An Annotation Protocol for Diachronic Evaluation of Semantic Drift in Disability Sources

Nitisha Jain, Chiara Di Bonaventura, Albert Meroño-Peñuela, Barbara McGillivray
King's College London, 30 Aldwych, London, UK
{nitisha.jain,chiara.di_bonaventura,albert.merono,barbara.mcgillivray}@kcl.ac.uk

Abstract

Annotating terms referring to aspects of disability in historical texts is crucial for understanding how societies in different periods conceptualized and treated disability. Such annotations help modern readers grasp the evolving language, cultural attitudes, and social structures surrounding disability, shedding light on both marginalization and inclusion throughout history. This is important as evolving societal attitudes can influence the perpetuation of harmful language that reinforces stereotypes and discrimination. However, this task presents significant challenges. Terminology often reflects outdated, offensive, or ambiguous concepts that require sensitive interpretation. Meaning of terms may have shifted over time, making it difficult to align historical terms with contemporary understandings of disability. Additionally, contextual nuances and the lack of standardized language in historical records demand careful scholarly judgment to avoid anachronism or misrepresentation. In this paper we introduce an annotation protocol for analysing and describing semantic shifts in the discourse on disabilities in historical texts, reporting on how our protocol's design evolved to address these specific challenges and on issues around annotators' agreement.

1 Introduction

Language constantly evolves and adapts to speakers' communicative needs and socio-cultural changes; understanding these shifts is crucial for grasping the dynamic nature of language and its intricate relationship with social and cultural phenomena. The semantics of words of a language shift due to influences from social practices, events, and political circumstances (Keidar et al., 2022; Castano et al., 2022; Azarbonyad et al., 2017). The *functioning and disability of individuals*,[1] such as those affecting their cognitive, developmental, intellectual, mental, physical or sensory functions, is a key area of study pursuing equitable access in society, and in which language is in constant motion: inappropriate use of language can contribute to the perpetuation of stereotypes, discrimination, and stigmatization (Andrews et al., 2022). For example, the word "lame" was historically associated with physical disabilities affecting a person's ability to walk or move normally; but over time, it has semantically changed to mean "socially inept or out of touch" (Oxford University Press, 2024b), shifting meaning from a physical disability context to a more casual and potentially derogatory usage. Therefore, development of techniques to annotate such semantic change within the disability domain is essential for ensuring accurate interpretation and fostering a deeper understanding of historical texts. Without such methods, there is a risk of misrepresenting or overlooking the evolving meanings and social implications of disability-related terms across different historical contexts.

In Natural Language Processing (NLP), the task of Semantic Shift Detection (SSD) focuses on detecting, interpreting, and assessing potential changes in the meaning of words over time (Montanelli and Periti, 2023). The International Workshops on Semantic Evaluation (SemEval) (Schlechtweg et al., 2020) and Ever Evolving NLP (EvoNLP[2]) have proposed various tasks and models. In the Semantic Web, ontology evolution (Stojanovic, 2004) studies how and why ontologies and knowledge graphs change over time; various works have proposed models based on heuristics (Stavropoulos et al., 2019) and machine learning models for semantic change in biomedicine (Pesquita and Couto, 2012) and generalised domains (Meroño-Peñuela et al., 2021), with some studies looking into the impact of seman-

[1] WHO disability classification standards.

[2] `https://sites.google.com/view/evonlp/home`.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 159–171
July 5–10, 2020. ©2020 Association for Computational Linguistics

tic change on reasoning and hierarchies (Pernisch et al., 2019, 2021). As explained in previous works (McGillivray et al., 2022; Hoeken et al., 2023), changes in language semantics over time can influence what is considered offensive. However, to the best of our knowledge no existing work facilitates resources for semantic change over large time spans (as these changes can be slow), considering both textual and semantic representations, and addressing discriminatory and harmful language in disability.

In this paper, we propose an annotation protocol for the analysis and evaluation of semantic change in the disability domain, which is built on two rounds of iteration. Our approach involves designing an annotation framework to capture both the descriptive and offensive nuances of historically relevant disability-related terms, accounting for their evolving connotations across different historical and social contexts. This includes structured guidelines for annotators to assess the perceived offensiveness, descriptive intent, and type of disability referenced in each instance. We present the quantitative and qualitative analyses on annotation disagreement that highlight the importance of capturing the nuanced and subjective nature of disability-related discourse, and discuss the four main challenges in annotating disability-related discourse over time. The annotation data and guidelines have been made available[3] to promote further research in this direction.

2 Background and Related Work

There are several previous studies directed towards the evolution of disability terminology across various mediums, including media representations, scholarly publications, and broader social discourse (Ferrigon and Tucker; Simon, 2017; Auslander and Gold, 1999). Importantly, these studies show the changing landscape of disability discourse, its impact on societal perceptions and attitudes, and the dynamic nature of language and its role in shaping perceptions of disability within diverse contexts (Andrews et al., 2022).

A number of research projects have addressed the issues of bias and representation in historical texts, developing several resources that focus on the language and portrayal of disability (Rahman, 2024; National Center on Disability and Journalism, 2021; DE-BIAS Project consortium, 2025). These initiatives aim to highlight and mitigate the marginalization of disabled individuals in historical records by providing analytical frameworks and lexical resources that bring attention to the social and cultural contexts in which disability-related terms were used in the past and how they should be used today.

Within the research area of Semantic Shift Detection, benchmark datasets and text corpora capable of supporting the analysis of word meaning change over time have been developed (cf. McGillivray et al. (2023) for an overview and Marongiu et al. (2024) for a discussion of this task in the context of semantic change research). The SemEval 2020 dataset (Schlechtweg et al., 2020) contains a multilingual set of annotated sentences from English, German, Latin, and Swedish historical texts; other gold standard datasets exist (Rodina and Kutuzov, 2020; Zamora-Reina et al., 2022). These datasets were all annotated by human experts, which ensures a high level of accuracy and contextual understanding, particularly important when dealing with nuanced and historically contingent language, but it is also a time-consuming and labor-intensive process. Ridge et al. (2024) present a dataset of historical British newspapers from the 19th century where the contexts of a number of terms related to vehicles were annotated with their meaning via voluntary crowdsourcing, leveraging the scalable, collective effort of non-expert contributors.

