Modeling Ambiguity with Many Annotators and Self-Assessments of Annotator Certainty

Melanie Andresen
Institute for Natural Language Processing
University of Stuttgart
melanie.andresen@ims.uni-stuttgart.de

Michael Vauth
Institute of Linguistics and Literary Studies
TU Darmstadt
vauth@linglit.tu-darmstadt.de

Heike Zinsmeister
Institute for German Language and Literature
University of Hamburg
heike.zinsmeister@uni-hamburg.de

Abstract

Most annotation efforts assume that annotators will agree on labels, if the annotation categories are well-defined and documented in annotation guidelines. However, this is not always true. For instance, content-related questions such as 'Is this sentence about topic X?' are unlikely to elicit the same answer from all annotators. Additional specifications in the guidelines are helpful to some extent, but can soon get overspecified by rules that cannot be justified by a research question. In this study, we model the semantic category 'illness' and its use in a gradual way. For this purpose, we (i) ask many annotators (30 votes per item, 960 items) for their opinion in a crowdsourcing experiment, (ii) ask annotators to indicate their certainty with respect to their annotation, and (iii) compare this across two different text types. We show that results of multiple annotations and average annotator certainty correlate, but many ambiguities can only be captured if several people contribute. The annotated data allow us to filter for sentences with high or low agreement and analyze causes of disagreement, thus getting a better understanding of people's perception of illness—as an example of a semantic category—as well as of the content of our annotated texts.

1 Introduction

Natural language is full of phenomena of ambiguity and uncertainty. However, we do not yet have a standard procedure for integrating ambiguities in formal models—or even for identifying ambiguities in the first place. Most supervised machine learning tasks assume that there is a ground truth—an inter-subjectively correct annotation. Algorithms rely on this as unambiguous training data. For the most part, annotation efforts assume that multiple annotators will agree on labels, if annotation categories are well-defined and the annotation guidelines are clear and comprehensive. Hence, low agreement scores are considered to indicate poor data quality. However, there is also an increasing awareness that this is not always the case (Poesio and Artstein, 2005; Beigman Klebanov et al., 2008; Morris, 2010; Rohde et al., 2016; Amidei et al., 2018; Pavlick and Kwiatkowski, 2019). This is backed up by findings in cognitive science and linguistics that suggest that language phenomena are predominantly gradual in nature instead of being discrete categories, for instance in prototype theory (Lakoff, 1987).

The 14th Linguistic Annotation Workshop, pages 48–59
Barcelona, Spain, December 12, 2020.

Two common possible ways of capturing ambiguity in text are asking more than one annotator to annotate a category and/or asking annotators to explicitly mark ambiguities or uncertainties. In the study conducted for this paper, we combine both strategies to model the semantic category 'illness' and its use in a gradual way. Illness lends itself to this type of analysis, because it is a highly socially constructed concept that makes ambiguities and disagreements likely. Furthermore, this topic is of interest to linguists, humanists and social scientists alike and thus facilitated cooperation in our project *hermA* (Gaidys et al., 2017). We use crowdsourcing in order to get the input of many annotators and use their overall vote as a measure of how clearly a sentence belongs to the topic 'illness'. In addition, we ask annotators to indicate their certainty with respect to their annotation. Our aim is to investigate to what extent multiple annotations and self-assessments yield similar results, deriving recommendations for future research. Furthermore, we identify reasons for disagreement in a qualitative analysis contrasting controversial and non-controversial sentences. Our data comprise sentences from German literary texts and transcripts of political debates, thus allowing for a comparison of text types.

2 Related Work

In this section we present previous research dealing with ambiguity in annotation settings with multiple annotators. With regard to agreement and reliability, Potter and Levine-Donnerstein (1999) (in the context of content analysis) differentiate three types of content to be annotated: '*manifest content* (directly observable events), *pattern latent content* (events that need to be inferred indirectly from the observations), and *projective latent content* (loosely said, events that require a subjective interpretation from the annotator)' (Reidsma and op den Akker, 2008, 8, summarizing Potter and Levine-Donnerstein, 1999, 261). The type of question asked in this paper ('Is this sentence about topic X?') involves projective latent content and some disagreement is to be expected. When annotating something of which people have an everyday understanding, it can be more helpful to rely on the 'coders' existing schema' instead of defining more and more detailed annotation rules (Potter and Levine-Donnerstein, 1999, 260).

In our study, we combine multiple annotations and self-assessment of the annotators' certainty, as has been done before by Poesio and Artstein (2005). They study the annotation of anaphora in dialogue data, where many references are vague without leading to misunderstandings. (See also Versley (2006) for a detailed analysis of causes of ambiguities in coreference annotation.) The study attempts to capture ambiguities by, on the one hand, letting 18 students annotate the same text and, on the other hand, giving annotators the option of marking more than one antecedent if they perceive the reference as ambiguous. They conclude that it is important to consider cases of 'implicit ambiguity' which only emerge in the disagreements of multiple annotators and the individual annotator is not aware of. Nedoluzhko and Mírovský (2013) report on the annotation of coreference and bridging relations in the Prague Dependency Treebank. They let annotators explicitly mark how certain they were about each annotated item. The analysis shows a correlation between annotator certainty and agreement, but also reveals many cases of disagreement despite high certainty. They agree with Poesio and Artstein (2005) that ambiguity can be captured more fully by using multiple annotators instead of only letting annotators mark it explicitly.

