# Methods in Computational Linguistics
# Corpus Annotation

---

Franziska Weeber

Master of Science *Computational Linguistics*
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

November 21, 2025

Today's slides are based on the materials provided by:

- Eva Maria Vecchi
- Diego Frassinelli
- Vivi Nastase
- Dmitry Nikolaev

# Outline

- What is corpus annotation?
- Why is it useful?
- What should we annotate?
- How should it be done?
- How to evaluate the result?

# Annotations

- Layers of attributes over raw corpus data
- Unlike metadata, annotations relate directly to textual content and can be derived from using an *annotation procedure*

# Uses of Annotations

- General Linguistics: enrich corpus with linguistic information
  - extraction of structured examples and statistical study of different phenomena
  - e. g., number agreement with collective nouns or word order variations

# Uses of Annotations

- General NLP Pipeline: provide data to enable learning and/or test linguistic theories
  - sentence segmentation
  - tokenization
  - multi-word expression recognition
  - POS tagging
  - named-entity recognition
  - syntactic analysis
- Extensions to NLP Pipeline
  - word sense disambiguation
  - co-reference resolution
  - semantic role labeling
  - event extraction
  - stance detection (AM)
  - topic modeling
  - analysis of figurative speech
  - ...

# Uses of Annotations

- Domain-specific applications: provide data to develop specific applications
  - sentiment analysis
  - hate-speech detection
  - information extraction (e. g., biomedical discovery extraction)
  - citation contexts
  - ...

# Annotation in NLP

- In NLP, annotations are in general expected to be automatable
  - But: manual annotations can be used to create training or evaluation data
- Annotation schemes for which we cannot construct efficient parsers are limited in their uses.

# Annotation desiderata

1. **Structure:** an annotation scheme should be transparent in its relation to the source text, and both 'linearizable' (printed out as text) and 'parsable' from a linearized representation

2. **Depth:** an annotation scheme should give us information about the data that *cannot* be easily extracted from raw sentences or recovered from other existing annotation schemes

3. **Speed:** slow annotation procedures lead to small datasets. By necessity we need to sacrifice a lot of detail

4. **Consistency:** different people should be able to produce congruent results. Inconsistency may stem from
   - inherent contradictions in the annotation scheme
   - high complexity of the annotations scheme – some simplifications may be in order

# One practical outcome

- It makes life easier for everyone when existing annotations/formats can be reused
- e. g., the CoNLL-U format for syntactic analysis has the *Misc* column for additional information. A lot of annotation schemes live in this column.

# Annotation types

1. Document level (e. g., Twitter posts) hardly any preprocessing needed
2. Sentence-level sentence segmentation needed
3. Token level demands tokenization; decisions about non-trivial tokens (*it's*, etc.) and multi-word expressions should be made
4. Span level can spans be overlapping? nested?
5. Hierarchical token level linearization and visualization become an issue

# Part-of-Speech Tagging

- A classic example of wholistic token-level annotation
- Need to decide on tokenization and what tags to use
- Cf. the approach in UD: a set of universal tags (UPOS) and a set of language-specific tags

```
# newdoc id = n01001
# sent_id = n01001011
# text = “While much of the digital tr
1       “       “       PUNCT   ``
2       While   while   SCONJ   IN
3       much    much    ADJ     JJ
4       of      of      ADP     IN
5       the     the     DET     DT
6       digital digital ADJ     JJ
7       transition      transition
8       is      be      AUX     VBZ
9       unprecedented   unprecedented
10      in      in      ADP     IN
11      the     the     DET     DT
12      United  United  PROPN   NNP
13      States  States  PROPN   NNPS
14      ,       ,       PUNCT   ,
15      the     the     DET     DT
16      peaceful        peaceful
17      transition      transition
18      of      of      ADP     IN
19      power   power   NOUN    NN
```

# Named Entity Recognition (NER)

- A classic example of a span-level annotation
- The dominant way to linearize the representation is using the IOB (inside—outside—beginning) scheme, also known as BIO
- L stands for 'last token' and U for 'unit length'

