Different Tastes of Entities: Investigating Human Label Variation in Named Entity Annotations

Siyao Peng[▲,🤖] **Zihang Sun**[▲] **Sebastian Loftus**[▲] **Barbara Plank**[▲,🤖]
[▲] MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[🤖] Munich Center for Machine Learning (MCML), Munich, Germany
`{siyaopeng, bplank}@cis.lmu.de` `{zihang.sun, s.loftus}@campus.lmu.de`

Abstract

Named Entity Recognition (NER) is a key information extraction task with a long-standing tradition. While recent studies address and aim to correct annotation errors via re-labeling efforts, little is known about the sources of human label variation, such as text ambiguity, annotation error, or guideline divergence. This is especially the case for high-quality datasets and beyond English CoNLL03. This paper studies disagreements in expert-annotated named entity datasets for three languages: English, Danish, and Bavarian. We show that text ambiguity and artificial guideline changes are dominant factors for diverse annotations among high-quality revisions. We survey student annotations on a subset of difficult entities and substantiate the feasibility and necessity of manifold annotations for understanding named entity ambiguities from a distributional perspective.

1 Introduction

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) (Yadav and Bethard, 2018). The task involves identifying named entities (NEs), such as *Justin Bieber*, *UNESCO*, and *Costa Rica*, and classifying them into semantic types, PER(son), ORG(anization), and LOC(ation), etc. Despite recent successes in achieving 93%+ strict F1 (Rücker and Akbik, 2023) on the English CoNLL03 benchmark (Tjong Kim Sang and De Meulder, 2003), recent research has observed that the percentage of noise in the data, particularly in the test partition, is comparable or even exceeding the error rates of state-of-the-art (SOTA) models (Wang et al., 2019; Reiss et al., 2020; Rücker and Akbik, 2023). They each conducted manual corrections or re-annotations, and model performances on their revised versions were higher than on the original. However, label variation in NEs, as shown in Table 1, remains an issue and hinders model performance.

Sentence
a. UK bookmakers [William Hill] ...
b. [ALPINE] SKIING ...
c. ... that there is no [God] .

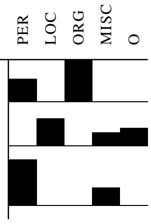

Table 1: Distribution of qualified student annotations on disagreed named entities in CoNLL03.

Human label variation (i.e., disagreement) refers to linguistically debatable cases where multiple labels are acceptable or appropriate in context (Plank et al., 2014; Jiang and de Marneffe, 2022). Recent studies that examine and benefit from disagreements among annotators challenge the conventional assumption of a single gold label. Learning from disagreements provides further insights into label distributions and preferences among human annotators (Uma et al., 2021a; Plank, 2022; Fetahu et al., 2023). However, there remains a gap for disagreement analyses on expert-labeled manifold NEs.

This paper presents quantitative and qualitative analyses of annotators' disagreements on labeling NEs in three Germanic variants: English, Danish, and Bavarian, in which multiple annotation efforts exist on the same documents. Unlike earlier studies that look at crowd-sourced data of unreliable quality (Rodrigues et al., 2014; Lu et al., 2023), we examine disagreements among expert annotations that went through iterations of published revisions and contrast them with the usual setting of independent annotators. §2 presents related work in disagreements and §3 demonstrates our setups. We analyze entity and label disagreements in §4, sources of disagreements in §5, and a student-surveyed annotation study in §6. §7 summarizes our work. We release our annotations and analyses on Github.[1]

[1] https://github.com/mainlp/NER-disagreements/

Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language, pages 73–81
March 21, 2024

2 Related Work

Despite disagreements between human judgments in subjective tasks (Prabhakaran et al., 2021; Davani et al., 2022; Fetahu et al., 2023; Leonardelli et al., 2023), annotation variation studies in NLP are recently on the rise (Uma et al., 2021a; Plank, 2022; Fetahu et al., 2023). These include part-of-speech tagging (Plank et al., 2014), anaphora and pronoun resolution (Poesio and Artstein, 2005; Poesio et al., 2019; Haber and Poesio, 2020), discourse relation labeling (Marchal et al., 2022; Pyatkin et al., 2023), word sense disambiguation (Passonneau et al., 2012; Navigli et al., 2013; Martínez Alonso et al., 2015), natural language inference (Nie et al., 2020; Jiang and de Marneffe, 2022; Liu et al., 2023), question answering (Min et al., 2020; Ferracane et al., 2021), to name a few.

