SUBJECTIVITY MINING (HATE SPEECH)

ISA MAKS

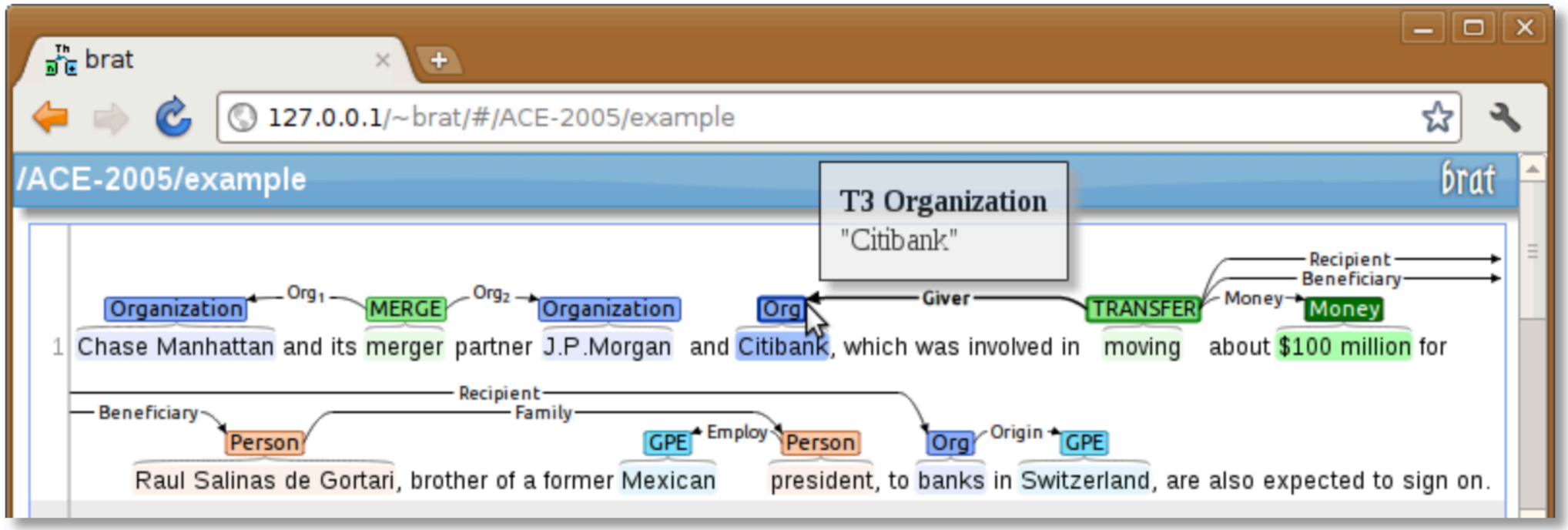
LINGUISTIC ANNOTATIONS FOR SUBJECTIVE EXPRESSIONS

WHY ANNOTATE

- Giving meaning and interpretation to signals (text)
- Leading to a systematic analysis
- Reproducibility
- Automatic annotations (training, testing)