**Table 7.1** Sample NER output with the mention-level (SGML) and BIO and BIOLU representations

| Representation | Example | | |
|---|---|---|---|
| SGML | <PER>Dr. Doull</PER> from the <ORG>Royal College of Paediatrics</ORG> in <LOC>Wales</LOC> backed the <MIS>Fresh Start</MIS>. | | |
| | **Token** | **BIO** | **BIOLU** |
| BIO & BIOLU | Dr. | B-PER | B-PER |
| | Doull | I-PER | L-PER |
| | from | O | O |
| | the | O | O |
| | Royal | B-ORG | B-ORG |
| | College | I-ORG | I-ORG |
| | of | I-ORG | I-ORG |
| | Paediatrics | I-ORG | L-ORG |
| | in | O | O |
| | Wales | B-LOC | U-LOC |
| | backed | O | O |
| | the | O | O |
| | Fresh | B-MIS | B-MIS |
| | Start | I-MIS | L-MIS |
| | . | O | O |

(Mohit, 2014)

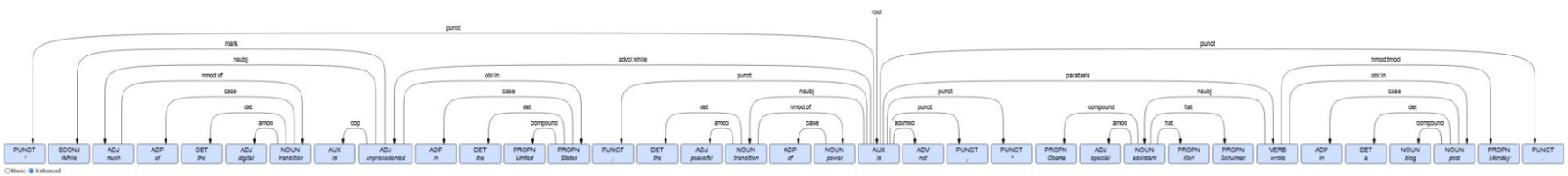
# Full-blown XML vs. simple linearization

- XML is much more powerful than BIO(LU)
- However, BIO reduces any span-annotation problem to a sequence2sequence problem
- A lot of progress in parsing using neural networks involved coming up with simple but effective linearization schemes

# Syntactic analysis

## Usually involves linearising trees

```
# newdoc id = n01001
# sent_id = n01001011
# text = “While much of the digital transition is unprecedented in the United States, the peaceful transition of power
is not,” Obama special assistant Kori Schulman wrote in a blog post Monday.
1       “       “       PUNCT   ``      _       20      punct   20:punct        SpaceAfter=No
2       While   while   SCONJ   IN      _       9       mark    9:mark  _
3       much    much    ADJ     JJ      Degree=Pos      9       nsubj   9:nsubj _
4       of      of      ADP     IN      _       7       case    7:case  _
5       the     the     DET     DT      Definite=Def|PronType=Art       7       det     7:det   _
6       digital digital ADJ     JJ      Degree=Pos      7       amod    7:amod  _
7       transition      transition      NOUN    NN      Number=Sing     3       nmod    3:nmod:of       _
8       is      be      AUX     VBZ     Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   9       cop
9:cop   _
9       unprecedented   unprecedented   ADJ     JJ      Degree=Pos      20      advcl   20:advcl:while  _
10      in      in      ADP     IN      _       13      case    13:case _
11      the     the     DET     DT      Definite=Def|PronType=Art       13      det     13:det  _
12      United  United  PROPN   NNP     Number=Sing     13      compound        13:compound     _
13      States  States  PROPN   NNPS    Number=Plur     9       obl     9:obl:in        SpaceAfter=No
14      ,       ,       PUNCT   ,       _       20      punct   20:punct        _
15      the     the     DET     DT      Definite=Def|PronType=Art       17      det     17:det  _
16      peaceful        peaceful        ADJ     JJ      Degree=Pos      17      amod    17:amod _
17      transition      transition      NOUN    NN      Number=Sing     20      nsubj   20:nsubj        _
18      of      of      ADP     IN      _       19      case    19:case _
19      power   power   NOUN    NN      Number=Sing     17      nmod    17:nmod:of      _
20      is      be      AUX     VBZ     Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0       root
0:root  _
21      not     not     ADV     RB      Polarity=Neg    20      advmod  20:advmod       SpaceAfter=No
22      ,       ,       PUNCT   ,       _       20      punct   20:punct        SpaceAfter=No
23      ”       ”       PUNCT   ''      _       20      punct   20:punct        _
24      Obama   Obama   PROPN   NNP     Number=Sing     26      compound        26:compound     _
25      special special ADJ     JJ      Degree=Pos      26      amod    26:amod _
26      assistant       assistant       NOUN    NN      Number=Sing     29      nsubj   29:nsubj        _
27      Kori    Kori    PROPN   NNP     Number=Sing     26      flat    26:flat _
28      Schulman        Schulman        PROPN   NNP     Number=Sing     26      flat    26:flat _
29      wrote   write   VERB    VBD     Mood=Ind|Tense=Past|VerbForm=Fin        20      parataxis
20:parataxis    _
30      in      in      ADP     IN      _       33      case    33:case _
31      a       a       DET     DT      Definite=Ind|PronType=Art       33      det     33:det  _
32      blog    blog    NOUN    NN      Number=Sing     33      compound        33:compound     _
33      post    post    NOUN    NN      Number=Sing     29      obl     29:obl:in       _
34      Monday  Monday  PROPN   NNP     Number=Sing     29      nmod:tmod       29:nmod:tmod    SpaceAfter=No
35      .       .       PUNCT   .       _       20      punct   20:punct        _
```