Further studies consider disagreements between annotators as valuable information. Rohde et al. (2016) conduct a crowdsourcing experiment to study which discourse adverbials licence which conjunctions (e. g. a clause with *instead* can start with the conjunctions *but*, *so*, or *and*). 28 annotators were presented with sentences with one of 20 adverbials and a gap for a possible conjunction. The results show that all adverbials have one to three conjunctions that participants considered acceptable, depending on context as well as individual preference. This variability could only be captured by a high number of annotators per item (similar: Scholman and Demberg, 2017). Morris and Hirst (2004) investigate subjectivity in text interpretation. They let five annotators identify semantically related word groups in a text and specify the semantic relations between the words. While the annotators agree on some core words, individual differences are large. Morris (2010) extends this method and concludes that '40% of the lexical cohesion perceived in text is subjectively interpreted' (Morris, 2010, 141) and therefore argues for a more reader-oriented modeling of text in computational linguistics.

Annotation with multiple annotators has been explored in literary studies, as polyvalence is an important textual characteristic for the definition of literariness. Gius and Jacke (2017) pose the question of how the falsifiability of interpretations can be guaranteed. Their proposal is to use computer-aided narratological annotation to document interpretative decisions, which in turn can make conflicting interpretations visible. While some textual ambiguities do not have consequences for the overall interpretation, some ambiguities result in more than one possible interpretation of a literary text. Hammond et al. (2013) target ambiguity in *To the Lighthouse* by Virginia Woolf. The novel makes extensive use of free indirect speech that cannot be attributed unambiguously to one character or the narrator. To capture possible attributions, they let three to four student annotators analyze the same text span. They reach a raw agreement of slightly less than 70%, however, for many text spans more than one analysis is valid.

Multiple annotations have also been exploited for machine learning. Plank et al. (2014) use the information of annotator agreement to improve POS-tagging. In training their classification model, they sanction mistakes on tags with high inter-annotator agreement more heavily than on tags with low agreement. Their experiments result in annotation improvements in several evaluation settings. Reidsma and op den Akker (2008) optimize their classifiers for high precision by allowing the classifier to make no decision on low-agreement parts of the data. Pavlick and Kwiatkowski (2019) work on entailment, i. e. the question whether the proposition of a sentence *B* can be inferred from the proposition of a sentence *A*. They ask 50 annotators for their judgment on several hundred sentence pairs and show that the disagreement in the annotations cannot be attributed to noise only, but indicates that different interpretations are possible for many sentence pairs. They argue that computational models for textual entailment should therefore produce a full distribution of possible human answers instead of just one aggregated score.

In a previous study of our own project, Adelmann et al. (2019) annotated illness and compared an approach based on semantic fields with manual annotations. The agreement between the two annotators was rather low. The attempt to rectify this problem by improving the annotation guidelines led to the inclusion of very detailed rules that for the most part could not be justified by the demands of the research questions. In truth, illness is simply not a discrete concept, but can be present in a sentence to varying degrees. For this reason, we decided to approach the annotation of illness in a different way. We conducted a crowdsourcing study on the decision task whether a sentence is about illness or not. Similar to Poesio and Artstein (2005) and Nedoluzhko and Mírovský (2013), we combine multiple annotators with a self-assessment of annotator certainty. Closer to Pavlick and Kwiatkowski (2019), we harvested a statistically relevant number of 30 judgements per sentence.

3 Annotation Experiment[1]

Data. The sentences that were presented to the participants were extracted from two corpora that were compiled in the digital humanities project *hermA* (Gaidys et al., 2017):

- Fiction Corpus: 40 novels from the dystopian genre (2000–2019), 135,000 sentences
- Transcript Corpus: written versions of speeches from the German federal parliament ('Bundestag'), filtered for texts that cover an aspect of the topic of telemedicine, 990,000 sentences

We chose these two corpora because we wanted to cover different text types that we assume to differ with respect to ambiguity: Literary texts are said to be especially ambiguous and authors play with ambiguity to create artistic value. Political speeches aim more at a common understanding and should be as clear as possible, but can also be deliberately ambiguous. Both corpora were split into sentences.

Sentence Selection. Sentences were sampled randomly. To ensure that the topic illness was present in a relatively large proportion of the sample, we filtered for sentences that contain a lexical item from the 'semantic field' (Lehrer, 1974; Vassilyev, 1974) of illness, which we realized as hyponyms of the term *illness* in GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010).[2] For each corpus, we included 380 sentences with a lexical item related to illness and 100 without such a word, resulting in 960 items in total.

[1] We published the annotations as a Zenodo dataset. See `https://doi.org/10.5281/zenodo.4088446`.

[2] We extended the original list of 586 hyponyms with inflectional variants using the SMOR (Schmid et al., 2004) derivate Zmorge (Sennrich and Kunz, 2014), resulting in a list of 2,026 wordforms.

