An IDR Framework of Opportunities and Barriers between HCI and NLP

Nanna Inie 🍆
IT University of Copenhagen
2300 Denmark
nans@itu.dk

Leon Derczynski 🍆
IT University of Copenhagen
2300 Denmark
leod@itu.dk

Abstract

This paper presents a framework of opportunities and barriers/risks between the two research fields Natural Language Processing (NLP) and Human-Computer Interaction (HCI). The framework is constructed by following an interdisciplinary research-model (IDR), combining field-specific knowledge with existing work in the two fields. The resulting framework is intended as a departure point for discussion and inspiration for research collaborations.

1 Motivation

Research has long suggested that the fields of Natural Language Processing (NLP) and Human-Computer Interaction (HCI) could both benefit from each other's methods, analyses, and findings, e.g. De Angeli and Petrelli (2000); Ozkan and Paris (2002); Green et al. (2015); Hung (2014). Despite this, it is also regularly pointed out that overlaps between the fields are still rare (Karamanis et al., 2009; Munteanu et al., 2013; Yang, 2017).

Reasons put forward for this gap have included differences of methods and evaluations (Ozkan and Paris, 2002), a perception that language technology is not advanced enough for the processing required in HCI work (Munteanu et al., 2013), as well as lack of programming skills required to work with NLP: *"Deploying an NLP system typically requires a substantial amount of time from an expert NLP developer – normally, applications do not generalize and must be rebuilt, retrained, enhanced, and re-evaluated for each new task"* (Chapman et al., 2011). HCI may even suffer from being perceived as a "soft science", lacking quantitative results allocated merit equal to that in more mathematically founded computer science, e.g. Wang (2013).

🍆: These authors contributed to the paper equally.

To better define this interface, this short paper presents a framework of opportunities and barriers/risks for both HCI and NLP in the combination of the two fields. We hope to inspire collaborations between researchers from both fields by presenting some practical gains in cross-pollination, as well as some of the potential pitfalls and challenges researchers might meet.

2 An IDR framework

Our interdisciplinary framework is presented in Table 1. It follows the approach described by CohenMiller and Pate (2019). They present a model for developing an interdisciplinary research (IDR) framework founded in two fields by identifying research topics, concepts, disciplines, theory, and terminology from each field. The objective is to create an integrated framework relevant to both.

The authors of this paper have backgrounds in HCI and NLP, respectively, and while we see potential in applying the distinct methods and analyses from both fields in the other field, we have focused especially on identifying common benefits; i.e. research *outputs* that benefit science and research in general, not only in one or the other field.

Our method is to debate and discuss research processes and methods, identifying differences (often in methods) and similarities (often in goals) between the fields. In the following, we attach some words to each of the topics in the framework.

3 Opportunities

3.1 Methods and analyses

It has been suggested before that NLP and HCI could benefit from applying each others' methods and analyses – by researchers in both fields, e.g. Munteanu et al. (2013); Ozkan and Paris (2002). Collaborations have been exemplified by specific cases, i.e. *applying NLP in clinical research* (Kara-

	Methods and analyses	Output
Opportunities, NLP	- Ethnographic methods - Annotation processes - Error analysis - New abstractions - Extra data layers	- Discoveries - Quality analyses - Artwork - Network
Opportunities, HCI	- Automatic transcription - Thematic classification and clustering - Novel analysis tools	
Barriers/risks, human	- Coding skills - Detaching from quantitative methods - Terminology - Critical thinking	- Biases in data
Barriers/risks, technical	- Limited modalities (speech and text processing) - Data access	

Table 1: An IDR framework of opportunities and barriers/risks in the intersection of NLP and HCI research

manis et al., 2009), *developing language-controlled interfaces* (Munteanu et al., 2013), or *designing interactive documents* (Ozkan and Paris, 2002). In this framework, we have tried to focus on a broad view of methods from NLP which may be useful in HCI research and vice versa. For clarity and overview, we present a brief selection of topics which we found important or obvious. The framework could and should be expanded in use by researchers from both fields.