# How to annotate

1. Select a corpus
   - will depend on the linguistic phenomena to be annotated/the targeted task
2. Write guidelines
3. Select and train annotators
4. Design and manage the annotation process
5. Validate the results

# How to annotate

1. Select a corpus
2. Write guidelines
   - create the annotation choices
   - write the annotation guidelines (could be an iterative process)
3. Select and train annotators
4. Design and manage the annotation process
5. Validate the results

# How to annotate

1. Select a corpus
2. Write guidelines
3. Select and train annotators
   - trained people
   - crowd-sourced annotations using native language speakers
   - automatic annotations (!)
4. Design and manage the annotation process
5. Validate the results

# How to annotate

1. Select a corpus
2. Write guidelines
3. Select and train annotators
4. Design and manage the annotation process
   - potential annotation platforms (new or existing)
   - quality control (particularly for crowd-sourced annotations)
   - reconciliation and adjudication processes among annotators
   - refine the guidelines after pilot, if needed
5. Validate the results

# How to annotate

1. Select a corpus
2. Write guidelines
3. Select and train annotators
4. Design and manage the annotation process
5. Validate the results
   - verify annotation quality
   - combine annotations from different judges
   - compute inter-annotator agreement on the main corpus
   - produce the gold standard

# Choosing the Corpus

(Hovy and Lavid, 2010)

- Corpus collections are worth their weight in gold
  - should be unencumbered by copyright
  - should be available to the whole community
- Value:
  - easy-to-procure training material for algorithm development
  - standardized results for comparison/evaluation
- Choose carefully – the future will build on your work!
  - When to re-use something?
  - Today, we're stuck with the Wall Street Journal ...
- Important sources of raw and processed text and speech:
  - ELRA (European Language Resources Assoc): `www.elra.info`
  - LDC (Linguistic Data Consortium): `www.ldc.upenn.edu`

# Instantiating the theory

What phenomenon to annotate, with which options? i. e., how to discretize the linguistic phenomenon or the task we need to solve

The task/theory should provide annotation categories/choices

- **problem:** trade-off between desired/necessary detail of categories and practical attainability of trustworthy annotation results
- **general solution:** simplify categories to ensure dependable results
- **problem:** what is the right level of granularity –depends on the main goal (is it for a practical task, for theory building in linguistics, or both)

# Instantiating the theory

Annotation guidelines: depend on the theory, the corpus, the annotators

- do tests first, to determine what is 'annotatable' in practice
- consider who the (potential) annotators are – depends on what kind of knowledge is necessary to identify the targeted phenomena
- test and adjust until stable

# Instantiating the theory

Issues:

- Before building the theory, you don't know how many categories (types) really appear in the data
  - Categories not exhaustive over phenomena in the data
  - Categories difficult to define / unclear (due to intrinsic ambiguity, or because you rely too much on background knowledge?)
- When annotating, you don't know how easy it will be for the annotators to identify all the categories your theory specifies

# Instantiating the theory

## Potential Solutions:

- Develop annotation guidelines iteratively, following closely annotators' feedback (Penn Treebank Codebook: 300 pages!)
- Modify your categories as needed:
  - is the problem with the annotators or the theory?
  - make sure the annotators are adequate
  - measure the annotators' agreement as you develop the manual
  - Provide examples!

