwise comparison of annotations ✓
- Which pairs do we compare?

Alignment

An alignment defines, which annotation of annotator 1 corresponds to which annotation of annotator 2 (if any)

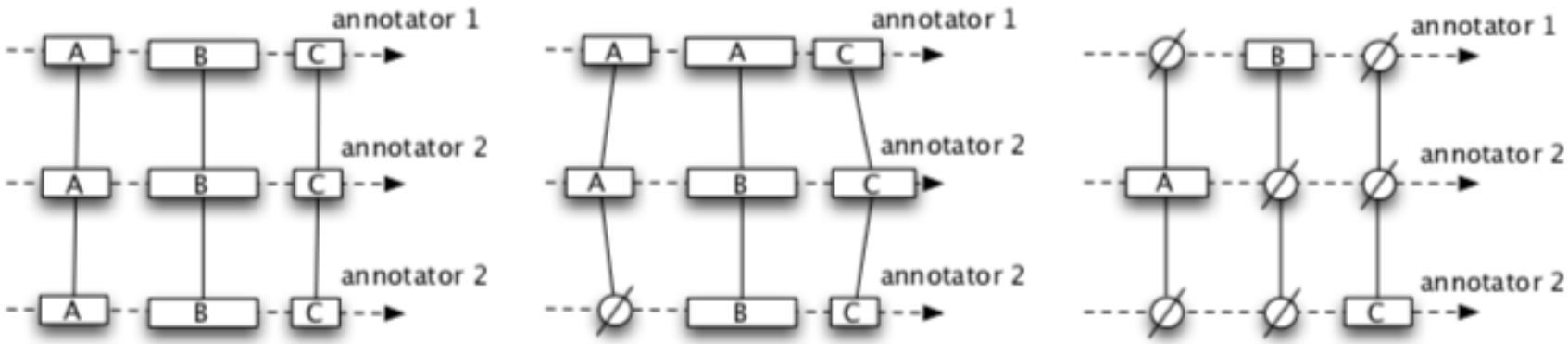


Figure: Different alignments between three annotators

Calculating Observed Agreement

Alignment: Two more ingredients

- Calculate disagreement over a set of aligned individual annotations: Average

Calculating Observed Agreement

Alignment: Two more ingredients

- Calculate disagreement over a set of aligned individual annotations: Average

$$\hat{\delta}(\hat{a}) = \frac{1}{|\hat{a}|} \sum_{(u,v)\in\hat{a}^2} d_{combi}(u, v)$$

with $\hat{a}$ being a set of aligned annotations

Calculating Observed Agreement

Alignment: Two more ingredients

- Calculate disagreement over a set of aligned individual annotations: Average

$$\hat{\delta}(\hat{a}) = \frac{1}{|\hat{a}|} \sum_{(u,v)\in\hat{a}^2} d_{combi}(u, v)$$

  with $\hat{a}$ being a set of aligned annotations
- Calculate disagreement over a set of annotators: Average

Calculating Observed Agreement

Alignment: Two more ingredients

- Calculate disagreement over a set of aligned individual annotations: Average

$$\hat{\delta}(\hat{a}) = \frac{1}{|\hat{a}|} \sum_{(u,v)\in\hat{a}^2} d_{combi}(u, v)$$

with $\hat{a}$ being a set of aligned annotations

- Calculate disagreement over a set of annotators: Average

$$\bar{\delta}(A) = \frac{1}{|x|} \sum_{i=1}^{|\hat{a}|} \hat{\delta}(\hat{a}_i)$$

with A being a set of annotators, and $|x|$ the mean number of annotations per annotator