While existing annotated datasets from semantic change detection research constitute a promising avenue for studying semantic change and improving the understanding of historical language use, the existing resources solely utilize corpora amassed from general domains. As a result, they often overlook specialized areas such as disability discourse, where terminology carries distinct social and cultural significance that requires focused analysis. On the other hand, previous studies on the language of disabilities have not looked specifically at the challenges of corpus annotation in historical texts. Our study addresses both these gaps by focussing on an annotation protocol specifically tailored to the annotation of disability terms whose semantics has changed in historical texts.

In addition to the semantic change literature, our work also intersects with annotation challenges explored in socially sensitive domains. Similar challenges have been discussed in the hate speech detection literature, where offensiveness and inflammatory intent often vary by context, speaker

[3] https://doi.org/10.6084/m9.figshare.29198132

identity, and target community (Sap et al., 2019; Pavlopoulos et al., 2020). Recent work has introduced graded offensiveness scales, soft-labeling approaches, and community-informed annotation schemes to better reflect the subjective and socially contingent nature of such language (Vidgen et al., 2019; Mostafazadeh Davani et al., 2022). Our annotation protocol draws on these developments by adopting a five-point offensiveness scale and encouraging annotators to consider both historical context and social intent when evaluating terms.

3 Data Sources

For designing the annotation protocol for measuring the semantic change in the disability domain, we selected texts for annotation from Gale's *History of Disabilities: Disabilities in Society, Seventeenth to Twentieth Century*[4], a collection of monographs, manuscripts, and ephemera documenting disability history (17th-20th centuries) through personal memoirs, accounts of care and rehabilitation, advocacy efforts, and policies impacting individuals with disabilities, thus examining society's evolving perceptions of disability. Additionally, we collected an initial list of terms used to refer to disabilities from Wikipedia[5] and the Disability at Stanford project.[6]

4 Annotation Protocol

The purpose of the annotation is to trace the evolution of selected terms related to disabilities over time in historical texts. We conducted two annotation rounds to assess the quality of the sources and refine the annotation protocol. The pilot round was carried out by a team of five annotators working in Digital Humanities and Natural Language Processing and from career levels ranging from doctoral students to senior lecturers. The aim of this pilot was to assess the quality of the source texts for the annotation task at hand. The annotation protocol was built and refined based on the feedback given by participants in the pilot.

In the first version of the protocol, each annotation line displayed a focus sentence with the disability term (one of the selected terms) in bold, along with the sentence before and after it for context. Annotators were tasked to choose from a drop-down menu whether the term was 'Derogatory', 'Not derogatory', 'Not referring to a disability', or 'Unclear due to illegible OCR'—a necessary option given the limitations of historical documents. If the term did refer to a disability, annotators also indicated whether it referred to a 'mental' or 'physical' disability. This distinction was important for understanding how different types of impairments were perceived and treated historically, as societal attitudes and institutional responses often varied between mental and physical disabilities.

Feedback from the pilot annotation round revealed several important insights and challenges that guided the updates to the following round of the protocol. Annotators noted, for example, that *demented* often appeared in medical texts to classify individuals deemed “mentally insane" by historical standards. Though medically framed at the time, the term would now be seen as stigmatizing. Similarly, *Downie* was sometimes used as a personal name rather than a reference to Down syndrome, and in certain cases, it appeared in affectionate or familiar contexts—underscoring the importance of contextual interpretation.

The term *cripple* also prompted discussion among annotators. While it was sometimes used descriptively in medical contexts, it often appeared in passages reflecting harsh or dehumanizing attitudes. *These examples highlighted the limitations of a binary classification (Derogatory vs. Not derogatory), which could not capture the nuance of tone and intent.* Annotators also found the mental vs. physical distinction for disability types too narrow, noting that many instances involved cognitive or sensory disabilities (e.g., blindness, deafness) that fell outside these categories.

Based on this feedback from the pilot, we modified the protocol to better account for the historical and contextual subtleties encountered in the data. Again, each annotation line presents a focus sentence with the disability term highlighted, preceded by the sentence before it and the sentence after. The annotation consists now in choosing from the drop-down menu the best category to which the term can be assigned according to the following dimensions.

The first decision annotators make is to determine whether the term is used as part of a 'formal diagnosis' or within 'common language'. This distinction helps clarify whether the term is functioning within an institutionalized medical discourse or in more casual, everyday speech.

Next, annotators assess whether the term is used

[4] Gale's Disabilities in Society, Seventeenth to Twentieth Century Collection.

[5] Wikipedia list of disabilities with negative connotations.

[6] Disability at Stanford project.

with a ‘descriptive’ or ‘offensive’ intent. To capture varying degrees of offensiveness and contextual appropriateness, we implemented a *graded scale*, allowing annotators to position the term along a five-point scale:

1. *Neutral/Descriptive*: Factually descriptive and still acceptable in contemporary usage.
2. *Outdated but Neutral*: Historically accepted and descriptive, but now considered outdated or replaced by person-first language.
3. *Mildly Pejorative / Stigmatizing*: Sometimes used negatively but not inherently offensive; may reflect stereotypical or patronizing attitudes.
4. *Strongly Pejorative / Insulting*: Clearly used offensively or with dehumanizing intent.
5. *Highly Offensive / Dehumanizing*: Explicitly used as a slur or in oppressive, violent, or cruel contexts.

This graded scale was introduced to replace the earlier binary classification of ‘Derogatory’ vs. ‘Not derogatory’, which proved inadequate in capturing the nuances of language and intent found in historical texts. With a more granular approach we acknowledge that offensiveness exists on a spectrum and is deeply influenced by context, authorial intent, and audience perception—particularly in diachronic corpora.

Further, if the term in context refers to a disability, annotators are asked to mark the ‘Type of Disability’ it pertains to. Annotators can select from *cognitive*, *sensory*, and/or *physical* categories. This refinement allows us to better track how different forms of disability were represented and discussed over time, and how terminology may have shifted in relation to different kinds of impairments.