ANNOTATION TOOLS

- Facilitate complex annotations
- Monitor agreement



Annotation visualization

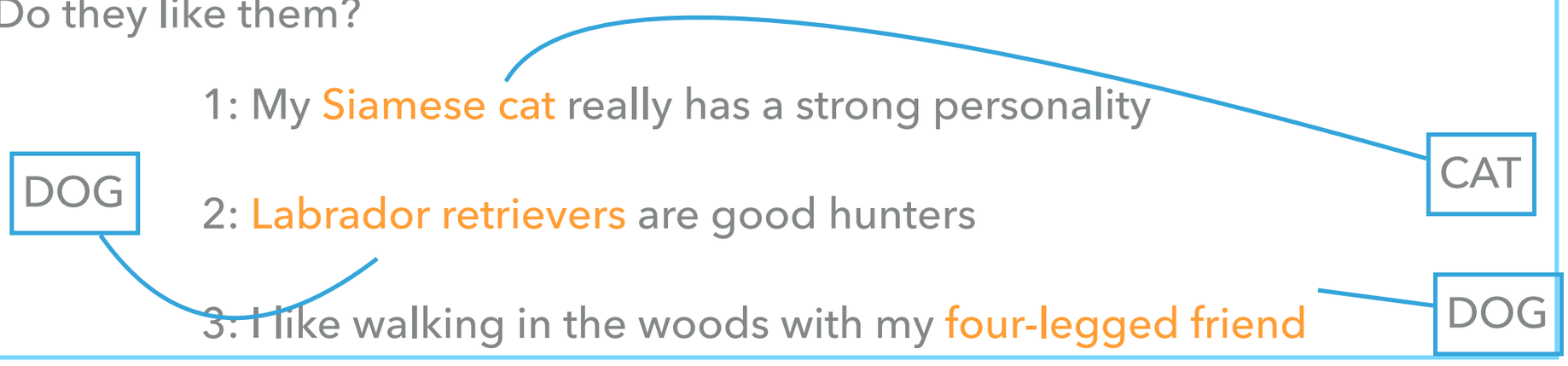
ANNOTATION OF TEXT

- Adding interpretive information into a (collection of) texts
  - In order to extract information
  - That can answer a research question

BG: *We know that animals are important for people. Cuddling a pet reduces stress, loneliness and anxiety. Cats and dogs are the most common pets*

RQ1: **Do they write about their cats and dogs?**

RQ2: Do they like them?



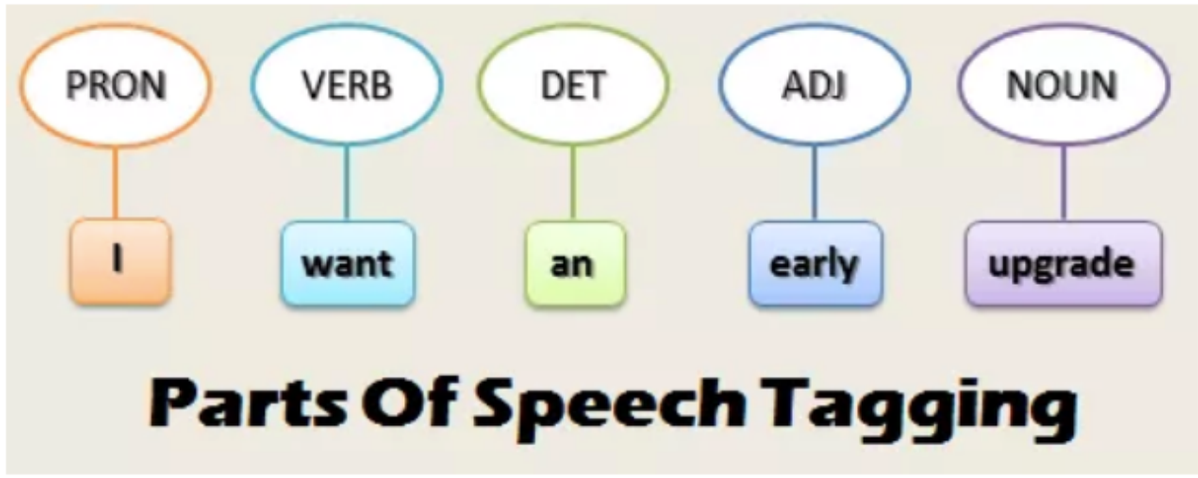
INSTANTIATING THE THEORY

- Every annotation instantiates some theory
- No theory is ever complete: not covering all phenomena
- Few theory developed to an equal degree for all variants they address

  - Verb: tells about an action or a state
  - Adjective: words that describe or modify other words
  - Noun: words that identify people, locations and things

This is a beautiful village not far from the sea. The rich and poor live together here.

PART OF SPEECH THEORY



By the end of the 2nd century BCE, grammarians had expanded this classification scheme into eight categories, seen in the *Art of Grammar*, attributed to Dionysius Thrax:[9]

1. Noun (*ónoma*): a part of speech inflected for case, signifying a concrete or abstract entity
2. Verb (*rhêma*): a part of speech without case inflection, but inflected for tense, person and number, signifying an activity or process performed or undergone
3. Participle (*metokhḗ*): a part of speech sharing features of the verb and the noun
4. Article (*árthron*): a declinable part of speech, taken to include the definite article, but also the basic relative pronoun
5. Pronoun (*antōnymíā*): a part of speech substitutable for a noun and marked for a person
6. Preposition (*próthesis*): a part of speech placed before other words in composition and in syntax
7. Adverb (*epírrhēma*): a part of speech without inflection, in modification of or in addition to a verb, adjective, clause, sentence, or other adverb
8. Conjunction (*sýndesmos*): a part of speech binding together the discourse and filling gaps in its interpretation

The door is closed (adjective/past participle)

IN CASE OF SUBJECTIVE PHENOMENA: IS THERE A THEORY?