Question Design. The selected sentences were presented to the annotators with a context of five sentences before and after the target sentence. Less context was included if the text began or ended in this span. The target sentence was printed in bold. The annotators were presented with two or. three questions: The first one asked for a binary judgment in response to the question: Is the topic of illness discussed in the sentence printed in bold? ('Wird im fett gedruckten Satz das Thema Krankheit thematisiert?'). Given the condition that the answer to this question was 'yes', there was a follow-up question about topic centrality: Annotators were asked whether illness is the central or rather a marginal topic of the sentence. The final question asked for the annotators' certainty: How certain are you about the answer to question 1? ('Wie sicher bist Du Dir bei der Antwort zu Frage 1?'). Possible answers were very certain ('sehr sicher'), rather certain ('eher sicher'), rather uncertain ('eher unsicher'), and very uncertain ('sehr unsicher').

Guidelines. As already mentioned above, the goal of this study was not to review strictly formalized annotation guidelines nor the creation of a consensual gold standard. Since we want to depict ambiguities through annotations, we decided to use extremely minimalist guidelines. In the task description, we even pointed out that we are interested in the subjective assessments of the annotators. However, we gave three example sentences: In the first, illness is the central topic of the sentence (a), in the second, illness is a marginal topic (b), and in the third, illness is not discussed (c).

(a) *Frank liegt schon seit einer Woche mit Fieber im Bett.*
'Frank has been in bed with a fever for a week.'

(b) *Ich freue mich sehr darauf, [. . .] meine Cousine zu treffen, die lange erkältet war.*
'I am very much looking forward to meeting my cousin [. . .], who has had a cold for a long time-.'

(c) *Zum Frühstück esse ich am liebsten Müsli.*
'For breakfast I prefer to eat cereal.'

The question whether a sentence deals with illness is an individual, conceptual decision. By using minimalist guidelines, we hope to cover disagreements caused by conceptual differences as to what the annotators consider to be illness as well as disagreements caused by grammar or style.

Annotation Procedure. We collected 30 judgments for each of the 960 items. For the annotation procedure, we used the crowdsourcing platform *Appen* The crowdworkers were paid $0.70 for each annotated page with ten sentences each. Based on a pretest, we assumed that this would result in more than a minimum wage of $10 for annotators working at average speed. During the annotation process it became apparent that the crowdsourcing platform could not supply us with a sufficient number of German speaking annotators. For this reason we additionally asked the students at our university to participate in the study. The students were paid €0.80 per page.

For purposes of quality control, a number of test questions were used to exclude annotators who either did not understand the task or did not intend to work seriously on the task. As test questions we selected sentences that we considered unambiguous with respect to the question whether they are about illness. However, some of the test sentences turned out to be more ambiguous than we expected. If the annotators gave arguments for their deviant answer, and thus showed that they were actively engaging with the task, we accepted their answers.

In total, 77 annotators participated in our study, 34 crowdworkers and 43 students. The extent of their work varies widely, ranging between 9 and 828 items. On average, participants annotated 374 items, with a huge standard deviation of 305 items. (See Figure 1 for the full distribution.) About half of our data set was annotated by crowdworkers, the other half by students. In order to identify a possible effect of annotator types, we compared the proportion of votes for illness per annotator from the two groups. There is a significant difference in the annotations of the sentences from the Fiction Corpus with semantic field words ($n = 380$, Mann-Whitney U test, $U = 259.0$, p< 0.001, students mean 0.64 ± 0.14, crowdworker mean 0.57 ± 0.07, rank-biserial correlation: -0.55). One possible reason for this difference is the fact that most of the students are studying language and literature and thus might approach the task differently than the crowdworkers. We do not consider this effect problematic and do not account for it in the analysis. There is no significant difference in the proportion of 'very sure' votes to the question on the annotators' certainty.

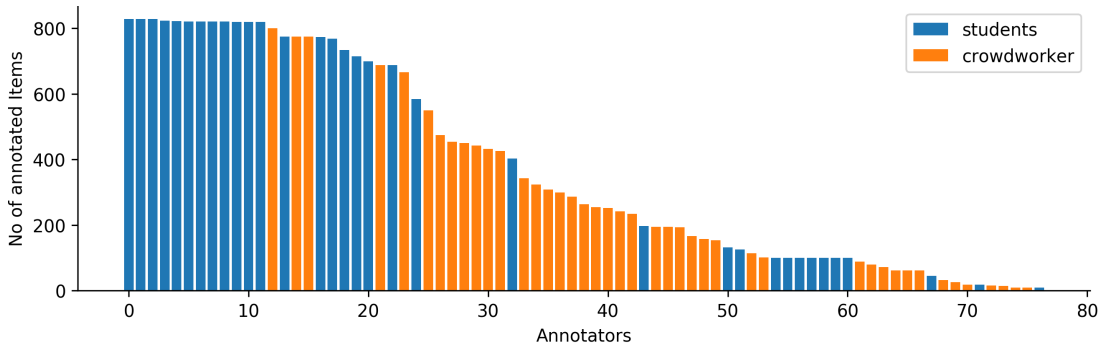
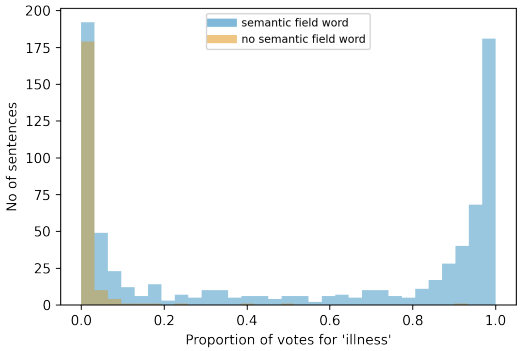
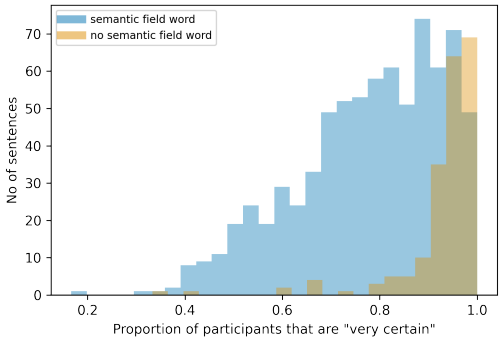