Finally, in an optional comment field, annotators can explain their decision or provide additional observations. These qualitative notes are crucial for later analysis of annotation disagreements and for understanding the reasoning processes behind individual annotations.

5 Annotation Process

In the pilot annotation round, we examined four terms (henceforth referred to as “keywords”): *abnormal*, *cripple*, *demented*, and *downie*. These were chosen for their historical relevance to disability and their shifting meanings and acceptability over time. The selection balanced terms referring to physical disabilities (*cripple*, *downie*) and cognitive or mental ones (*abnormal*, *demented*) to explore varied linguistic representations.

Abnormal, derived from Latin *abnormis* (“irregular”), was commonly used in 19th- and early 20th-century clinical texts to describe physical or mental deviations from a perceived norm. Though often descriptive, the term has accumulated negative connotations, reinforcing ideas of deviance and stigma.

Cripple once served as a general descriptor for individuals with physical disabilities, especially mobility impairments. While historically common in both medical and everyday language, it is now widely viewed as offensive due to its reductive and dehumanizing implications. Some activists have attempted to reclaim the term in recent years to subvert its derogatory implications (Wikipedia contributors, 2025).

Demented, from Latin *demens* (“out of one’s mind”), was used in medical contexts to describe cognitive and psychiatric impairments. Though originally clinical, it has since acquired derogatory connotations and is often used pejoratively in modern speech.

Downie, a colloquial term sometimes aimed at individuals with Down syndrome, appeared in both derogatory and affectionate contexts. However, its frequent use as a personal surname made annotation difficult due to ambiguity and low inter-annotator agreement.

In the first round of annotation, for each keyword, we selected three textual excerpts from monographs and one from manuscripts through advanced search throughout the *Gale’s History of Disabilities* collection (as described in §3). This approach aimed to capture both institutional and personal uses of the terms while accounting for sources’ distributions.

In the subsequent annotation round, we excluded *downie* from the dataset due to its ambiguity. Most occurrences were personal surnames unrelated to disability, resulting in non-relevant instances and inconsistent annotator agreement. Additionally, the limited context in some documents made it difficult to determine whether the term was used derogatorily or descriptively. As a result, we selected the word *blind* for further analysis. The term *blind* has a long history, originating from Old English meaning “sightless” or “obscured” (Oxford University Press, 2024a). Historically, *blind* was commonly used to describe individuals with significant visual impairments. Although originally a neutral descriptor, modern disability discourse has raised concerns about its use, particularly in metaphorical contexts where it can perpetuate negative stereotypes (e.g., “blind to the truth”). In disability advocacy, there is increasing emphasis on person-first language (e.g., “person who is blind”) or identity-first language (e.g., “blind person”), depending on individual and community preferences.

For this second round, we aimed to curate a larger annotation corpus for a more detailed analysis. For each of the four keywords, we first identified 15 monographs and 10 manuscripts from the collection through advanced keyword search. From these, a list of 40 sentences were randomly selected for each keyword (along with the previous and next sentences for context), resulting in a curated annotation corpus of 120 textual excerpts in total. The annotation workshop comprised 12 annotators from research teams within the authors’ University. One annotator had a background in Linguistics and all others had background in Computer Science. The levels of experience ranged from early career researchers (doctoral students, postdocs) to senior lecturers. During the workshop, participants were first introduced to the annotation protocol and guidelines. Then, they worked in small groups of three to annotate the selected sentences along the dimensions discussed in §4 following a structured approach[7].

6 Analysis of annotations

In this section, we analyse the results of the annotation process described in §5. Specifically, we present a quantitative analysis regarding annotators’ agreement in §6.1. In addition, we present a qualitative analysis discussing the challenges and some of the interesting cases that were observed during the annotation process in §6.2[8].

6.1 Quantitative Analysis

The total size of the annotation corpus in terms of the actual sentences to be annotated, measured as count of words is 6717 (*Abnormal* - 1581, *Blind* - 1359, *Cripple* - 1749, and *Demented* - 2028). Firstly, we show in Figure 1 the distribution of the curated annotation corpus over time[9] in terms of number of texts from each decade with respect to the different keywords. The corpus contains texts from a varied range of time periods, starting from 1860s to 1980s. We notice that there is a peak in the 1910s, primarily driven by the word *cripple*, followed by *abnormal*. After this peak, there is a decline in document mentions during the 1920s and 1930s, with a slight resurgence in the 1950s and 1960s. The word *blind* sees a significant rise in the 1950s, while demented appears more frequently in the 1960s and 1980s. Early decades from the 1860s to 1900s show consistent but lower occurrences of these terms.

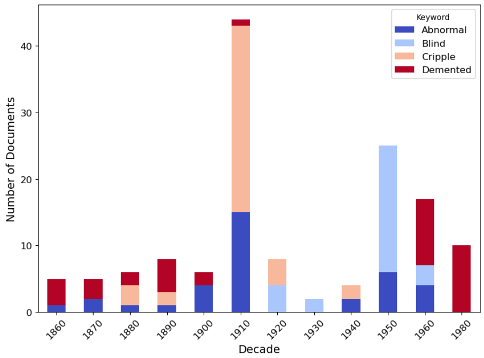

Figure 1: Publication dates of the documents in the annotation corpus (grouped by decades).

Figure 2 presents the distribution of labels obtained from the annotations (cumulative for all annotators) for three different annotation tasks across multiple keywords. The distribution of labels for the first task reveals how medical terms transfer into common discourse, and conversely, how colloquial expressions find their way into formal diagnostic contexts. In our dataset, *cripple* appears to lean more heavily into common language usage, while the other keywords maintain a more balanced representation between diagnostic and everyday speech. In the second task, at the neutral end (level 1), the terms begin with a relatively descriptive, clinical approach. As the labels progress through values 2 and 3, we see the gradual introduction of more pejorative and stigmatizing language. The transition is particularly striking for *cripple* and *demented*, which shows a significant shift towards more negative characterizations. Finally, in the third task we see a substantial agreement among annotators, with *blind* being recognised as predominantly sensory-focused, *demented* as heavily weighted towards cognitive characteristics, and *cripple* with strong physical connotation. *Abnormal* stands out as displaying a more polysemous