Calculating Observed Agreement

Alignment: Two more ingredients

- Calculate disagreement over a set of aligned individual annotations: Average

$$\hat{\delta}(\hat{a}) = \frac{1}{|\hat{a}|} \sum_{(u,v)\in\hat{a}^2} d_{combi}(u, v)$$

  with $\hat{a}$ being a set of aligned annotations
- Calculate disagreement over a set of annotators: Average

$$\bar{\delta}(A) = \frac{1}{|x|} \sum_{i=1}^{|\hat{a}|} \hat{\delta}(\hat{a}_i)$$

  with A being a set of annotators, and $|x|$ the mean number of annotations per annotator
- Alignment is created such that $\bar{\delta}(A)$ is minimal

Calculating Observed Agreement

Summary

- Gamma combines alignment and agreement calculation
- Core: Compare annotations pairwise, w.r.t.
  - their position
  - their categories
- Settable parameters
  - Dissimilarity of categories
  - Weighting between dissimilarity types
  - Position metric (SANTA: token numbers)
- Computationally expensive
- Implementation by Mathet et al. (2015) using ILP
  `https://gamma.greyc.fr`

Calculating Expected Agreement

- Random annotations need to be *realistic* w.r.t. several criteria
  - Distribution of units per annotator
  - Distribution of categories
  - …
- □ 's expected disagreement is based on real annotations
  1. Take the annotations created by a real annotator
  2. Split the text at a random point
  3. Permute the two parts
  4. Repeat multiple times and calculate disagreement
- This doesn't work if the text only contains a single annotation that spans the entire text

Other Criteria for Annotation Quality

- IAA not the only criterion
- Perfect IAA is easy to achieve
  (by cheating)

Other Criteria for Annotation Quality

- IAA not the only criterion
- Perfect IAA is easy to achieve
  (by cheating)

Three dimensions

Gius et al. (to appear)

- Conceptual coverage: How much of a theory is represented in the guidelines?
- Applicability: How well can the guidelines be applied?
- Usefulness: How useful are annotations based on these guidelines for interpretation/subsequent analysis steps?

Other Criteria for Annotation Quality

- IAA not the only criterion
- Perfect IAA is easy to achieve
  (by cheating)

Three dimensions

Gius et al. (to appear)

- Conceptual coverage: How much of a theory is represented in the guidelines?
- Applicability: How well can the guidelines be applied?
- Usefulness: How useful are annotations based on these guidelines for interpretation/subsequent analysis steps?

Contradictory: Optimising one dimension hurts at least one other

Subsection 3

How to Write Guidelines

How to Write Guidelines (Gius/Willand/Reiter) I

- Preliminaries
  1. If your guideline is based on specific concepts or theories, specify them by referring to the concepts/theories and their authors.
  2. Give definitions for the phenomena you are addressing. Demarcate the phenomena from each other explicitly. This also may help to facilitate a scholarly discussion about the concepts or other people's decisions about whether re-using your guideline or data that has been annotated according to it.

How to Write Guidelines (Gius/Willand/Reiter) II

- Annotation instructions
  Defining the Annotation Span
  3. Define the span of text an annotation typically covers
     - E.g., a sentence, word, paragraph or something different)
  4. Define the borders of the annotations as exact as possible
     - E.g., specify whether to include/exclude interpunctuation, blanks at the beginning or end of a span etc.

How to Write Guidelines (Gius/Willand/Reiter) III

- Auxiliary indications
  5. Give positive and, if possible, also negative examples for each phenomenon. Text examples might help as well as schematic illustrations does.
  6. Name markers that indicate the presence of the phenomenon, if applicable
     - Think about syntactical, grammatical, semantical and other features that are typically connected to the phenomenon. E.g., specific words (as verbs with a specific semantic meaning, pronouns of a specific type etc.), tense, changes in mode or tense, preceding or subsequent phenomena etc.
  7. Provide tests the annotators can perform in order to detect the phenomena
     - E.g., when replacing X with Y…; when paraphrasing it to Z…;

How to Write Guidelines (Gius/Willand/Reiter) IV

- Organization of the Annotation Process
  8. Provide an overview of the annotation categories (or overviews of subsets of related annotation categories)
  9. If possible, organize the annotation routine from simple to complex phenomena
  10. Where present, point out dependencies between phenomena (and consider them in the ordering in step 9)