Figure 1: Our 77 annotators, sorted by the number of annotated items (crowdworkers in orange, students in blue)



(a) Distribution of sentences by proportion of annotators that voted for 'illness' (n=980)



(b) Distribution of sentences by proportion of annotators that indicated to be 'very certain' (n=980)

Figure 2: Response distributions (blue bars: semantic field word, orange bars: not a semantic field word).

4 Results

Distribution of Annotation Categories. Figure 2a shows the distribution of answers to the first question: 'Is this sentence about illness?'. A value of 0 on the x axis means that all 30 annotators said that this sentence is not about illness, a value of 1 means that all 30 annotators said that the sentence is about illness. We can see that for most of the sentences without a semantic field word (orange bars), all annotators agreed that this sentence is not about illness. Only 21 of 200 sentences got at least some votes for illness. On the other hand, responses for sentences with a semantic field word are distributed much more widely (blue bars). About one quarter of the sentences is unanimously considered to be about illness (181), another quarter is unanimously considered not to be about illness (192). About half of all sentences with a semantic field word caused some degree of disagreement between the annotators. We conclude that the absence of a semantic field word is a good indicator that the sentence is not about illness. However, the presence of a semantic field word still leaves us with about a 50:50 chance for the sentence to be about illness or not. This decision appears to be non-trivial for human annotators.

The second question ('How certain are you about your answer to question 1?') reveals that the annotators were, overall, rather confident about their answers: When looking at all 28,800 answers to this question, 81.9% of the time the annotators said they were 'very certain' about their assessment of the sentence, in 16.5% they were 'rather certain'. Only in very rare cases did the annotators indicate that they were 'rather uncertain' (1.5%) or even 'very uncertain' (0.2%). Aggregated to sentences this results in the distribution in Figure 2b. Despite the high number of 'very certain' votes overall, only 49 of 760 sentences with a semantic field word and 69 of 200 sentences without a semantic field word get 'very certain' votes only. This is because the annotators had very individual certainty profiles: The proportion of 'very sure' votes per annotator ranges between 0 (one annotator) and 1 (eight annotators), with a mean

of 0.78 ± 0.20. In 362 cases an annotator (55 different annotators) asserted to be very sure about the topic annotation, but annotated against the majority vote. 17 different annotators declared at least once to be very sure while all other annotators were of the opposite opinion.

Agreement. For the calculation of inter-annotator agreement, we use the coefficient Krippendorff's α (Krippendorff, 1980; Krippendorff, 2013)[3]. This coefficient does not require all items to be annotated by the same annotators (see Artstein and Poesio, 2008, for an overview). Following the recommendations of Artstein (2017, 304), we additionally calculate the agreement scores for our two subcorpora and two conditions (semantic field word vs. non semantic field word) individually.

Regarding the first question ('Is this sentence about illness?'), the agreement for the full data set is 0.690. This value indicates substantial agreement (Landis and Koch, 1977). As Figure 2a suggests, the agreement varies depending on the presence of a semantic field word: The agreement on all sentences without a semantic field word is very high with 0.896. Sentences with a semantic field word achieve a much lower agreement of 0.658. This can be explained by the fact that most sentences without a semantic field word are totally unrelated to illness, making the question a trivial one. There is also a moderate difference in agreement between the two corpora: The agreement for the sentences of the Transcript Corpus is 0.756 and the agreement for the sentences of the Fiction Corpus is 0.637.

We did not calculate the agreement for the second question ('How certain are you about your answer to question 1?'), because in answering this question, the annotators do not judge the text shared by all annotators, but their individual annotation experience.

Overall, the agreement scores show a high annotation consistency, given the fact that the annotators did not get further instructions by annotation guidelines. There is a core idea of illness that is shared by most annotators. At the same time we see a considerable number of sentences where the annotators disagree. These sentences can be said to cover the peripheral understanding of illness that is only shared by subgroups of annotators. In the following section we will look more closely at how this relates to annotator certainty and possible causes of disagreement in the sentences themselves.