In NER, Rodrigues et al. (2014) crowd-sourced problematic annotations from 47 Turkers on CoNLL03, scoring F1 of 17.60% the lowest and ~60% on average against CoNLL03 annotations, considerably under-performing the 90%+ inter-annotator agreement among expert annotators and SOTA model performances (Lu et al., 2023). Recently, Rücker and Akbik (2023) brought forward the newest CoNLL03 correction and thoroughly compared it with previous versions (Tjong Kim Sang and De Meulder, 2003; Wang et al., 2019; Reiss et al., 2020). However, many corrections are due to project-dependent guideline alternations and 2.34% of entities remain unresolved due to ambiguities. Thus, an onlooker assessment of NE disagreements and label variations is missing, particularly for expert annotations.

3 Datasets & Preprocessing

We analyze label variations in CoNLL03-styled PER/LOC/ORG/MISC NE annotations in three Germanic languages: English, Danish, and Bavarian (a Germanic dialect without standard orthography), where multiple annotation efforts on the same text documents are (or will be) available. Since the English CoNLL03 (Tjong Kim Sang and De Meulder, 2003) and Danish DDT (Plank, 2019) texts underwent iteration(s) of re-annotations or corrections by subsequent scholars, we conduct a diachronic comparison of the revisions for English and Danish. We also analyze disagreements on an in-house NE dataset for Bavarian German to distinguish disagreements among full-fledged corpora from independent unadjudicated annotations.

English The seminal English CoNLL03 dataset (henceforth `original`, Tjong Kim Sang and De Meulder 2003) presents the renowned NLP task to label flat and named entity spans into four major semantic types (PER, LOC, ORG, MISC) using (B)IO-encoding. The dataset includes 14.04K, 3.25K, and 3.45K sentences in its train, dev, and test partitions sourced from Reuters News between 1996-1997. Despite achieving 93%+ F1 score of the best systems on `original`, CoNLL03 annotations underwent several revisions (Wang et al., 2019; Reiss et al., 2020; Rücker and Akbik, 2023).

Wang et al. (2019) (`conllpp`) manually corrected 186 (5.38%) test sentences. Reiss et al. (2020) (`reiss`) used a semi-automatic approach to flag a larger quantity of error-prone labels (3.18K) in the entire dataset, and manually corrected 1.32K, including 421 in the test, as well as fixing tokenization and sentence splitting. They categorize these errors into six types: `Tag`, `Span`, `Both`, `Wrong`, `Sentence`, and `Token`. Rücker and Akbik (2023) (`clean`) present the most comprehensive relabeling effort by correcting 7.0% of all labels and adding a novel layer for entity linking. Though 5%+ of annotation errors were fixed compared to `original`, 2.34% of entities in `clean` remain ambiguous.

To establish fair comparisons, we manually align tokenization in the test partitions of `original`, `conllpp`, `reiss`, and `clean`. These include removing redundant line breaks, splitting hyphenized compounds, etc. Our alignment results in 46,738 test tokens across the four versions and 5,629, 5,683, 5,636, 5,725 annotated entities respectively.

Danish Plank (2019) annotates NEs on the dev and test partitions of the Danish Universal Dependencies (DDT, Johannsen et al. 2015). Plank et al. (2020) (`plank`) revise annotations, expand to more data and genres, and add -part/deriv suffixed labels and second-level nesting. Hvingelby et al. (2020) (`hvingelby`) re-annotate the dev and test sets of Plank (2019) by adding POS-marked proper nouns as NEs, resulting in ~0.75 and ~3.0 times more ORG and MISC NEs, such as nationalities and derived adjectives. We focus on the test partition (10,023 tokens) and compare `hvingelby` to the more recent `plank`, removing nesting and -part/deriv entities for cross-lingual analogy, leading to 564 and 531 NEs in `hvingelby` and `plank`.