[7] the annotations will be made publicly available

[8] Note that these results correspond to the second round of annotations, the pilot was only leveraged to refine the annotation protocol and no agreement measurements were made

[9] wherever this information was explicitly available in the metadata from the collection

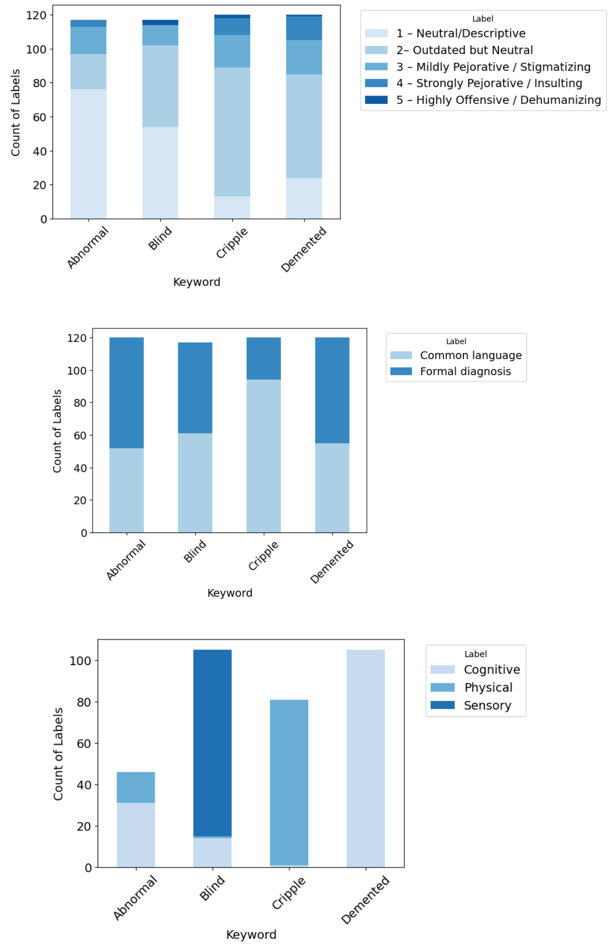

Figure 2: Distribution of annotation labels across different tasks and datasets. The subfigures show the label distributions for three annotation tasks: `Intent of Term`, `Use of Term` and `Type of Disability`.

profile, including both cognitive and physical interpretations[10].

6.1.1 Measuring annotator agreement

To assess the consistency of the annotations and the degree to which annotators agree on the interpretation of the terms, we calculated Cohen's Kappa (Cohen, 1960) and Fleiss' Kappa scores (Joseph and Fleiss, 2023) (Table 1). We also calculated Spearman's rank correlation (Spearman, 1961) to measure the agreement and variance among annotators who classified terms with varying degrees of offensiveness.

Annotation Task	**Keyword**	$\overline{C\kappa}$	$F\kappa$	$\overline{S\rho}$
Intent of Term	Abnormal	0.18	0.17	0.22
	Blind	0.26	0.24	0.30
	Cripple	-0.12	-0.13	0.02
	Demented	0.06	0.02	0.52
Use of Term	Abnormal	0.26	0.25	-
	Blind	0.06	0.04	-
	Cripple	-0.05	-0.08	-
	Demented	0.36	0.36	-
Type of Disability	Abnormal	0.27	0.19	-
	Blind	-0.08	-0.15	-
	Cripple	0.33	-0.01	-
	Demented	1.00	1.00	-

Table 1: Average Cohen's Kappa ($\overline{C\kappa}$) and Fleiss' Kappa ($F\kappa$) for each annotation task and keyword. Averaged Spearman's Rank Correlation ($\overline{S\rho}$) for the `Intent of Term` annotations.

Cohen's Kappa ($C\kappa$). Cohen's Kappa (κ) was used to measure pairwise agreement between annotators, calculated as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o is the observed agreement and P_e is the agreement expected by chance. It is to be noted that in cases with highly skewed label distributions, P_e can be close to or equal to P_o, resulting in *low or even zero Kappa scores* despite frequent agreement between annotators. In extreme cases where both annotators used only a single class, $P_e = 1$, making the denominator zero and rendering Kappa *undefined* (NaN). For reporting purposes, we replaced such values with 1.00 to reflect perfect agreement in these cases. Keeping this in mind, the averaged Cohen's Kappa results in Table 1 reveal varying levels of agreement across annotation tasks and keywords. For the 'Intent of Term' task, the agreement is generally low, with *blind* showing the highest value (0.26), and *cripple* showing a negative Cohen's Kappa value (-0.12) indicating poor or no agreement between raters. In the 'Use of Term' task, the highest agreement is seen with the keyword *demented* (0.36), while the keyword *blind* has a low agreement (0.06). The keyword *cripple* shows a negative value (-0.05). In the 'Type of Disability' task, the agreement is stronger, particularly for *demented* (1.00 indicating complete agreement), suggesting a higher level of consistency in annotating this keyword. On the other hand, other keywords show much lower agreement, with *blind* showing the lowest score (-0.08) as the annotators chose differently among the *cognitive*, *sensory*, and *physical* categories. Overall, these results suggest

[10] This figure illustrates the overall distribution of labels across all annotators, but does not reflect inter-annotator agreement and should not be interpreted as indicative of consistency between annotators. Due to label imbalance and varied interpretation of terms, high label frequency does not necessarily imply high agreement, which is instead captured through chance-corrected metrics like Cohen's or Fleiss' Kappa.

that the annotators show varied levels of agreement when categorizing disability-related keywords[11]. Keywords like *demented* are more clearly interpreted by annotators, leading to higher agreement, whereas *cripple* and *blind* are perceived as more ambiguous or context-dependent, highlighting the challenges in achieving a consistent understanding of these terms, particularly in contexts that might be socially or culturally sensitive.

Fleiss' Kappa ($F\kappa$). Fleiss' Kappa (κ) was used to assess agreement across multiple annotators, using the same chance-corrected formulation:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

where $\bar{P}$ is the *mean observed agreement* and $\bar{P}_e$ the expected agreement by chance. As with Cohen's Kappa, *skewed label distributions* can lead to low or undefined (NaN) scores. We replaced undefined values with 1.00 in cases of unanimous single-class agreement. These scores generally indicates low-to-moderate agreement across the keywords. In the 'Use of Term' task, *demented* stands out with the highest Fleiss' Kappa, suggesting better consensus among annotators, while *cripple* and *blind* show much lower Fleiss' Kappa values, indicating significant disagreement. Notably, *cripple* has a negative Fleiss' Kappa in all tasks, reflecting widespread discord [12].