Topic Centrality. Annotators who said the sentence was about illness had to additionally specify whether illness is the central topic or only a marginal topic of the sentence. For this question, our annotators reach an agreement of 0.246, which is hardly above chance level. For this reason we will not analyze the data for this question in detail. The answers are correlated with the second question about annotator certainty: If the annotators considered illness the central topic, 83% were 'very certain' about their answer. If illness was only a marginal topic, only 58% were 'very certain'. We therefore assume that the (lacking) centrality of the topic is one of many possible reasons for uncertainty. The comprehensive assessment of causes of uncertainty would require a more complex question design.

Disagreement and Uncertainty. Our hypothesis was that if many annotators indicate a high uncertainty, the agreement for these items would be low. For the measurement of agreement per item we chose the highest proportion of participants that agree on one answer. As we have two categories, this is a value between 0.5 and 1 with 0.5 indicating that half of our annotators chose one answer and the other half the other answer, 1 indicating that all annotators agree (on either category).[4] Figure 3 shows the relationship between the agreement proportion on question 1 ('Is this sentence about illness?') on the y axis and the proportion of participants that were very certain about their answer on the x axis. Every data point is one possible value combination and the size corresponds to the number of sentences that match this value combination (between 1 and 129). We can see a clear correlation that is confirmed by a Pearson's correlation coefficient of 0.68 ($p < 0.001$):[5] If many participants were unsure about their answer, there is also much variation in their answers.

The correlation indicates that a considerable amount of variation can be captured by self-assessment of the annotators. However, there still remains a substantial amount of variation that is only captured

[3]As implemented in the R package 'irr' (`https://cran.r-project.org/web/packages/irr/irr.pdf`).

[4]This is similarly captured by the very common measure of entropy, however, we consider a linear measure more appropriate for our interpretation.

[5]This effect is robust even if all items with an agreement of 1 are excluded ($r = 0.57$).

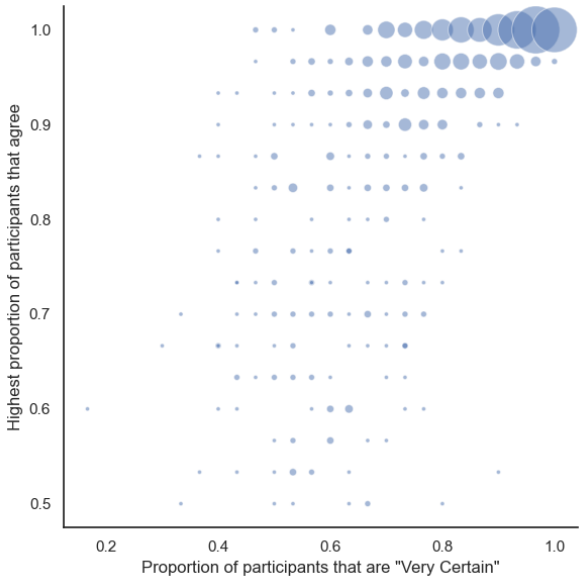


Figure 3: Relationship between the proportion of participants that agree on an answer and the proportion of participants indicating to be 'very certain' (n=960)

by the combination of multiple annotators. We also have to keep in mind that, while the proportion of annotators that are very certain correlates with the answers of the group, this must not be true for the individual annotators. If the annotation targets a phenomenon where ambiguity is expected and the research question makes it desirable to capture it, multiple annotators are highly beneficial.

Reasons for Disagreement. To explore the reasons for disagreement, we evaluate the semantic field words and the sentences on which the annotators disagreed most. We calculate the average agreement of all sentences in which the semantic field words occur, see Table 1. Among the semantic field words that occur in sentences with low average agreement is a group of words which refer to psychological states: *Paranoia* ('paranoia'), *Wahn* ('madness'), *Traumata* ('trauma'), *Sucht* ('addiction'), *Schock* ('shock') and *Anfall* ('seizure'). We assume that the low agreement values in sentences with these terms are caused by different opinions about whether these states have the status of a disease. Additionally, some of the terms describe only short-lived states that therefore have a debatable status: *Anfall* ('seizure'), *Schock* ('shock') and *Husten* ('cough'). We additionally inspected the ten words with the highest average agreement. For five of these words the annotators (almost) agreed that the sentences deal with illness. These words are specific names of diseases: *Tuberkulose* ('tuberculosis'), *Diabetes* ('diabetes'), *Malaria* ('malaria'), *Krebs* ('cancer'), *Leukämie* ('leukemia'). For the other five terms, the annotators agreed that they do not address any disease: *Flechten* ('lichens'), *Attacke(n)* ('attack(s)'), *Verdrängung* ('repression'), and *Abhängigkeiten* ('dependencies') are highly ambiguous, because they can describe a pathological state, but also have a completely separate semantic dimension.