- We do not have a comprehensive linguistic theory about how exactly people express their private states (beliefs, opinions, attitudes, emotions, stances, etc.)
- We do have initial models or schemas
- We do have a lot of research though that tells us about lot of characteristics of subjective language for which we know what the issues are
- =>*The classes emerge from the material*

CHARACTERISTICS OF SUBJECTIVE LANGUAGE

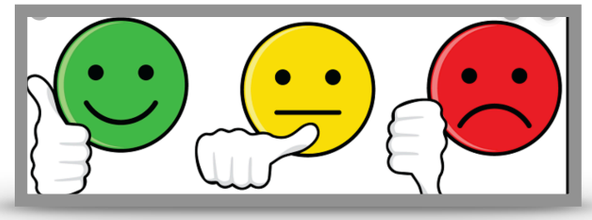
- Concepts and their meaning: sad, beautiful, enjoy, idiot
- Words with connotations: a doctor vs. A quack selling medicines
- Adjectives: this is fantastic
- Gradation: this is very nice
- Modal expressions : you should go home
- Opinion introducing verbs (he thinks that)
- Alignment verbs (I agree with you)
- Positive and negative concepts (deaths, parties)
- Repetition of words (this really really fantastic)
- Informal language and personal language : damn you

This is not a theory

SUBJECTIVITY ANNOTATIONS

Simple (simplified) models

- Sentiment



- Stance

- Opinions

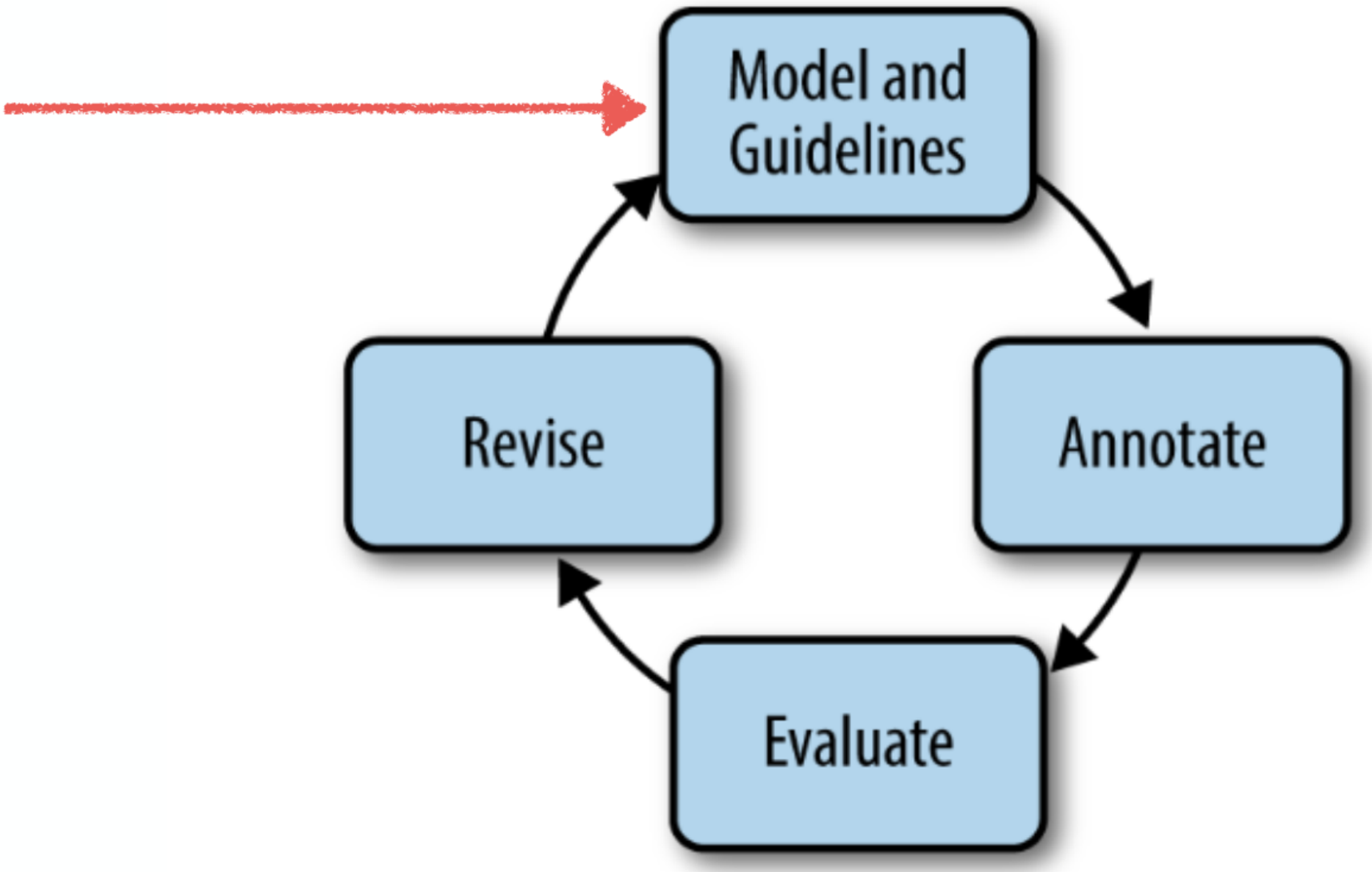
WHO THINKS WHAT ABOUT WHAT/WHOM

- Emotions

- Hatespeech: ??offensive, hateful, harmful, abusive ??

The process of *annotation*

The full annotation process can be represented by the MEAR cycle (Pustejovsky and Stubbs, 2012)



WHO IS THE BEST ANNOTATOR?

- Self reporting : subjects are asked to describe situations associated with a specific emotion (Scherer and Walbott, 1994)
- Distant supervision (star ratings, hashtags)
- Crowdsourcing
- Expert annotators, trained annotators
- In group vs. out group annotators

ANNOTATION PERSPECTIVE

- Writer: what emotion is expressed by the writer
- Reader: what does the text evoke? What do you feel after reading the text? (Poetry, lyrics)
- *Text, Between reader and writer t: What emotion is conveyed by the text
- *Annotator?*
  - *The Hotel is situated in. The city center near to all main attractions*
  - *Italy defeats France in World Cup Final;Italy wins the World Cup Final*
  - *House prices are rising*