Bavarian We additionally analyze the test partition of an in-house Bavarian NE dataset with ~12K tokens and ~400 entities on Wikipedia and Twitter

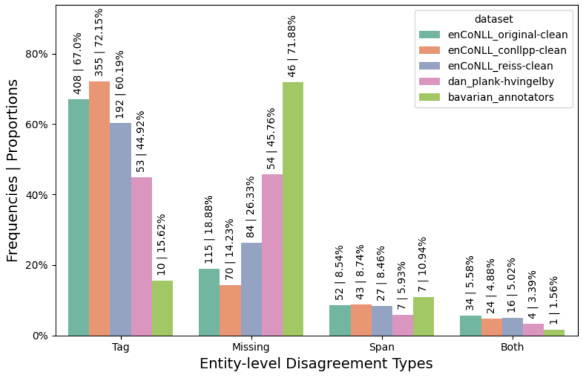

Figure 1: Proportions of entity-level disagreements in English `original-clean`, `conllpp-clean`, `reiss-clean`, Danish `plank-hvingelby`, and Bavarian.

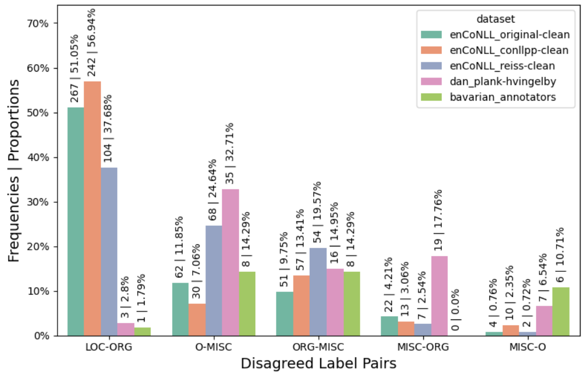

Figure 2: Proportions of top 5 label pairs in `Tag` and `Missing` disagreements in English, Danish, and Bavarian.

(X) annotated in 2023. Compared to the more established and iteratively revised English and Danish datasets, our Bavarian corpus represents the more common scenario of disagreements between two independent and unadjudicated annotations.

4 Entity-level Disagreements

Given our manually aligned tokenization across datasets, we modify Reiss et al. (2020)'s six error types into four entity-level disagreement types:

- `Tag`: same span selection, but different assigned tags, e.g., *[a b]*LOC vs. *[a b]*ORG;
- `Span`: different overlapping spans but the same tag, e.g., *[a b]*LOC vs. *[a]*LOC *b*;
- `Both`: overlapping spans with different tags, e.g., *[a b]*LOC vs. *[a]*ORG *b*;
- `Missing`: one annotator misses the entity completely, e.g., *[a b]*LOC vs. *a b*.

Figure 1 presents the frequencies and proportions of entity-level disagreements in five paired comparisons: English `original-clean`, `conllpp-clean`, `reiss-clean`, Danish `plank-hvingelby`, and between two Bavarian annotators. `Tag` disagreements contribute to most cases among repeatedly developed English corpora. On the other hand, Danish and Bavarian contain more `Missing` disagreements. Nevertheless, combining `Tag` and `Missing` accounts for 85%+ of disagreements in all comparisons across three languages. That is, entity tagging remains a bigger issue compared to span selection.

`Tag` and `Missing` disagreements are comparable in that both concern tagging the same entity span with different labels: the former with two different entity types (i.e., two non-O labels), and the latter with one entity type (a non-O label) and an O. Figure 2 displays the proportions of the top 5 disagreed label pairs in `Tag` and `Missing` disagreements across the five comparison scenarios (see Appendix A for a full list of label pairs). LOC-ORG, O-MISC and ORG-MISC are the most frequently disagreed label pairs in English comparisons, totaling 70%+ label disagreements. On the other hand, most (80%+) of Danish label disagreements concern MISC, whereas O-related (i.e., `Missing`) disagreements donate the majority (70%+) to Bavarian. To understand which factors trigger these label disagreements, §5 qualitatively analyzes the sources of human label variations in three languages.