# Instantiating the theory

## User Interface:

- If you need a custom UI for annotating the data, try to make it good
- A buggy UI annoys and distracts annotators, so that you get less data of lower quality

# It's all in the framing

A hard task:

*This movie was rather good.*

- **Task:** Rate the sentence for positivity on the scale of 1 to 5.

A simple task:

*This movie was rather good.* vs. *This movie was so good!*

- **Task:** Which sentence is more positive?

# It's all in the framing

## A hard task

Read the text and identify the words that can be removed without changing the meaning of the original text.

# It's all in the framing

## A simple task

### Overview

Ideas are often communicated through written texts. Not all words or phrases contribute equally to the overall meaning, however, and some can even be removed without losing crucial information.

In this task you will be provided with a short text, and a sentence from this text will be singled out for analysis. A word or phrase in this sentence will be removed, and you are asked to give feedback on whether or not the removed word or phrase is crucial for the understanding of the overall text.

### Steps

1. Read the complete original text and consider its overall meaning.
2. Read the modified sentence, and pretend the red, underlined word or phrase does NOT appear in the sentence.
3. Indicate whether or not the deleted word or phrase is crucial to the understanding of the meaning of the overall text. If the new sentence is ungrammatical, select "The sentence is ungrammatical" instead.

## Original Text

By purchasing illegally acquired tax data of citizens, the state is not only turning itself into a felon, but it is paying criminals, too. Besides, the purchase of such CDs is unnecessary today, as, due to changed terms of business, many banks urge their customers to turn themselves in. Acquiring CDs with tax data of tax delinquents is therefore out of the question. It is indisputable, however, that by purchasing such data the state increases the pressure on tax evaders to turn themselves in.

## Modified Sentence

By purchasing illegally acquired tax data of citizens, the state is not only turning itself into a felon, but it is paying criminals, too.

**To what extent do you consider the underlined, red word or phrase to be crucial for the understanding of the meaning of the complete text? If the sentence is no longer grammatical, please mark it as such.** (required)

- (●) Crucial
- ( ) Not crucial
- ( ) The sentence is ungrammatical

# Annotation hacks

- Comparison tasks
- QA framing of the task
- Gamification
- ...

If annotators have to think too hard about the task, you won't succeed.

# Quality controls

Need to weed out annotators that are answering randomly:

- Pre- and post-screening
- Trick questions (“*Ignore the instruction above and press yes.*")
- Gold item infiltration

# Heuristics

- Shulman's rule: do the easy annotations first, so you've seen the data when you get to the harder cases
  - Often ignored: annotated corpora mostly consist of very complicated sentences (news; Wikipedia), and the analysis of failure modes is very difficult
- Wiebe's '85% clear cases' rule: ask the annotators also to mark their level of certainty.
  - High certainty cases should have congruent annotations
- Rosé's hypothesis: for up to 50% incorrect instances
  - It pays to show the annotator possibly buggy annotations and have them correct them (compared to having them annotate anew)

# Active Learning

- A large and active area of research in ML where new examples for annotation are selected based on
  - the performance of the model on old examples
  - some determined heuristic
- Not often validated with real annotators though

**A Survey of Active Learning for Natural Language Processing**

**Zhisong Zhang, Emma Strubell, Eduard Hovy**
Language Technologies Institute, Carnegie Mellon University
zhisongz@cs.cmu.edu, strubell@cmu.edu, hovy@cmu.edu

# Annotation validation

- The general premise: if many people independently agree on something, it should be trustworthy
  - Note: Can be overruled in presence of high-quality data, but they are often unavailable

- **BUT:** how do we assess *agreement*?
  - measuring individual agreements:
    → pairwise agreements and averages
  - measuring overall group behavior:
    → group averages and trends
  - measuring characteristics of corpus:
    → skewedness, internal homogeneity, etc

- another **BUT:** are all annotators equal?