  - *I believe the problem is Europe wide, but I feel Malta is harder hit simply because of its size. Malta simply has not got the room to keep receiving these people[1]*
  - *We just need to teach our children to respect each other whoever they are. Teaching gender diversity is too complicated and abnormal for their standards. I'm afraid it would effect their perception that lgbti is the norm instead of a minority[1]*

1: from MaNeCo corpus

TEXT

Contextual knowledge
Maclachlan and Reid (1994)

MacLachlan, G. and Reid, I. (1994) Framing and Interpretation. Melbourne: Melbourne University Press.

- Intratextual: any knowledge extracted from the text (names, people, events) including grammatical knowledge (parts-of-speech and their order and internal structure

- Extratextual: any reference to world knowledge: people, events, etc. (entity oriented knowledge found outside the document)

- Circumtextual: circumstances through which the text has come into existence (time, place, author, platform, publisher)

- Intertextual: cross-document knowledge (a stream of information or news, a twitter thread, a conversation)

COMPONENTS OF ANNOTATION GUIDELINES

- Background and motivation
- Identification and description of the phenomenon
- Specification of the model (tags, definitions, relations between tags)
- Rules for annotations (defining the text span of annotations)
- Examples of annotations
- (Description of the dataset to be annotated)

RULES FOR ANNOTATION

- Conceptual rules:
  - `indicate per document whether it is PROSE or POETRY
  - If the genre is POETRY you have to indicate whether it is FREE or NON FREE
  - In case you cannot decide easily, choose POETRY (!!) *
- Practical rules:
  - Select all words of the document and assign the label PROSE or POETRY

INTER-ANNOTATOR AGREEMENT STUDY

- Inter-coder reliability refers to the extent to which two or more coders agree on the coding of the **content of interest** with an application of the same **coding scheme** applied **independently** from each other

- The extent to which a measurement procedure yields the same answer however and whenever it is carried out