5 Sources of Disagreements

Taxonomy We attribute NE label variations to three sources (Aroyo and Welty, 2015; Jiang and de Marneffe, 2022): 1) *text ambiguity* for uncertainties in the sentence meaning, 2) *guideline update* where NE type definitions vary across different guideline versions, and 3) *annotator error*. *Text ambiguity* could be caused by different interpretations with or without enough context that hinders pinpointing a definitive reference. *Guideline update* occurs when one annotation version is incoherent with another guideline. This is dominant in our analyses since annotation projects consist of iterations of guidelines and annotation revisions. For instance, whether proper noun-derived adjectives, e.g., *[ALPINE]* in Table 1, should be LOC, MISC, or not an entity (i.e., O); whether polysemous LOC/ORG entities are labeled LOC or ORG depending on context, or always as MISC. The

Source types	English		Danish		Bavarian	
text ambiguity	19	9.5%	7	6.0%	10	15.6%
guideline update	160	80.0%	62	52.5%	11	17.2%
annotator error	21	10.5%	49	41.5%	43	67.2%
Total	200	100.0%	118	100.0%	64	100.0%

Table 2: Sources of label disagreements and their distributions in English, Danish, and Bavarian samples.

last category, *annotator error*, refers to annotations that differ from a single deterministic ground truth. Closer inspections could fix annotators' attention slip errors, whereas special cultural knowledge is needed for resolving knowledge gap disagreements. We manually annotate a small sample of disagreements in three languages using these source categories to separate guideline changes and textual ambiguities from annotators' mistakes.

Setup For English, we sample 200 disagreed test entities between the `original` and the most recent `clean` annotation. Since the Danish `plank-hvingelby` comparisons and the Bavarian double annotations have much smaller test sets, we sample all test disagreements in the two languages, 118 entities in Danish and 64 in Bavarian. Each language sample is assessed by one computational linguist who speaks that language. Table 2 presents the source of disagreement results.

Additionally, we measure inter-annotator agreement (IAA) on source classes between two assessors[2] on 50 ambiguous English `original-clean` test entities and achieve 61.73% Cohen's kappa. Assessors find the hardest differentiating whether the lack of contextual information resulted from annotators' personal knowledge backgrounds (*annotator error)* or the settings behind text segments (*text ambiguity*). Even though surrounding sentences are provided, NE annotators tend to focus on the nearer context for NE tagging.

English In the `original-clean` comparison, most (80.0%) of disagreements stem from differences in *guideline update*. To disambiguate inconsistent cases in `original`, `clean` updated the guideline to be less context-dependent: 1) ORG instead of LOC for national sports teams as well as public facilities, even for *the flight to [Atlanta]*ORG; 2) MISC is used for more abstract institutions and adjectival affiliations e.g., *[Czech]*MISC *politics*; 3) instead of further correcting tokenizations and splitting hyphenated compounds, they assign labels that are relevant to part of the compound to the entirety, e.g., *[German-born]*MISC. Aside from *guideline update*, ambiguities occur for religious deities, such as whether *[Allah]* or *[God]* should be PER, MISC or O (see Table 1). Previous automatic conversions from IO-encodings in `original` to BIO in `clean` also caused disagreements since it is hard to tell apart if a sequence of I-tags is one entity or multiple, e.g., *[Spanish]*MISC *[Super Cup]*MISC or *[Spanish Super Cup]*MISC.

[2] We use "assessors" to refer to our source of disagreement coders and differentiate from "annotators" of the NE datasets.

Danish Akin to the English analysis, we found that large parts (52.5%) of the ambiguous cases in Danish stem from *guideline updates*, e.g., frequently mentioned ferry routes are labeled LOC in `hvingelby` but MISC in `plank`. Besides, we found 41.5% of disagreements are *annotator errors*, and the majority are ORG-MISC disagreements and concern a single hyphen-joint token with two sports clubs, e.g., *[Vejle-Ikast]*. This points out a disadvantage of the current cross-lingual comparable analysis — compounding morphology prevails in Danish and Bavarian, and removing -part/deriv labels leads to information loss.