In order to determine what causes disagreement above the lexical level, we manually inspected the 50 sentences with the lowest agreement scores and 50 random sentences with complete agreement. 38 of the sentences with low agreement belong to the Fiction Corpus. Some disagreements can be explained by the lexical phenomena described before: mental phenomena and short-lived states. In addition, some *grammatical* and *stylistic* phenomena prove to be important. One example are negations, which are either explicitly marked by a negator as in example (1), but can also be realized in a syntactically more complex way, as example (2) shows:

(1) *Kein Husten, kein Lebenszeichen.*
'No cough, no sign of life.'

(2) *Das heißt jedoch auch, dass beispielsweise eine Frau, die vor vierunddreißig Jahren geboren wurde, keine persönlichen Erinnerungen an körperliches Leiden besitzt.*

semantic field word	translation	frequency	agreement mean	proportion pos. annotations
Flechten	lichens/eczemas	5	1.00	0.00
Tuberkulose	tuberculosis	5	1.00	1.00
Attacke	attack	9	1.00	0.00
Diabetes	diabetes	8	1.00	1.00
Attacken	attacks	11	0.99	0.10
Malaria	malaria	5	0.99	0.99
Verdrängung	displacement/repression	11	0.99	0.01
Krebs	cancer	10	0.99	0.99
Abhängigkeiten	addictions/dependencies	6	0.99	0.01
Leukämie	leukemia	8	0.99	0.99
...	...	...	...	...
Pilz	mushroom/fungus	5	0.91	0.09
Anfall	seizure/fit	11	0.86	0.45
Leiden	suffering	8	0.85	0.60
Schock	shock	29	0.80	0.24
Pickel	pimples	6	0.75	0.25
Husten	cough	15	0.73	0.68
Sucht	addiction	5	0.73	0.48
Traumata	traumas	9	0.73	0.71
Wahn	delusion/madness	6	0.72	0.34
Paranoia	paranoia	7	0.70	0.60

Table 1: Semantic field words with the lowest and highest agreement scores

'However, this also means that, for instance, a woman born forty-three years ago has no personal memories of physical suffering.'

Among the stylistic phenomena, metaphorical uses of disease symptoms in a non-medical context are the most frequent:

(3) *Diese Schizophrenie findet sich auch in der Öffentlichkeit.*
'This schizophrenia is also found in public.'

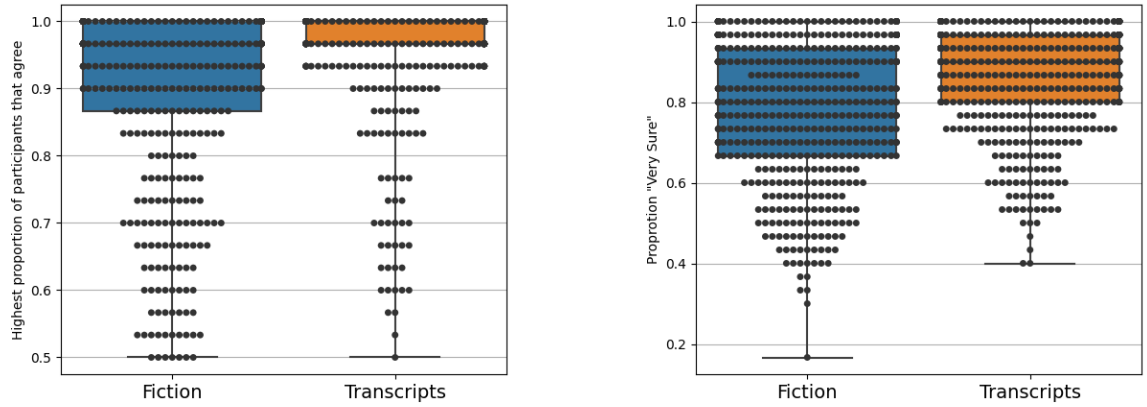
Finally, there are cases where the narrative representation of events may have led to low agreement values. In some sentences, the narrative instance makes speculative statements (example (4)) or presents the perspective of a narrated character (example (5)):

(4) *Das Ergebnis mochte dann am Ende ein Wahn sein, wie er Branagorn befallen hatte.*
'In the end, the result might have been a delusion, as it had infested Branagorn.'

(5) *Man erklärte es sich dann teils mit dem Schockzustand des Kindes und teils mit der erst heraufziehenden Dämmerung [...].*
'This was explained partly by the child's state of shock and partly by the approaching dawn [...].'

The examples also show that several features which potentially cause disagreement among annotators can co-occur in a single sentence. Thus, in example (5) the behaviour of a character is attributed to a temporary state of shock and at the same time the narrator distances himself from this explanation.