Spearman's rank correlation ($S\rho$). For the 'Intent of Term', since the annotators rate terms across categories from neutral/descriptive to highly offensive, Spearman's correlation provides insight into how consistently these annotators align in their evaluations. The average correlation scores highlight differences in annotator agreement across keywords. *Demented* has the highest overall agreement (0.52), suggesting that annotators had a more consistent understanding of how to classify this term. *Blind* (0.30) and *abnormal* (0.23) show moderate agreement. In contrast, *cripple* has the lowest agreement (0.02), indicating substantial variation in interpretation, possibly due to its historical connotations and evolving societal perceptions. This suggests that certain terms may be more prone to subjective interpretation, impacting annotation reliability[13].

6.2 Qualitative Analysis

This section presents a qualitative analysis of annotator disagreements during dataset annotation, with a selection of particularly insightful examples, which reflect the subjective nature of interpreting complex socio-linguistic constructs, especially in ethically and historically sensitive domains like disability-related language. Following the framework proposed by Röttger et al. (Röttger et al., 2022), who distinguish between descriptive and prescriptive annotation paradigms for subjective NLP tasks, we adopted the descriptive paradigm in our annotation process. This approach encourages annotator subjectivity, allowing us to capture a range of valid interpretations rather than enforcing a single normative viewpoint. Specifically, we discuss the unique challenges in time-sensitive annotations, that we group into four categories: (1) subjectivity in the interpretation, (2) contextual influence on the annotation, (3) Historical and linguistic evolution, (4) Categorisation challenges[14].

6.2.1 Subjectivity in the interpretation

Offensiveness vs. Stigmatization. The assessment of offensive language varied significantly across annotators. Although disability-related terms were not explicitly offensive in isolation, the surrounding context often conveyed stigmatizing messages. Annotators frequently highlighted portrayals of disability that reinforced harmful stereotypes—for example, associating blindness with poverty, abnormality with criminality, or framing disabled individuals as obstacles to social and economic progress. Such implicit negativity influenced how terms were judged, leading to disagreement about their offensiveness. For example, in the sentence *"The so-called 'cripples' were confined to a separate wing of the institution"*, one annotator viewed the term 'cripples' as mildly pejorative due to its stigmatizing undertones, while another interpreted it as neutral, reflecting historical norms. A third annotator took an intermediate position, recognizing the term's outdated but non-hostile nature. These differences underscore the subjective nature of assessing offensive language, particularly in historical texts where social norms have evolved.

[11] pairwise Kappa scores are presented in the Appendix (Table 2)

[12] A visual representation of the Fleiss' Kappa scores and their variation across different terms is presented in the Appendix (Figure 3)

[13] detailed analysis and visualization in the Appendix

[14] Further discussion and examples in Appendix B

Value of Qualitative Comments. The notes provided by annotators offered valuable insight into their reasoning and highlighted the complexity of the task. For instance, one annotator remarked that while 'abnormal' could be interpreted as informal, the historical context suggested it carried diagnostic weight. Another commented that the term 'cripple' felt stigmatizing but did not appear intended to insult. Such reflections underscore the importance of qualitative comments in resolving ambiguity and improving consistency in annotation.

6.2.2 Contextual influences on the annotation

Focus sentence vs. Whole context. In some cases, annotators reported that the ratings of intent of use would have been different based on whether they should have considered just the focus sentence or the whole context. Indeed, annotators found instances in which the use of a word was mildly offensive or not offensive at all, but their context was very offensive or contained other offensive words. For example, one original sentence concerning 'demented' said that *"dementia concerned mental retrogression"*, but the immediate context after discussed *"the intelligence of idiots* and that *idiocy in all its degrees means arrested or retarded development"*. Such discrepancies contributed to annotator disagreement, as some focused on the standalone sentence while others considered the full passage. This variability reveals the limitations of narrow-span annotation when assessing offensive language, especially in historical texts where offensive intent or stigma may accumulate across sentences. It also underscores the importance of supporting larger-span annotations to better capture temporally sensitive shifts in language use and meaning.

Unique Challenges in Semi-Structured Content. The annotators felt that the task of annotating uses of the potentially offensive words in titles, references, and citations was fundamentally different from working on free text, mostly due to the limited context.

6.2.3 Historical and linguistics evolution

Influence of Historical Context on Meaning. The historical context of language significantly influenced annotators' decisions. Terms like 'abnormal' and 'cripple' have undergone shifts in meaning over time, from clinical or neutral descriptors to terms with potential stigmatizing connotations. Annotators' varied responses reflect the difficulty of balancing the original historical context with modern understandings of disability language.

Semantic Change and the Origin of Slurs. Prompted by the cross-analysis of their annotations, the annotators openly discussed about the origin of slurs and how offensive language comes into existence in the first place. One annotator said that slurs have "only appeared recently" and that "it made no sense to have them back then, it is a newer phenomenon". The discussion focused on the fact that there are probably no "intentional" slurs in the dataset (because of the medical domain, and because of the time at which the text of the dataset was published), hypothesising that it is the post-hoc use of medical terms in discourse what prompts their semantic drift into offensive language.

6.2.4 Categorisation challenges

Formal Diagnosis vs. Common Language. Annotators faced challenges in classifying disability-related terms, particularly when distinguishing between formal medical diagnoses and common or colloquial usage. For instance, the sentence *"The child was described as abnormal in both behavior and appearance, requiring constant supervision"* was interpreted differently. While one annotator classified it as common language, reflecting everyday usage, others marked it as a formal diagnosis. This highlights the challenge of distinguishing between colloquial and medical language, especially when historical shifts in meaning blur the boundaries. For future time-sensitive annotations in disability sources we suggest practitioners to expand these two categories including, for instance, 'medical use but not formal diagnosis'.