Bavarian We present the less developed but more common scenario of disagreements between two unadjudicated annotations in Bavarian. Though achieving 85%+ Span IAA, *annotator error* (67.2%) remains the highest source of disagreements. Apart from local entities, e.g., *[Feucht]*$_{\text{loc}}$ (a small town in Bavaria), that require geographical knowledge or detailed search, many of these *annotator errors* classified based on the Bavarian guideline are indeed acceptable under certain versions of the English CoNLL guidelines. For example, when *[Edeka]* (a supermarket chain) functions as a destination, the disagreement between LOC-ORG is classified as *annotator error* in Bavarian, but would rather be a *guideline update* in English.

6 Surveying Student Annotations

Though NE guidelines can be meticulously different from each other, the underlying concepts of PER, LOC, ORG are cognitively straightforward. To inspect the distribution of multiple interpretations, we follow Liu et al. (2023) to survey annotations from 27 bachelor and master students in computational linguistics at LMU Munich. We gave them a 7-minute introduction to NEs, walked through the CoNLL03 guideline,[3] and showed some examples of type ambiguities in NE annotations. Students

[3] `www.cnts.ua.ac.be/conll2003/ner/annotation.txt`

were instructed in the classroom to annotate entity types in English and Bavarian selected from difficult examples in §5.[4] We further sample 10 representative English CoNLL entities for the qualitative evaluation below.[5] To ensure the quality of student-surveyed annotations, we only keep an annotation if 80%+ of entity labels match any of the four CoNLL annotations. Table 1 demonstrates the distribution of 14 qualified student annotations on three examples (see Appendix B for the ten representative English CoNLL entities).

Results demonstrate that label variation across annotation projects are also prevalent in the student-surveyed annotations. On one side, even with a brief training, students were able to disambiguate the contextual interpretations between [the away team]ORG and [the home team]LOC in *[LA CLIPPERS]*ORG *AT [NEW YORK]*LOC. Our participants also recognize the collectiveness of *[White House]*ORG, *[Australia]*ORG, etc., and the fixedness of *[EST]*MISC (Eastern Standard Time). On the other hand, knowledge gap or insufficient context contribute to the high variance of *[William Hill]*, whether it refers to [the businessman]PER or [the gambling bookstore he created]ORG. Annotators also diverge in marginal cases: whether *[God]* is PER or MISC and whether nominal derivatives *ALPINE* and *Fascist* are NEs.

7 Conclusion

This paper examines named entity disagreements across expert annotations and contrasts them with the more common setting of individual annotations. We demonstrate that human label variation, e.g., LOC-ORG and ORG-MISC, contribute to most English, Danish, and Bavarian disagreements. We also discover that *guideline updates* and *text ambiguities* are leading sources of disagreements in established English and German datasets, whereas *annotator errors* remain the dominant cause for the new Bavarian corpus. Lastly, we survey student annotations and encourage more researchers to explore NE label variations to narrow the gap to model performance.

Though modeling NER from label variation is out of the scope of this paper, we embrace the prospect of learning from disagreements (Uma et al., 2021b). Particularly, we look forward to conducting annotations on a much larger scale in terms of both the number of participants and annotated instances to provide more statistically meaningful NE distributions for NER models. Future work also includes separating valid label variations from true annotation mistakes by leveraging Automatic Error Detection (AED) methods (Klie et al., 2023; Weber and Plank, 2023). We hope tackling NER through label variations can remedy the conflicts among versions of annotation guidelines.

[4] Students acknowledge that their annotations could be used for research purposes.

[5] The full English and Bavarian student-surveyed annotations are available on GitHub.

Acknowledgements

We would like to thank Verena Blaschke for giving feedback on earlier drafts of this paper. This project is supported by ERC Consolidator Grant DIALECT 101043235.

References

Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1):15–24. Number: 1.

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.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2021. Did they answer? Subjective acts and intents in conversational discourse. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1644, Online. Association for Computational Linguistics.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023. SemEval-2023 Task 2: Fine-grained Multilingual Named Entity Recognition (MultiCoNER 2). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2247–2265, Toronto, Canada. Association for Computational Linguistics.

Janosch Haber and Massimo Poesio. 2020. Classification of low-agreement Pronouns through Collaborative Dialogue: A Proof of Concept.