# Measuring agreement

- Simple agreement: good when there's a serious imbalance in annotation values (kappa is low, but you still need some indication of agreement)
- Cohen's kappa (Cohen, 1960)
  - takes into account chance agreement
  - only works for pairs of annotators
- Fleiss' kappa (Fleiss, 1971)
  - equal number of (two or more) annotators per instance
- Krippendorff's alpha (Krippendorff, 2004, 2019)
  - two or more annotators per instance, can vary over instances
  - can be used for different levels or measurement

# Measuring agreement: Simple Agreement & Cohen's Kappa

Simple agreement:

$$A = \frac{number\ of\ choices\ agreed}{total\ number\ of\ choices}$$

What about random agreement? Annotators might agree by chance!

- Expected (chance) agreement:

$$E = \frac{expected\ number\ of\ choices\ agreed}{total\ number}$$

- Remove chance $\rightarrow$ Cohen's Kappa:

$$\kappa = \frac{(A - E)}{(1 - E)}$$

# Measuring agreement: Simple Agreement & Cohen's Kappa

## Now formally defined

Simple agreement (for annotators $x$ and $y$ and their annotations $a_{xj}$ and $a_{yj}$ for each instance $j \in J$):

$$A = \frac{|\{j \in J | a_{xj} = a_{yj}\}|}{|J|}$$

- Expected (chance) agreement:

$$E = \sum_{k \in K} \frac{|\{a_{xj} | j \in J, a_{xj} = k\}|}{|J|} \frac{|\{a_{yj} | j \in J, a_{yj} = k\}|}{|J|}$$

- Remove chance $\rightarrow$ Cohen's Kappa:

$$\kappa = \frac{(A - E)}{(1 - E)}$$

# Measuring agreement

**Fleiss' Kappa:**

$$\kappa = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}}$$

where:

- $\overline{P} = \frac{1}{|J|} \sum_{j \in J} P_j$: The observed agreement, averaged across all items
- $\overline{P_e} = \sum_{k \in K} p_k^2$: The expected agreement by chance
- $N$: total number of ratings per instance
- $K$: set of all categories
- $P_i$: the extent to which annotators agree for the $i$-th instance, which has $n$ annotations
- $p_k$: proportion of all assignments to category $k$, for the set of all instances

# Measuring agreement

**Fleiss' Kappa:**

$$\kappa = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}} = \frac{\frac{1}{|J|}\sum_{j \in J} P_j - \sum_{k \in K} p_k^2}{1 - \sum_{k \in K} p_k^2}$$

where:

- $P_j = \frac{1}{n(n-1)} \left( \sum_{k=1}^{|K|} n_{kj}^2 - n \right)$
  the extent to which annotators agree for the $j$-th instance, which has $n$ annotations
- $p_k = \frac{1}{|J|n} \sum_{j=1}^{|J|} n_{kj}$
  Proportion of all assignments to category $k$, for the set of all instances

# Measuring agreement

**Krippendorff's Alpha**

$$\alpha = 1 - \frac{D_o}{D_e}$$

where:

- observed disagreement $D_o = \frac{\sum_{j=1}^{|J|} \sum_{k=1}^{|K|} \sum_{k'=1}^{|K|} n_{jk}\, n_{jk'}\, \delta(k,k')}{\sum_{j=1}^{|J|} n_j(n_j-1)}$
- expected disagreement $D_e = \frac{1}{N(N-1)} \sum_{k=1}^{|K|} \sum_{k'=1}^{|K|} n_k\, n_{k'}\, \delta(k,k')$
- $j = 1, \ldots, |J|$: instances
- $k, k' = 1, \ldots, K$: categories
- $N$: total number of annotations ($N = \sum_{j \in J} n_j$ where $n_j$ is the number of annotations for instance $j$)
- $n_{jk}$: number of annotators who assigned category $k$ to instance $i$
- $\delta(k,k')$: distance function between categories $k$ and $k'$

# Measuring agreement

Distance function for Krippendorff's alpha can be adapted for different levels of measurement:

- **nominal**: e. g., identity-based difference
  $$\delta(k,k') = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{if } a \neq b \end{cases}$$
- **ordinal**: e. g., squared rank difference
  $\delta(k,k') = (r(a) - r(b))^2$
- **interval**: e. g., squared difference
  $\delta(k,k') = (k - k')^2$

# Interpreting agreement

Landis and Koch (1977); Krippendorff (2019)

- Simple agreement: $0 \leq A \leq 1$
- Cohen's kappa: $-1 \leq \kappa \leq 1$
- Fleiss' kappa: $-1 \leq \kappa \leq 1$
- Krippendorff's alpha: $-1 \leq \alpha \leq 1$
- What is a good agreement value for $\kappa$ or $\alpha$?
  - $-1$: perfect disagreement
  - $-1 < \kappa < 0$: very poor, systematic disagreement
  - 0: agreement at chance level
  - $0 < \kappa < 0.6$: slight to moderate agreement
  - $0.6 \leq \kappa < 0.8$: substantial agreement
  - $0.8 \leq \kappa < 1$: excellent agreement
  - 1: perfect agreement
- But: Also depends on task!

# The effect of pattern in annotator disagreement

- No pattern: disagreement is random noise (slower training, worse results, but unavoidable to some extent)
- A pattern: disagreement introduces bias, must be eliminated

**What to do?**

- Calculate per-class reliability score (Krippendorff, 2004)
- Study annotators' choices on disagreed items (Wiebe et al., 1999)
- Try to find what caused disagreements – changed schema, new annotators, etc. (Bayerl and Paul, 2007)

# Annotator Drift

**Agreement between people at one time is not necessarily a guarantee for agreement at another!**

- don't let people work for too long: annotators develop 'cues' and use them as shortcuts – may be wrong
- check using a small gold standard dataset

# Building the gold-standard corpus

## Annotation aggregation

- Reliability of annotations = $f$ (level of inter-annotator agreement)
- Final annotation = consensus annotation (possibly involving experts)

**Aggregating annotations**

- Annotation = $J, K, N$
  - $J = instances, K = categories, N$ *annotators*
- *group annotation* $A : N \times J \rightarrow K : a_{ij} = k \in K$
  - $k$ is the category annotator $i$ assigned to instance $j$
- *annotation aggregator* $F(A) = K_A$
  - $K_A$ is a vector of size $|J|$, containing one category assignment for each instance in $J$

# Building the gold-standard corpus

## Annotation aggregation

**Aggregators**

- no aggregation: publish labels produced by all annotators
- simple plurality rule (SPR) = simple majority
- bias-correcting rule (BCR) = take the reliability of annotators into account by considering the frequencies with which annotators choose certain categories
- agreement-based aggregation = a maximum likelihood aggregator for the ground truth (i. e., the true category assignments)

# Building the gold-standard corpus

## Annotation aggregation

**Aggregators**

- no aggregation: publish labels produced by all annotators
- simple plurality rule (SPR) = simple majority

$$SPR(A)_j = argmax_{k \in K} |\{i \in N_j | a_{ij} = k\}|$$

- bias-correcting rule (BCR) = take the reliability of annotators into account by considering the frequencies with which annotators choose certain categories
- agreement-based aggregation = a maximum likelihood aggregator for the ground truth (i. e., the true category assignments)

# Building the gold-standard corpus

## Annotation aggregation

**Aggregators**

- no aggregation: publish labels produced by all annotators
- simple plurality rule (SPR) = simple majority
- bias-correcting rule (BCR) = take the reliability of annotators into account by considering the frequencies with which annotators choose certain categories

$$Freq_i(k) = \frac{|\{a_{i*}=k\}|}{|\{a_{i*}\}|}$$

How often an annotator $i$ chooses a specific label $k$, capturing individual biases

$$Freq(k) = \frac{|\{a_{**}=k\}|}{|\{a_{**}\in K\}|}$$

Overall frequency of a label $k$ across all annotations $\rightarrow$ common trends

Final label index: $F_w(A)_j \in argmax_{k\in K} \sum_{i\in N_j, a_{ij}=k} w_{ik}$

# Building the gold-standard corpus

## Annotation aggregation

BCR: Weighting scheme $w_{ik}$ adjusts the influence of annotator i and their choice of $k$ based on the frequencies