Influence of Text Type. We have seen above that the inter-annotator agreement for the sentences from the Transcript Corpus (0.756) was higher than for those from the Fiction Corpus (0.637). In Figure 4 we compare the two corpora from two additional perspectives: Figure 4a shows one box plot per corpus for the agreement per sentence, measured as the highest proportion of participants that agree. Visual inspection gives an indication that the annotators disagreed more in the literary texts than in the debate transcripts on the question of whether illness was addressed in the sentences. The mean agreement is 0.91 (± 0.13) for the Fiction Corpus sentences and 0.96 (± 0.09) for the Transcript Corpus sentences. The Mann-Whitney rank test confirms that this is a significant difference ($U = 89024.0$, $p < 0.001$). With a rank-biserial correlation of 0.23, the effect size is rather small. As Figure 4b shows, there are also differences between the corpora in how confident the annotators are. The mean proportion of 'very



(a) Agreement per sentence

(b) Certainty per sentence

Figure 4: Distribution of agreement and annotation certainty for sentences from the Fiction Corpus (n=480) and the Transcript Corpus (n=480) in contrast

certain' annotators is $0.78(\pm 0.17)$ for the Fiction Corpus sentences and 0.86 (± 0.13) for the Transcript Corpus sentences. This is a statistically significant difference ($U = 83232.5$, $p > 0.001$) and the effect size is slightly larger (rank-biserial correlation: 0.28).

The differences between the corpora can be explained by the vocabulary associated with the two text types. While the sentences for both corpora were selected by the same semantic field, the corpora largely cover different parts of the semantic field: Of 183 semantic field word types in our sentences, only 50 occur in sentences of both corpora. Table 2 presents the most common semantic field words for the two corpora. Only word forms of the general term *Krankheit* ('illness') are frequent in both corpora. Among the most frequent semantic field words in the transcripts are many abstract terms that have very general meanings: *Missbrauch* ('abuse/misuse'), *Abhängigkeit* ('addiction/dependence'), *Vermeidung* ('avoidance'), *Komplex* ('complex'). These are part of the more technical language that characterizes political discussions compared to literary texts. These words often occur in contexts that are not related to illness. This is also reflected in the overall annotation patterns: In the Transcript Corpus data, 59% of all sentences with full agreement were annotated as not being about illness while this is only true for 40% of sentences from the Fiction Corpus. Beyond that, the top ten include two very specific disease terms, *Aids* ('aids') and *Krebs* ('cancer') that will hardly cause disagreement.

The sentences from the Fiction Corpus, on the other hand, are more concrete. Novels tell the story of a character or a small group of characters, depicting the inner life of these characters. To this end, these texts tend to describe mental states that cannot be clearly classified as symptoms of illness. Furthermore, novels also include words that refer to (mostly) minor symptoms like *Husten* ('cough'), whose status as an illness is debatable and which would usually not be discussed in parliament.

5 Discussion and Future Work

In our study, we tested two ways of capturing and modeling ambiguity in texts: By asking many annotators for their judgment and by asking the annotators for a meta annotation about their annotation certainty. We found that low annotator certainty and high variation in judgments are highly correlated. However, many data points that individual annotators were certain about did display variation. In addition, the correlation need not be given for every annotator or even any annotator individually. We conclude that multiple annotations are a useful means to identify and document ambiguity in texts.

Disagreement between our annotators was caused by 1) different concepts of what qualifies as illness (mental phenomena and very short-lived states are controversial), 2) grammatical phenomena like negation, and 3) stylistic properties of the text such as metaphors. Especially in the Fiction Corpus, the information conveyed by the texts can also be ambiguous because the narrative is imprecise, vague or

Fiction			Transcripts		
word	translation	frequency	word	translation	frequency
Krankheit	disease	34	Missbrauch	abuse/misuse	61
Schock	shock	28	Krankheiten	diseases	44
Krankheiten	diseases	18	Abhängigkeit	addiction/dependence	28
Sommersprossen	freckles	15	Krankheit	disease	26
Husten	cough	14	Vermeidung	avoidance	22
Pilze	fungus/mushrooms	13	Komplex	complex	14
Anfall	seizure	9	Attacken	attacks	9
Tief	low/depression	9	Verdrängung	repression/displacement	9
Seuche	plague	8	Aids	aids	7
Attacke	attack	7	Krebs	cancer	7

Table 2: Most common semantic field terms per corpus (ambiguous terms are marked in the translation)

character-driven, as is typical of literary narratives. In order to further differentiate the causes of disagreement and annotator uncertainty, a more comprehensive study would be necessary that gives the annotators more space to give reasons for why they think a sentence is (not) about illness.

Some of the causes of disagreement could easily be avoided by annotation guidelines. We decided against the use of guidelines because the aim of our study was to explore the whole range of possible views on illness in our data. This range could be used for inductively specifying categories for specific guidelines. In addition, if the research objective allowed for a clear position on whether mental phenomena are supposed to be annotated as illness or whether negated mentions of illness are supposed to be annotated, guidelines clarifying these points are highly recommended. However, beyond the definitions we can derive inductively or from a research question there will most likely be space for individual interpretation. We would like to encourage researchers to regard this variation not as a problem to be fixed but something that can be incorporated into our modeling of the world and in our analyses.

The semantic field as a tool to search for specific topics is ambivalent: While the absence of a semantic field word was a good indicator that the sentence is not about illness, the presence of a semantic field word only resulted in a 50% chance of the sentence being about illness. Based on our annotations we can derive a weighted semantic field that can, for instance, be filtered to get a core semantic field. This allows for a reduction of hits to only those sentences that most people consider to be about illness. However, this would systematically exclude phenomena from the data set, as especially psychological phenomena led to disagreements.