Rasmus Hvingelby, Amalie Brogaard Pauli, Maria Barrett, Christina Rosted, Lasse Malm Lidegaard, and Anders Søgaard. 2020. DaNE: A Named Entity Resource for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4597–4604, Marseille, France. European Language Resources Association.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating Reasons for Disagreement in

Natural Language Inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374. Place: Cambridge, MA Publisher: MIT Press.

Anders Johannsen, Héctor Martínez Alonso, and Barbara Plank. 2015. Universal dependencies for danish. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 157–167.

Jan-Christoph Klie, Bonnie Webber, and Iryna Gurevych. 2023. Annotation Error Detection: Analyzing the Past and Present for a More Coherent Future. *Computational Linguistics*, 49(1):157–198.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 Task 11: Learning with Disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. We're Afraid Language Models Aren't Modeling Ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are Emergent Abilities in Large Language Models just In-Context Learning? ArXiv:2309.01809 [cs].

Marian Marchal, Merel Scholman, Frances Yung, and Vera Demberg. 2022. Establishing Annotation Quality in Multi-label Annotations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3659–3668, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Héctor Martínez Alonso, Anders Johannsen, Oier Lopez de Lacalle, and Eneko Agirre. 2015. Predicting word sense annotation agreement. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 89–94, Lisbon, Portugal. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering Ambiguous Open-domain Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What Can We Learn from Collective Human Opinions on Natural Language Inference Data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Rebecca J. Passonneau, Vikas Bhardwaj, Ansaf Salleb-Aouissi, and Nancy Ide. 2012. Multiplicity and word sense: evaluating and learning from multiply labeled word sense annotations. *Language Resources and Evaluation*, 46(2):219–252.

Barbara Plank. 2019. Neural Cross-Lingual Transfer and Limited Annotated Data for Named Entity Recognition in Danish. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 370–375, Turku, Finland. Linköping University Electronic Press.

Barbara Plank. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. DaN+: Danish Nested Named Entities and Lexical Normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6649–6662, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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. Association for Computational Linguistics.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A Crowdsourced Corpus of Multiple Judgments and Disagreement on Anaphoric Interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789, Minneapolis, Minnesota. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On Releasing Annotator-Level Labels and Information in Datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Valentina Pyatkin, Frances Yung, Merel C. J. Scholman, Reut Tsarfaty, Ido Dagan, and Vera Demberg. 2023. Design Choices for Crowdsourcing Implicit Discourse Relations: Revealing the Biases Introduced by Task Design. ArXiv:2304.00815 [cs].

Frederick Reiss, Hong Xu, Bryan Cutler, Karthik Muthuraman, and Zachary Eichenberger. 2020. Identifying Incorrect Labels in the CoNLL-2003 Corpus. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 215–226, Online. Association for Computational Linguistics.

Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. 2014. Sequence labeling with multiple annotators. *Machine Learning*, 95(2):165–181.

Susanna Rücker and Alan Akbik. 2023. CleanCoNLL: A Nearly Noise-Free Named Entity Recognition Dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8628–8645, Singapore. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Alexandra Uma, Tommaso Fornaciari, Anca Dumitrache, Tristan Miller, Jon Chamberlain, Barbara Plank, Edwin Simpson, and Massimo Poesio. 2021a. SemEval-2021 Task 12: Learning with Disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online. Association for Computational Linguistics.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021b. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. 2019. CrossWeigh: Training Named Entity Tagger from Imperfect Annotations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5154–5163, Hong Kong, China. Association for Computational Linguistics.

Leon Weber and Barbara Plank. 2023. ActiveAED: A Human in the Loop Improves Annotation Error Detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8834–8845, Toronto, Canada. Association for Computational Linguistics.