Difficulties in Identifying Implied Disabilities. In some cases, annotators differed in marking implied disability types. For example, the sentence "*The blind man had remarkable memory and navigated the town with ease*" was identified as referring to sensory disability by two annotators, while another overlooked the implication. This suggests that implicit references to disability, especially when not explicitly stated, pose challenges for consistent annotation and require greater sensitivity to context.

Multiple Dimensions of Medical Conditions. The annotators notes highlighted the difficulty in assigning one single category to some medical conditions. For example, for contexts that mentioned the condition *epilepsy* the annotators were unclear on whether this is a "cognitive" or a "sensory" condition; they would have perhaps selected both. This might change across different conditions.

7 Observations and Conclusions

The annotation disagreements described in §6 reflect the inherent subjectivity in interpreting historical texts that contain socially charged language. Annotators brought divergent perspectives on the historical role of terminology, the socio-political context of the sentences, and the contemporary implications of stigmatizing language. These divergences align with observations in prior research that annotation of socio-psychological constructs often entails subjective and multidimensional judgments (Pavlick and Kwiatkowski, 2019).

The annotation guidelines provided to annotators did not fully account for these interpretive differences. Future annotation tasks involving socially sensitive language would benefit from clearer operational definitions, explicit guidance on balancing historical and modern interpretations, and perhaps more granular label schemes. Another key challenge, also noted in hate speech annotation literature, is the variation in perceived offensiveness based on the background of the annotators and their relationship to the communities referenced in the texts (Vidgen et al., 2019). This is especially relevant for disability discourse, where community preferences around person-first versus identity-first language and perceptions of terms as outdated or offensive can differ widely. While our annotators were trained to reflect on historical and social context, future annotation efforts would benefit from including individuals with lived experience of disability or from adopting participatory annotation approaches that foreground community perspectives. Additionally, methods that embrace annotation disagreement such as soft labeling (Wu et al., 2023) may better reflect the inherent subjectivity of such tasks than traditional majority vote approaches. Other annotation disagreement challenges, such as different readings of a sentence's tone, remain outside the capabilities of textual representations and we consider them much harder to address through annotation protocols alone.

The findings from this analysis suggest several implications for the development of annotation schemes in the context of socio-political constructs and sensitive domains such as disability discourse. First, annotation tasks involving socio-psychological or politically charged constructs should acknowledge that disagreements are not necessarily indicative of noise, but may instead reflect valid differences in perspective that offer richer interpretive possibilities (Mostafazadeh Davani et al., 2022). Second, annotation protocols might benefit from incorporating structured reflection or justification fields, prompting annotators to explicitly state the reasoning behind their choices. Finally, our study highlights the need for methodological innovations in annotation aggregation. Majority voting may obscure valuable minority perspectives that offer critical insights into the data. Alternative approaches such as adjudication by discussion or perspectivist approaches (Cabitza et al., 2023) may be better suited to capturing the complexities inherent in the annotation of multidimensional socio-linguistic phenomena. Our analysis shows the deeply subjective nature of such annotation tasks. Where social and ethical considerations intersect with linguistic analysis, disagreements may be inevitable and even desirable, provided they are systematically analysed and leveraged.

Acknowledgments

This work was partially supported by the HE project MuseIT, co-funded by the European Union under Grant Agreement No. 101061441. The views and opinions expressed in this work are those of the authors and do not necessarily reflect the views of the European Union or the European Research Executive Agency. We thank MuseIT partners including Nasrine and Sandor for their valuable feedback in the pilot round, and Lloyd May for initial discussions. Special thanks to Andrea for her detailed feedback on the first draft.

Authors' contributions

BMcG designed the study, acquired the data for the annotation, developed the annotation protocol and wrote sections 1, 2, 4, 5, and B. NJ helped with design of study, data collection and formatting for the annotation session, co-led the annotation session and collection of data, performed the quantitative analysis of the dataset and the annotations, wrote section 5 and contributed to 6.1, 6.2 and 7 and refined the paper overall. CDB helped with the data collection, co-leading the annotation session, and writing section 6.2. AMP helped with the data collection and writing section 6.2.

Limitations

We are aware of the following limitations. **(1)** We only focused on English using readily available resources. However, exploring the applicability of this annotation protocol to other languages would be an important direction for future work, which could show interesting patterns about disability over time across languages. **(2)** We investigated a limited number of disability keywords. Although we diversified our data selection to account for multiple sources, multiple centuries, multiple intent of term, use of term and types of disability, future work should expand this annotation protocol to more disability keywords. **(3)** We did not conduct a fine-grained annotation analysis based on annotators' background. This was out-of-scope for this paper but we acknowledge the importance of this analysis for future work centered around subjectivity, especially given that domain expertise (e.g., in historical or medical texts) could influence annotation quality and help address cases of low agreement.

References

Erin E Andrews, Robyn M Powell, and Kara Ayers. 2022. The evolution of disability language: Choosing terms to describe disability. *Disability and Health Journal*, 15(3):101328.

Gail K Auslander and Nora Gold. 1999. Disability terminology in the media: a comparison of newspaper reports in Canada and Israel. *Social Science & Medicine*, 48(10):1395–1405.

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1509–1518.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Silvana Castano, Alfio Ferrara, Stefano Montanelli, Francesco Periti, et al. 2022. Semantic shift detection in vatican publications: a case study from leo xiii to francis. In *CEUR WORKSHOP PROCEEDINGS*, volume 3194, pages 231–243. CEUR-WS.

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

DE-BIAS Project consortium. 2025. De-bias: Vocabulary – english.

Phillip Ferrigon and Kevin Tucker. Person-first language vs. identity-first language: An examination of the gains and drawbacks of disability language in society. *Journal of Teaching Disability Studies*, 1:1–12.

Sanne Hoeken, Sophie Spliethoff, Silke Schwandt, Sina Zarrieß, and Özge Alaçam. 2023. Towards detecting lexical change of hate speech in historical data. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 100–111.

L Joseph and Levin Fleiss. 2023. *Statistical methods for rates and proportions*. Wiley-Blackwell.

Daphna Keidar, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. Slangvolution: A causal analysis of semantic change and frequency dynamics in slang. *arXiv preprint arXiv:2203.04651*.