- Reliability tells you how consistently a method measures something. When you apply the same method to the same sample under the same conditions, you should get the same results. If not, the method of measurement may be unreliable. https://www.scribbr.com/methodology/types-of-reliability/

WHY

- Are the results reliable enough to :
  - Draw conclusions regarding the research questions
  - Use the annotations for machine learning
  - Use the schema and model again

DIFFERENT TYPES OF TESTS

- Test-retest reliability (stability) :
  - *the extent to which a coding procedure yields the same results on repeated trials. An annotator annotates the same document again after some time (ex.: the door is closed)*
- Internal consistency reliability:
  - *the extent to which an annotator is consistent in applying the coding guidelines, treating similar textual items in the same way throughout the corpus (ex.: the rich and poor live together (noun or adjective))*
- =>Inter-rater reliability:
  - *the extent to which the process can be replicated*

COMPONENTS OF AN IAA STUDY

- Corpus statistics:
  - An overview of the numbers of texts, the numbers of annotations per annotator
- Inter-coder agreement scores and analysis
  - An overview of the inter-annotator agreement scores for the different annotation categories (i.e. “markables” in CAT)
  - A description and interpretation of the results
- Error analysis:
  - Compare the annotations of the independent coders and find out the differences
  - Group the errors in error types and discuss them (ie. find out why they occurred)
- Summary
  - Make suggestions for improvements of the annotation model and guidelines

CONFUSION MATRIX

Sentiment

\ Annotator 1 Annotator 2 \	pos	neg	ntr	mix
pos	24	12	3	5
neg	4	124	11	9
ntr			6	3
mix				

Ntr is confused easily with neg (11 of 20 for Ann1)

Mix is confused easily with all other categories

Pos and neg not easily confused although there are differences between annotators

PERCENTAGE AGREEMENT

	(B)Blue	(B)Yellow	
(A)Blue	**20**	**5**	25
(A)Yellow	**10**	**15**	25
	30	20	50

#agreements/items

(20+15)/50=35/50=0.7

KAPPA SCORES

- Simple example (https://en.wikipedia.org/wiki/Cohen%27s_kappa)
- Takes into account agreement occurring by chance

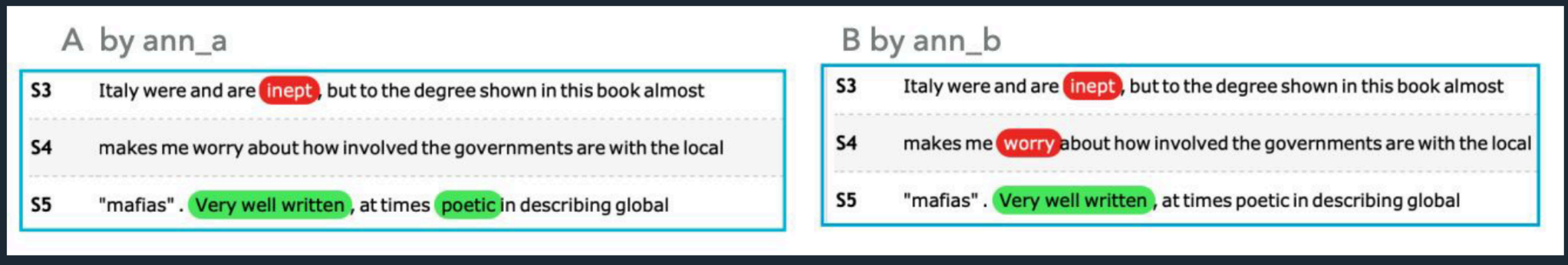
$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

pO observed agreement

pE expected agreement

Agreement Metrics

- Fleiss' Kappa: Generalizes Cohen's Kappa to multiple raters. The basic idea is to consider each pairwise agreement of raters and average over all items.
- Weighted Kappa: Not all disagreements are equally bad.
- Dice: Can be applied with different annotation spans.



EVALUATING HIERARCHICAL ANNOTATIONS

EVALUATION OF HIERARCHICAL SCHEMAS

- Different numbers per level

	Percentage Agreement	Cohen's Kappa	N
de Gibert et al., 2018	77.3%	0.382	44
Kumar et al., 2018	77.3%	0.621	44
Zampieri_Offensive	79.5%	0.573	44
Zampieri_Target	86.4%	-0.065	22
Zampieri_Identification	84.2%	0.708	19
Zampieri_all	65.9%	0.520	44

Table 1: Inter-annotator agreement scores for the three annotation schemes on the dataset. Note that since the annotation scheme by Zampieri et al., 2019 is hierarchical, per-tag agreement has also been computed.