3.1.1 Opportunities for NLP

Ethnographic methods. While NLP studies *natural* language and often reports findings about culture and human interaction, NLP rarely studies the contexts and realities of people. HCI has a strong tradition of researching to understand people and their behaviors and realities through methods of inquiry (Blomberg et al., 2009), as well as viewing behaviors as *situated* in a context which influences and is influenced by human actions (Suchman, 1987). We argue that NLP analyses could gain depth and quality if combined with ethnographic studies of the people whose language are processed.

Understanding how and where to deploy NLP technology in an effective way is a challenge (e.g. Tolmie et al. (2017); Mellinger (2017)) which may be tackled using ethnographic methods. NLP researchers are adept at optimising, so it is important to make sure the right goal is being optimised for. For example, what makes a machine translation output good depends on stakeholder and situation, regardless of the (e.g.) chrF score that indicates "objective" translation performance.

Annotation processes. The creation of NLP datasets hinges on quality human annotation. A core challenge in NLP annotation is to develop schemas that allow computation, that model a real-world phenomenon, and that are understandable by annotators (Pustejovsky and Stubbs, 2012; Ferro et al., 2005). These schemas and their description benefit from having a situated understanding the phenomena well and having clear interactions around it. After all, machine learning models tend to learn the behaviours represented their training data – which may not always be desirable (Reidsma and Carletta, 2008; Bender et al., 2021). However, while these schemas are typically developed with both linguistic and computational constraints in mind, the fact that they are to be applied to language by humans in order to generate data is seldom recognised, despite this being the critical part of the process.

Another challenge in annotation processes is assessing quality: annotator disagreement is often a valuable signal, indicating e.g. rich, multi-phenomena instances (Das et al., 2017; Sommerauer et al., 2020) or cases where only an annotator minority has the knowledge to complete the task (Derczynski et al., 2016; Vidgen and Derczynski, 2020). This disagreement, which signals a problem in the interface between human annotator and the dataset goal, should be investigated: its sources are likely to be interesting, and at the least, disagreement can help downstream applications (Plank et al., 2014). HCI offers the tools for exploring and better understanding that interface.

HCI, concerned with assessing qualitative data, offers numerous methods, discussions, and reflections on reliability and when to use (and when *not* to use) inter-rater reliability (IRR) as a measure (McDonald et al., 2019). Analysis work in HCI is often *iterative*, with several steps of testing and inclusion of relevant stakeholders (Zimmerman et al., 2007). In HCI, there is no one correct way to assess reliability – different studies invite different frames and methods. Therefore, HCI research must be *transparent* and *reflective* about its methodology and how analysis is performed, including how analysis categories and schemas are developed (Siegert et al., 2014; Olson and Kellogg, 2014).

We believe methodologies for developing NLP datasets should be discussed and detailed with a similar level of iterativeness and reflection.

Error analysis. Understanding and describing how NLP models and systems go wrong requires a different toolkit to the quantitative approaches typically used (e.g. feature ablation, layer freezing, BERTology). NLP has tended to focus more on quantitative rather than qualitative data, leaving potential gaps of understanding, where thorough and methodical qualitative analyses stand to yield new research, especially when combined with existing quantitative methods (e.g. (Bornakke and Due, 2018); also Section 4.1.1).

Qualitative error analysis opens paths to both better understanding of language and to developing new technical results. HCI, building on cognitive and social science, offers many methods and frameworks for establishing such an understanding, e.g. *cognitive modelling* (Olson and Olson, 1995) and *chains of cognitive breakdowns* (Ko and Myers, 2005). These framings and methodologies could readily improve NLP error analyses.