Paola Marongiu, Barbara McGillivray, and Anas Fahad Khan. 2024. Multilingual workflows for semantic change research. *Journal of Open Humanities Data*.

Barbara McGillivray, Malithi Alahapperuma, Jonathan Cook, Chiara Di Bonaventura, Albert Meroño-Peñuela, Gareth Tyson, and Steven Wilson. 2022. Leveraging time-dependent lexical features for offensive language detection. In *Proceedings of the The First Workshop on Ever Evolving NLP (EvoNLP)*, pages 39–54.

Barbara McGillivray, Anas Fahad Khan, and Paola Marongiu. 2023. A new clarin resource family for lexical semantic change – final report. Technical report, Zenodo.

Albert Meroño-Peñuela, Romana Pernisch, Christophe Guéret, and Stefan Schlobach. 2021. Multi-domain and explainable prediction of changes in web vocabularies. In *Proceedings of the 11th Knowledge Capture Conference*, pages 193–200.

Stefano Montanelli and Francesco Periti. 2023. A survey on contextualised semantic shift detection. *Preprint*, arXiv:2304.01666.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

National Center on Disability and Journalism. 2021. National center on disability and journalism. National Center on Disability and Journalism.

Oxford University Press. 2024a. blind (adj., n.1, & adv.))). Oxford English Dictionary.

Oxford University Press. 2024b. lame, (adj. & n.). Oxford English Dictionary.

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

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Romana Pernisch, Daniele Dell'Aglio, Matthiew Horridge, Matthias Baumgartner, and Abraham Bernstein. 2019. Toward predicting impact of changes in evolving knowledge graphs. ISWC.

Romana Pernisch, Daniele Dell'Aglio, and Abraham Bernstein. 2021. Beware of the hierarchy—an analysis of ontology evolution and the materialisation impact for biomedical ontologies. *Journal of Web Semantics*, 70:100658.

Catia Pesquita and Francisco M. Couto. 2012. Predicting the Extension of Biomedical Ontologies. *PLoS Computational Biology*, 8(9):e1002630.

Labib Rahman. 2024. Disability language guide. Stanford University.

Mia Ridge, Nilo Pedrazzini, Miguel Vieira, Arianna Ciula, and Barbara McGillivray. 2024. Language of mechanisation crowdsourcing datasets from the living with machines project. *Journal of Open Humanities Data*.

Julia Rodina and Andrey Kutuzov. 2020. Rusemshift: a dataset of historical lexical semantic change in russian. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1037–1047.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. *Preprint*, arXiv:2007.11464.

Cecilia Capuzzi Simon. 2017. Disability studies: A new normal. In *Beginning with Disability*, pages 301–304. Routledge.

Charles Spearman. 1961. The proof and measurement of association between two things.

T.G. Stavropoulos, S. Andreadis, E. Kontopoulos, and I. Kompatsiaris. 2019. Semadrift: A hybrid method and visual tools to measure semantic drift in ontologies. *Journal of Web Semantics*, 54:87–106. Managing the Evolution and Preservation of the Data Web.

Ljiljiana Stojanovic. 2004. *Methods and Tools for Ontology Evolution*. Ph.D. thesis, University of Karlsruhe.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Wikipedia contributors. 2025. List of disability-related terms with negative connotations — Wikipedia, the free encyclopedia. [Online; accessed 18-March-2025].

Ben Wu, Yue Li, Yida Mu, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Don't waste a single annotation: improving single-label classifiers through soft labels. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5347–5355, Singapore. Association for Computational Linguistics.

Frank D Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. Lscdiscovery: A shared task on semantic change discovery and detection in spanish. *arXiv preprint arXiv:2205.06691*.

A Additional Results for Inter-annotator Agreement

Cohen's Kappa. The detailed pairwise results for Cohen's Kappa are shown in Table 2. With respect to Cohen's Kappa, we observe the following:

- Use of Terms: The "Use of Term" category shows mixed agreement among the annotators. For example, the term "Abnormal" has moderate agreement between A1 and A3 (0.50), but very low agreement between A1 and A2 (0.16). The terms "Blind" and "Cripple exhibit negative or low values in some comparisons, indicating weak or no agreement in those cases.

- Intent of Terms: The "Intent of Term" category shows a more consistent, although still low, agreement between annotators. The term

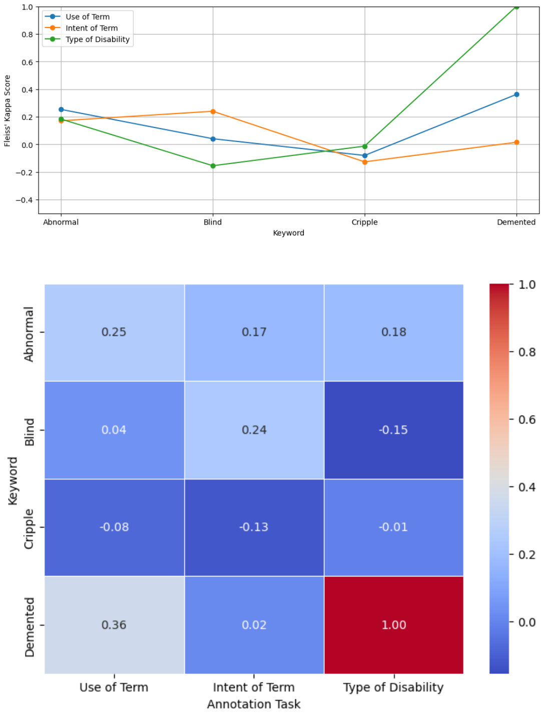

Figure 3: Comparative analysis of the Fleiss' Kappa scores across different keywords and annotation tasks.

"Blind" shows the strongest agreement between A1 and A3 (0.50), but the other terms exhibit lower kappa scores, suggesting more disagreement on the intent behind terms like “Abnormal” and “Cripple”.

- Type of Disability: This category shows somewhat better agreement, especially for the terms “Demented” and “Cripple”, which have full agreement or expected agreement scores between all pairs of annotators. In contrast, the term “Blind” shows negative or weak kappa scores across all pairs, suggesting minimal consensus on its classification as a type of disability.