	OFF	NON
OFF	22	4
NON	5	13

	TARGET	NO TARGET
TARGET	19	1
NOTARGET	2	0

	Other	Group	Indiv.
Other	1	1	0
Group	1	10	1
Indiv.	0	0	5

Table 4: Confusion tables for the annotations made using the Zampieri et al., 2019 annotation scheme.

ERROR ANALYSIS

- Systematic analysis
- Agreement and disagreement per category
- Try to explain what has gone wrong
- Start from confusion matrix and IAA scores

- Take 100 cases of disagreements
  - Try to make categories of disagreements
  - And calculate the number of disagreements per category

WHAT CAN WE DO WHEN AGREEMENT IS TOO LOW

- The guidelines are not clear
  - Re-write the guidelines
  - Provide examples for difficult cases

- The task is too hard
  - Re-define and simplify the task
  - Merge problematic categories
  - Remove problematic categories
  - Use a different set of values (e.g. smaller scale)

- Annotators need more training

- The guidelines are biased to one set of texts (and applied on another type)
  - Expand the guidelines

- Include only annotations with high agreement (?!)

- What should we NOT do: genuine ambiguity should be preserved

SENTIMENT ANALYSIS OF HOTEL REVIEWS: 2 LABELS

raters	number of reviews	kappa	percentage
REV-R1	1172	81.1	90.5
REV-R2	1172	81.6	90.7
R1-R2	1172	91.6	95.8

REV=reviewer;R1=reader1;R2=reader2

Kappa and percentage agreement between reader and reviewer ratings (REV-R1 and REV-R2) and between mutual reader ratings(R1-R2)

Table 5.2: Interannotator-agreement on hotel reviews

- Binary task : positive or negative
- Hotel reviews

  e.g. *The hotel seems rather outdated. The breakfast room is not big enough to cope with the Sunday morning crowds.*

Maks (2018) The automatic acquisition of a Dutch lexicon for opinion mining

SENTIMENT IN HEADLINES: 3 LABELS

- Boukes et al. (2020) 300 headlines ; 3 annotators; positive /negative / neutral
- Average kappa agreement 0.75
- *Greeks are stone-broke again*
- *House prices are rising in the Netherlands*

GERMEVAL TASK 2 (2019)

- Identification of offensive language
- 4 categories: profane, insult, abusive, other
- GERMEVAL (2018) 0.66 240 tweets ; (2019) 0.59 206 tweets
- 4 annotators; fleiss kappa

		training set		test set	
categories		**freq**	**%**	**freq**	**%**
coarse-grained	OFFENSE	1287	32.2	970	32.0
	OTHER	2707	67.8	2061	68.0
fine-grained	ABUSE	510	12.8	400	13.2
	INSULT	625	15.6	459	15.1
	PROFANITY	152	3.8	111	3.7
	OTHER	2707	67.8	2061	68.0
total		3994	100.0	3031	100.0

Table 2: Class distribution on the 2019 training and test set

Struss et al.(2019) https://www.researchgate.net/publication/337441720_Overview_of_GermEval_Task_2_2019_Shared_Task_on_the_Identification_of_Offensive_Language

FORTUNA ET AL. (2019)

- Hierarchical labeling
-

Class
Sexism
Body
Origin
Homophobi
Racism
Ideology
Religion
Health
Other-Lifestyle

Table 1: Direct

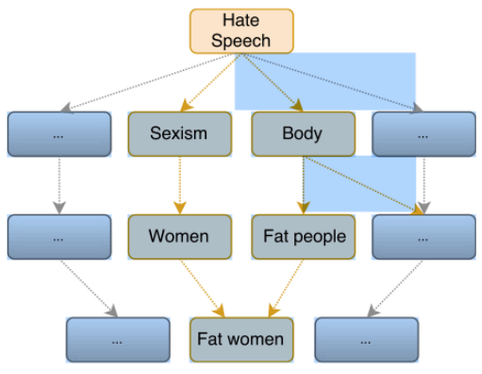


Figure 2: Part of the rooted directed acyclic graph used for hate speech classification.