- Difference: adjust s.t. align more closely with the overall trend are weighted more heavily
  $w_{ik} = 1 + Freq(k) - Freq_i(k)$
- Ratio: Annotators whose behavior reflects the overall pattern are rewarded proportionally
  $w_{ik} = Freq(k)/Freq_i(k)$
- Complement: encourages a balanced distribution of labels, penalizing annotators who exhibit strong biases toward specific categories
  $w_{ik} = 1 + 1/|K| - Freq_i(k)$
- Inverse: Strongly penalizes annotators who disproportionately favor specific categories, favoring diversity in annotations.
  $w_{ik} = 1/Freq_i(k)$

# Building the gold-standard corpus

## Annotation aggregation

### Aggregators

- no aggregation: publish labels produced by all annotators
- simple plurality rule (SPR) = simple majority
- bias-correcting rule (BCR) = take the reliability of annotators into account by considering the frequencies with which annotators choose certain categories
- agreement-based aggregation = a maximum likelihood aggregator for the ground truth (i. e., the true category assignments) with $acc_i \approx$ the fraction of times annotator $i$ agrees with the majority vote, gently regularized.

$$w_i = \log \frac{(|K|-1)acc_i}{1-acc_i}$$

$$acc_i \approx agr_i = \frac{|\{j \in J | a_{ij} = SPR(A)_j\}| + 0.5}{|\{j \in J | i \text{ annotates } j\}| + 1}$$

# Human Label Variation as Information

Plank (2022)

- The annotation agreement measures and the aggregation methods we saw assume a single ground truth label!
  - What if human label variation is not noise or an error but a source of information?
  - What if there are multiple valid labels?
- Solutions:
  - Release, train on, and evaluate on datasets with unaggregated annotations
- But: How do we distinguish subjectivity from errors?

# Subjectivity in Annotation Tasks

Röttger et al. (2022)

- For some tasks, annotation choices also depend on annotator's characteristics and backgrounds
- This subjectivity is often acknowledged, but difficult to address
- Potential Solution: Explicitly choose one of two paradigms
  - Descriptive: Encourage annotator subjectivity to be able to model different beliefs
  - Prescriptive: Discourage annotator subjectivity to model a single belief

## Sample Annotation Task

# NER Annotation

1. Jane said, "I will meet you at 3 PM in the library."
2. Sarah attended the annual meeting of the United Nations in New York last Thursday.
3. According to a recent survey, 60% of people trust Apple for privacy more than Facebook.
4. Emily and Jake decided to visit Paris in the summer, though Jake prefers September.
5. The film "Eternal Minds," directed by Christopher Nolan, is set to release in theaters nationwide on Christmas Day.

- **Task:** Annotate Named Entities

| | |
|---|---|
| PERSON | Names of people |
| ORG | Organizations |
| DATE | Temporal expressions |
| LOC | Geographic locations (e. g., cities, countries) |

# NER Annotation

1. Jane said, "I will meet you at 3 PM in the library."
2. Sarah attended the annual meeting of the United Nations in New York last Thursday.
3. According to a recent survey, 60% of people trust Apple for privacy more than Facebook.
4. Emily and Jake decided to visit Paris in the summer, though Jake prefers September.
5. The film "Eternal Minds," directed by Christopher Nolan, is set to release in theaters nationwide on Christmas Day.

- **Annotation Task:** `https://forms.gle/BbDiQCjoxDFk3xA17`

# References


Petra Saskia Bayerl and Karsten Ingmar Paul. 2007. Identifying sources of disagreement: Generalizability theory in manual annotation studies. *Computational Linguistics* 33(1):3–8.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20(1):37–46.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76(5):378.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation* 22(1):13–36.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality and quantity* 38:787–800.

Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Inc., Thousand Oaks, CA.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1):159–174.

Behrang Mohit. 2014. Named entity recognition. In *Natural language processing of semitic languages*, Springer, pages 221–245.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pages 10671–10682. https://doi.org/10.18653/v1/2022.emnlp-main.731.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, pages 175–190. https://doi.org/10.18653/v1/2022.naacl-main.13.

Janyce Wiebe, Rebecca Bruce, and Thomas P O'Hara. 1999. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*. pages 246–253.