Summary

- Annotation: Metadatum
- Annotation guidelines, parallel annotations
  - Existing guidelines for many (linguistic) tasks Ide and Pustejovsky (2017)
  - Non-linguistic tasks not well covered
- Creating guidelines: Iterative process, increasing IAA
- Annotated data, annotation guidelines etc. are fundamental for anything you can do automatically

Part IV

Appendix

Bögel, Thomas, Michael Gertz, Evelyn Gius, Janina Jacke, Jan Christoph Meister, Marco Petris, and Jannik Strötgen. "Collaborative Text Annotation Meets Machine Learning: heureCLÉA, a Digital Heuristic of Narrative". In: *DHCommons* 1 (2015).

Cohen, Jacob. "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46.

Finlayson, Mark A. "ProppLearner: Deeply Annotating a Corpus of Russian Folktales to Enable the Machine Learning of a Russian Formalist Theory". In: *Digital Scholarship in the Humanities* (Dec. 2015). eprint: `http://dsh.oxfordjournals.org/content/early/2015/12/15/llc.fqv067.full.pdf`. URL: `http://dsh.oxfordjournals.org/content/early/2015/12/15/llc.fqv067`.

Fleiss, Joseph L. "Measuring nominal scale agreement among many raters". In: *Psychological Bulletin* 76.5 (1971), pp. 420–428.

Fournier, Chris and Diana Inkpen. "Segmentation Similarity and Agreement". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montrèal, Canada: Association for Computational Linguistics, 2012, pp. 152–161. URL: `http://aclweb.org/anthology/N12-1016`.

Gius, Evelyn, Nils Reiter, and Marcus Willand. "A Shared Task for the Digital Humanities: Evaluating Annotation Guidelines". In: *Cultural Analytics* Special Issue "Developing Guidelines for Annotating Narrative Levels as a Shared Task" (to appear).

Hovy, Eduard and Julia Lavid. "Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics". In: *International Journal of Translation Studies* 22.1 (Jan. 2010). URL: https://www.cs.cmu.edu/~hovy/papers/10KNS-annotation-Hovy-Lavid.pdf.

Ide, Nancy and James Pustejovsky, eds. *Handbook of Linguistic Annotation*. Springer, 2017. URL: https://www.springer.com/de/book/9789402408799.

Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier. "The Unified and Holistic Method Gamma ( ) for Inter-Annotator Agreement Measure and Alignment". In: *Computational Linguistics* 41.3 (2015), pp. 437–479.

Pagel, Janis, Nils Reiter, Ina Rösiger, and Sarah Schulz. "A Unified Text Annotation Workflow for Diverse Goals". In: *Proceedings of the Workshop on Annotation in Digital Humanities, co-located with ESSLLI 2018*. Ed. by Sandra Kübler and Heike Zinsmeister. Sofia, Bulgaria, Aug. 2018. URL: http://ceur-ws.org/Vol-2155/pagel.pdf.

Reiter, Nils, Evelyn Gius, Jannik Strötgen, and Marcus Willand. "A Shared Task for a Shared Goal - Systematic Annotation of Literary Texts". In: *Digital Humanities 2017: Conference Abstracts*. Montreal, Canada, Aug. 2017. URL: https://dh2017.adho.org/abstracts/192/192.pdf.

Reiter, Nils, Marcus Willand, and Evelyn Gius. "A Shared Task for the Digital Humanities: Introduction to Annotation, Narrative Levels and Shared Tasks". In: *Cultural Analytics* (to appear).

Siewert, Stephanie and Nils Reiter. "The Explorative Value of Computational Methods: Rereading the American Short Story". In: *American Studies* 63.2 (Oct. 2018). to appear.

Willand, Marcus, Jens Beck, and Nils Reiter. "Reading Data: On Digital Reception Studies". In: *Abstracts of EADH: Data in the Digital Humanities*. Galway, Ireland, Dec. 2018.