Classes	K	Annotator 1	Annotator 2
Lesbians	0.879	59	53
Health	0.856	3	4
Homofobia	0.823	69	61
Disabled people	0.799	2	3
Refugees	0.763	13	13
Migrants	0.751	15	14
Sexism	0.669	134	104
Trans women	0.662	6	9
Men	0.657	12	15
Women	0.642	109	75
Fat women	0.637	30	16
Body	0.637	32	17
Fat people	0.637	32	17
Ideology	0.609	14	15
Feminists	0.581	13	14
Hate speech	0.569	245	213
Racism	0.501	18	13
Religion	0.493	5	11
Black people	0.435	11	7
Origin	0.329	3	3
Islamists	0.329	2	10
Gays	0.300	4	9
Ugly women	0.276	24	4

Table 2: Annotator agreement by class, with the number of messages annotated by each annotator.

Fortuna et a.(2019) A hierarchically labeled Portuguese hate speech dataset; https://aclanthology.org/W19-3510.pdf

GIBERT ET AL.: HATE VS. NO HATE

- Internet forum posts: Sentences extracted from Stormfront ( ), a white supremacist forum. (largest online community of white nationalists)
- Definition:
  - ✦ A deliberate attack
  - ✦ Directed towards a specific group of people
  - ✦ Motivated by actual or perceived aspects that form the group’s identity
- Annotations:
  - Skip: not written in English
  - Hate: see definition (all premises are true)
  - noHate: see definition
  - Relation: only the combination of sentences conveys hate speech

FORTUNA ET AL. (2019)

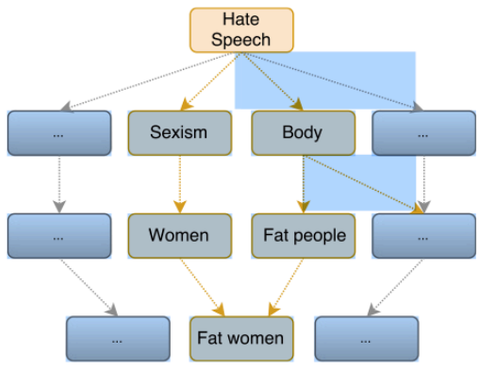


Figure 2: Part of the rooted directed acyclic graph used for hate speech classification.

Classes	K	Annotator 1	Annotator 2
Lesbians	0.879	59	53
Health	0.856	3	4
Homofobia	0.823	69	61
Disabled people	0.799	2	3
Refugees	0.763	13	13
Migrants	0.751	15	14
Sexism	0.669	134	104
Trans women	0.662	6	9
Men	0.657	12	15
Women	0.642	109	75
Fat women	0.637	30	16
Body	0.637	32	17
Fat people	0.637	32	17
Ideology	0.609	14	15
Feminists	0.581	13	14
Hate speech	0.569	245	213
Racism	0.501	18	13
Religion	0.493	5	11
Black people	0.435	11	7
Origin	0.329	3	3
Islamists	0.329	2	10
Gays	0.300	4	9
Ugly women	0.276	24	4

Table 2: Annotator agreement by class, with the number of messages annotated by each annotator.

MANECO CORPUS (2020)

1. Does the post communicate a positive, negative or neutral attitude? [**Positive / Negative / Neutral**]

2. If negative, who does this attitude target? [**Individual / Group**]

   (a) If it targets an individual, does it do so because of the individual's affiliation to a group? [**Yes / No**] If yes, **name the group**.

   (b) If it targets a group, **name the group**.

3. How is the attitude expressed in relation to the target group? Select all that apply. [**Derogatory term / Generalisation / Insult / Sarcasm (including jokes and trolling) / Stereotyping / Suggestion / Threat**][4]

(a) If the post involves a suggestion, is it a suggestio that calls for violence against the target group [**Yes / No**]

TOXIC SPAN DETECTION TASK

- Semeval 2021

Post	Offensive Spans
Stupid hatcheries have completely fucked everything	[0, 1, 2, 3, 4, 5, 34, 35, 36, 37, 38, 39]
Victimitis: You are such an asshole.	[28, 29, 30, 31, 32, 33, 34]
So is his mother. They are silver spoon parasites.	[]
You're just silly.	[12, 13, 14, 15, 16]

Table 1: Four comments from the dataset, with their annotations. The offensive words are displayed in red and the spans are indicated by the character position in the instance.

QUESTIONS