Vikas Yadav and Steven Bethard. 2018. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

A Proportions of Disagreed Label Pairs

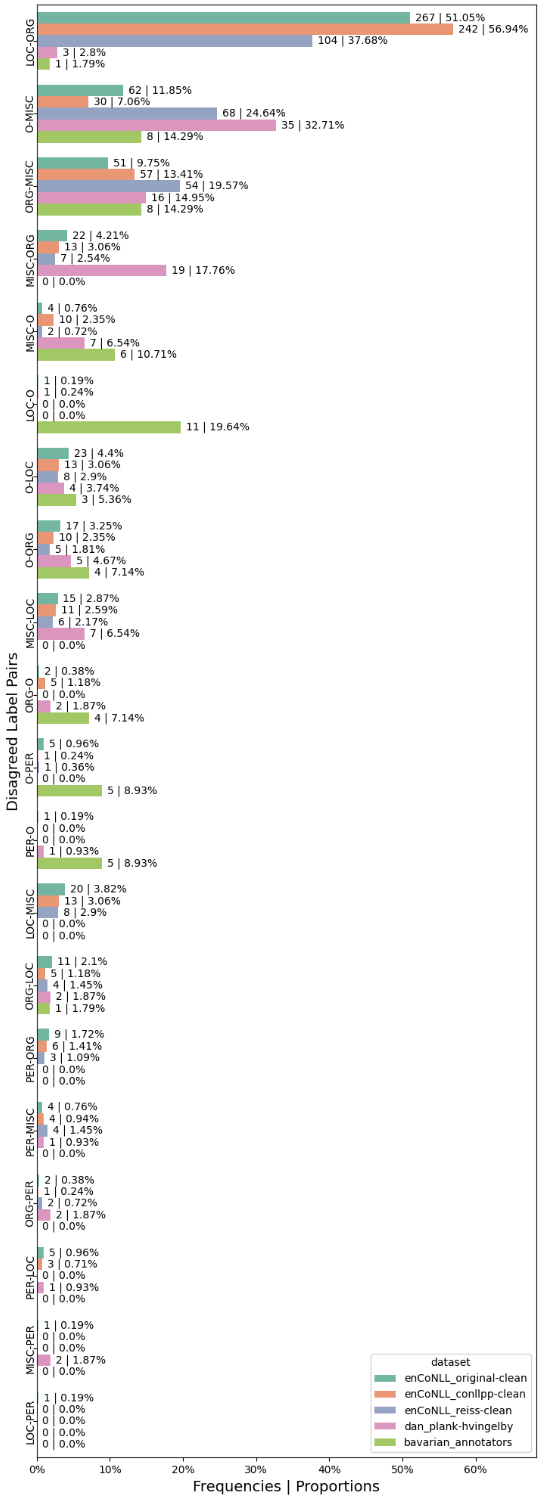

Figure 3: Proportions of label pairs (full) in `Tag` and `Missing` disagreements in English, Danish, and Bavarian.

B Student Surveyed NE Annotations

Sentence	PER	LOC	ORG	MISC	O	*abstained*

[ALPINE] SKIING		6
		`clean`
				3	4
					`original`
					`conllpp`
					`reiss`
						1

[LA CLIPPERS] AT NEW YORK			13
			`original`
			`conllpp`
			`reiss`
			`clean`
						1

LA CLIPPERS AT [NEW YORK]		14
		`original`
		`conllpp`
		`reiss`
			0
			`clean`			

[White House] spokesman Mike McCurry said Clinton plans to have regular news conferences during his second term .		2
		`original`
		`conllpp`
		`reiss`
			11
			`clean`
				1		

UK bookmakers [William Hill][6] *said on Friday they have lengthened the odds of a Conservative victory .*	5
	`original`
	`conllpp`
	`reiss`
			9
			`clean`			

The man who kicked [Australia] to defeat with a last-ditch drop-goal in the World Cup quarter-final in Cape Town .		5
		`original`
		`conllpp`
		`reiss`
			9
			`clean`			

The years I spent as (soccer team) manager of the [Republic of Ireland] were the best years of my life .		4
		`original`
		`conllpp`
		`reiss`
			9
			`clean`
				1		

I bear witness that there is no [God] .	10
	`original`
	`conllpp`
	`reiss`
				4
				`clean`		

The granddaughter of Italy's [Fascist][7] *dictator Benito Mussolini*			3	3
				`clean`
					8
					`original`
					`conllpp`
					`reiss`	

at about 3 A.M. local time / 1:30 A.M. [EST]				10
				`clean`
					2
					`original`
					`conllpp`
					`reiss`
						2

Table 3: 14 classroom surveyed and qualified annotations on difficult disagreement cases in CoNLL03 test.