Fleiss' Kappa. Figure 3 shows the Fleiss' Kappa scores and their variation across different terms.

Spearman's rank correlation. The results are visualized in Figure 4. Based on the results, we make the following observations for the annotations obtained for each keyword:

- Abnormal: The correlation between A1 and A2 (0.59) is moderate, indicating that their annotations show some alignment. However, A1 and A3 (0.19) and A2 and A3 (-0.10) show weak to negative correlations, suggesting discrepancies in the way these annotators interpreted the terms.

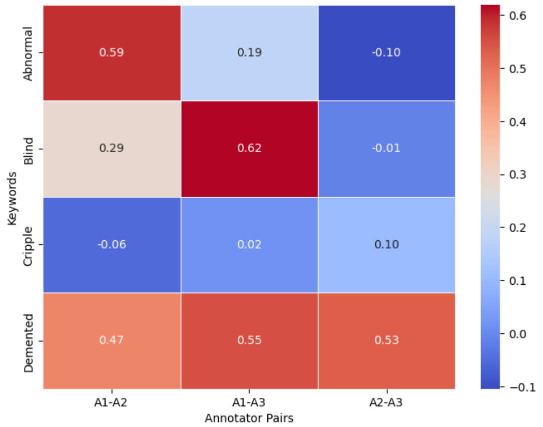

Figure 4: Comparative analysis of the Spearman's Rank Correlation scores across different keywords for the Intent annotations.

- Blind: The correlation between A1 and A3 (0.62) is relatively strong, indicating agreement between these two annotators. A1 and A2 (0.29) and A2 and A3 (-0.01) show weaker correlations, with A2 and A3 almost having no agreement at all.

- Cripple: All correlations are weak, with A1 and A2 (-0.06), A1 and A3 (0.02), and A2 and A3 (0.10), showing minimal or negative alignment. This suggests significant divergence in how these annotators approached the classification of terms.

- Demented: The correlations are generally higher, with A1 and A2 (0.47), A1 and A3 (0.55), and A2 and A3 (0.53) indicating a moderate to strong agreement across all annotators, suggesting more consistency in how these annotators rated the terms.

B Cases of Low Annotator Agreement

Here we present three examples of low annotator agreement.

Example 1: “Joe Hanlon, a cripple, had tits, and Cronin asked him for a match.” This is an account from a journal, most likely documenting conditions in an institutional setting—perhaps a psychiatric hospital, asylum, or another care facility. The narrator describes instances of abuse by a person named Cronin, presumably a staff member or attendant, towards several patients. The journal writer's tone is matter-of-fact, possibly reflecting either the norms

Term/Disability Type	Cohen's Kappa (A1 vs A2)	Cohen's Kappa (A1 vs A3)	Cohen's Kappa (A2 vs A3)	Fleiss' Kappa
Abnormal (Use of Term)	0.16	0.50	0.11	0.25
Blind (Use of Term)	0.21	-0.42	0.40	0.04
Cripple (Use of Term)	0.15	-0.07	-0.22	-0.08
Demented (Use of Term)	0.34	0.20	0.55	0.36
Abnormal (Intent of Term)	0.37	0.15	0.02	0.17
Blind (Intent of Term)	0.15	0.50	0.14	0.24
Cripple (Intent of Term)	-0.13	-0.15	-0.09	-0.13
Demented (Intent of Term)	0.08	-0.11	0.22	0.02
Abnormal (Type of Disability)	0.05	0.74	0.03	0.19
Blind (Type of Disability)	-0.25	0.00	0.00	-0.15
Cripple (Type of Disability)	1.00	0.00	0.00	-0.01
Demented (Type of Disability)	1.00	1.00	1.00	1.00

Table 2: Kappa scores for different terms and types of disability.

of the time or an attempt to objectively record events. The language reflects the historical attitudes toward the term *cripple* are likely seen today as offensive, though they may have been considered clinical or neutral by the writer. In this sentence, the annotators unanimously categorized the use of term *cripple* as common language. However, their assessments of Intent diverged substantially. One annotator interpreted the intent as Outdated but Neutral, while another annotator labeled it Mildly Pejorative or Stigmatizing, and the third annotator classified it as Strongly Pejorative or Insulting. This variation may be attributed to different readings of the sentence's tone. For one annotator, the use of *cripple* in this context may have reflected outdated but descriptive language, whereas another annotator may have perceived the sentence structure and reference as dehumanizing, intensifying the perceived stigma. The third annotator's annotation fellsbetween these extremes, reflecting uncertainty about whether the term is merely descriptive or carries additional pejorative force.

Example 2: "In the heat of their technical testimony they forgot the cripple seated at the far end of the room." In this case, two annotators labeled Use of Term as Formal Diagnosis, while the third annotator categorized it as Common Language. The Intent annotations again showed marked variation: one annotator perceived the term as Outdated but Neutral, whereas another annotator assigned Mildly Pejorative or Stigmatizing, and the third annotator assigned Strongly Pejorative or Insulting. The second annotator's notes indicate that their decision was guided by the broader context of the sentence, which they felt framed the reference to the *cripple* in a neutral, factual manner. The third annotator, on the other hand, appeared to prioritize the contemporary offensiveness of the term. The disagreement over Use suggests differing interpretations of whether *cripple* was historically considered a formal medical designation or a colloquial term, showing the difficulty of aligning modern sensibilities with historical usage.

Example 3: "The poor, the lame, the blind, the crippled, the outcast." This sentence generated consistent annotations for Use of Term (all three annotators selected Common Language), but Intent annotations were highly variable. The second annotator labeled it Neutral/Descriptive, suggesting an understanding that the sentence was listing marginalized groups without pejorative intent. In contrast, the first annotator classified it as Mildly Pejorative or Stigmatizing, and the third annotator as Strongly Pejorative or Insulting. The inclusion of *outcast* alongside terms for disability may have contributed to the third annotator's interpretation of heightened stigma. Furthermore, this annotator's detailed notes, distinguishing between different types of disabilities referenced in the sentence (e.g., *lame* as physical, *blind* as sensory), suggest an analytic focus on the cumulative social exclusions implied by the sentence structure.