With respect to text types, our study revealed more ambiguous instances in the Fiction Corpus than in the Transcript Corpus. While one might consider metaphors to be the cause of ambiguity in literary text, this is not what the inspection of sentences with low agreement indicates. Instead, the transcripts are characterized by many abstract terms like *Abhängigkeiten* ('dependencies, addictions') which are mostly used in a way that is clearly unrelated to illness and do not cause any disagreement. The Fiction Corpus, on the other hand, names many minor symptoms (*Husten*, 'cough') as everyday situations are described and can also report the inner life of characters.

In the future, we plan to explore possibilities for training machine learning models on the data presented here. By showing how ambiguity levels can be represented by multiple annotations, we hope to prepare for the creation of a complex gold standard that incorporates conflicting evidence (Reidsma and op den Akker, 2008; Passonneau and Carpenter, 2014).

Acknowledgements

The work on this paper was funded by the *Landesforschungsförderung Hamburg* (LFF-FV 35) in the context of the project *hermA* (Gaidys et al., 2017) at Universität Hamburg and Hamburg University of Technology. We thank Piklu Gupta and Carla Sökefeld for proofreading. All remaining errors are our own.

References

Benedikt Adelmann, Melanie Andresen, Anke Begerow, Lina Franken, Evelyn Gius, and Michael Vauth. 2019. Evaluation of a Semantic Field-Based Approach to Identifying Text Sections about Specific Topics. In *DH 2019. Book of Abstracts*.

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3318–3329, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, September.

Ron Artstein. 2017. Inter-annotator agreement. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 297–313. Springer.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. 2008. Analyzing Disagreements. In *Coling 2008: Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 2–7, Manchester, UK, August. Coling 2008 Organizing Committee.

Uta Gaidys, Evelyn Gius, Margarete Jarchow, Gertraud Koch, Wolfgang Menzel, Dominik Orth, and Heike Zinsmeister. 2017. hermA: Automated modelling of hermeneutic processes. *Hamburger Journal für Kulturanthropologie*, (7):119–123.

Evelyn Gius and Janina Jacke. 2017. The Hermeneutic Profit of Annotation. On preventing and fostering disagreement in literary text analysis. *International Journal of Humanities and Arts Computing*, 11(2):233–254.

Adam Hammond, Julian Brooke, and Graeme Hirst. 2013. A Tale of Two Cultures: Bringing Literary Analysis and Computational Linguistics Together. In *Proceedings of the Workshop on Computational Linguistics for Literature*, pages 1–8. Association for Computational Linguistics.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet – a Lexical-Semantic Net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT – The GermaNet Editing Tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pages 2228–2235, Valletta, Malta.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Number 5 in The Sage Commtext Series. Sage, Beverly Hills, California.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. Sage, Los Angeles, third edition.

George Lakoff. 1987. *Women, Fire, and Dangerous Things. What Categories Reveal about the Mind.* The University of Chicago Press, Chicago, London.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Adrienne Lehrer. 1974. *Semantic fields and lexical structure*. North-Holland Publishing Company, Amsterdam.

Jane Morris and Graeme Hirst. 2004. The Subjectivity of Lexical Cohesion in Text. *AAAI Spring Symposium - Technical Report*, 20.

Jane Morris. 2010. Individual differences in the interpretation of text: Implications for information science. *Journal of the American Society for Information Science and Technology*, 61(1):141–149.

Anna Nedoluzhko and Jiří Mírovský. 2013. Annotators' Certainty and Disagreements in Coreference and Bridging Annotation in Prague Dependency Treebank. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 236–243.

Rebecca J. Passonneau and Bob Carpenter. 2014. The Benefits of a Model of Annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326, December.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751.

Massimo Poesio and Ron Artstein. 2005. The Reliability of Anaphoric Annotation, Reconsidered: Taking Ambiguity into Account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan, June.

W. James Potter and Deborah Levine-Donnerstein. 1999. Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, 27(3):258–284, August.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 8–16, Manchester, UK, August. Coling 2008 Organizing Committee.

Hannah Rohde, Anna Dickinson, Nathan Schneider, Christopher N. L. Clark, Annie Louis, and Bonnie Webber. 2016. Filling in the Blanks in Understanding Discourse Adverbials: Consistency, Conflict, and Context-Dependence in a Crowdsourced Elicitation Task. In *Proceedings of the 10th Linguistic Annotation Workshop Held in Conjunction with ACL 2016 (LAW-X 2016)*, pages 49–58, Berlin, Germany, August.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Merel Scholman and Vera Demberg. 2017. Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 24–33, Valencia, Spain, April. Association for Computational Linguistics.

Rico Sennrich and Beat Kunz. 2014. Zmorge: A German morphological lexicon extracted from Wiktionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1063–1067, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Leonid M. Vassilyev. 1974. The theory of semantic fields: a survey. *Linguistics*, 12(137):79–94.

Yannick Versley. 2006. Disagreement Dissected: Vagueness as a Source of Ambiguity in Nominal (Co-)Reference. In *Proceedings of the Ambiguity in Anaphora Workshop (ESSLLI 2006)*, pages 83–89.