New abstractions. HCI is a broad field with many interactions and concepts within its aegis. These can lead to new abstractions regarding the use and structure of NLP technology. For example, the abstract concept of a design material – a conceptual, tangible, or other item used in or by a design process – can be applied to machine learning, improving understanding of how machine learning (or ML-based NLP) can be used or useful (Dove et al., 2017). Similarly, the role played by linguistic actions (e.g. conversation analysis (Norman and Thomas, 1991; Hirst, 1991)) may be differently understood in various HCI frameworks, giving new interfaces that may lead to a deeper problem understanding. A concrete example of such work is VoxML's use of affordances and embodiment for semantic disambiguation (Pustejovsky and Krishnaswamy, 2020).

Extra data layers. HCI research produces vast amounts of data, a lot of it transcribed. Interviews, field notes, observations, and video transcripts could all serve as readily accessible data for NLP training datasets and analyses. This novel modality gives an interface for applying NLP to support HCI research while presenting novel NLP tasks and text genres.

3.1.2 Opportunities for HCI

Automatic transcription. One of the major potential gains for HCI in the automatic processing of language is automatic transcription of qualitative data, for instance interviews and video recordings. While some tools are already available for this, they are usually substandard to human transcription, and only work for English language.

Thematic classification and clustering. Both NLP and HCI analyze text by tagging it with themes or classes. HCI researchers spend years training to identify important themes in qualitative data, in HCI known as *thematic coding* (Gibbs, 2007), *grounded theory development* (Strauss and Corbin, 1997) or *affinity diagramming* (Lucero, 2015). Applying NLP methods to identify themes in qualitative data may reveal themes (or prevalence of themes) that researchers would not have otherwise identified.Concrete examples of successful application of automatic coding of qualitative data is described in Marathe and Toyama (2018) and Crowston et al. (2012)'s work, which both achieved promising IRR scores combining semi-automatic NLP techniques with human coders.

Novel analysis tools. NLP offers an array of analysis tools which can be used in HCI to analyze text in a consistent fashion in large quantities of data. This opens up new possibilities for analyzing and comparing language, discovering facets of data which were otherwise not apparent.

Preliminary HCI research has, e.g., explored available NLP tools such as sentiment analysis to explore how frustrated students are while progressing through a design process (Frich et al., 2018).

3.2 Output opportunities

New methods and types of analyses are intriguing, but it is valuable to think specifically about the results and outputs one would like to achieve in applying different methods. We believe it is useful to think about these outputs *before* choosing methods from a novel field. Relevant methods may, e.g., vary depending on career stage, and whether one is conducting basic or applied research.

Discoveries. The most apparent goal of interdisciplinary research should of course be novel research results. Using the same pen and paper to draw every day will increase one's drawing skills, but it will probably not generate an entirely new artistic expression. Using paint and a canvas may be challenging, but it may produce surprising results. Therefore, the first output topic is *discoveries*, covering any form of novel research output which would not have been otherwise obvious. Examples of research discoveries in the intersection of NLP and HCI include: creating and evaluating automated feedback to psychotherapists (Hirsch et al., 2018); exploring cultural biases in media coverage (El Ali et al., 2018); using NLP to support creative journalism (Maiden et al., 2018); and integrating afforded behaviours with lexical semantics to model motion (Krishnaswamy and Pustejovsky, 2016).

Quality analyses. Another clear output of applying NLP and HCI methods to each others' datasets is increasing the quality of analyses. While it is not given that randomly matching methods from two different research fields will produce novel results, we believe NLP and HCI can uniquely augment each other. Both fields are concerned with *communication* as a core concept, and humans' most natural forms of communication, speech and language, are also among the most difficult modalities for machines to process (Ozkan and Paris, 2002; Munteanu et al., 2013).

Artwork. HCI conferences have a long-standing history of exhibiting novel technology demonstrations, sometimes in the intersection between utility and art. Traditionally, the realm of computational art has predominantly featured audio and visual works. Language modelling has driven some artwork, such as automatically generated poetry and, with more recent models, a broader range of artworks (e.g. Inie et al. (2020); Rubin (2002)). The range and depth sophistication in NLP is a potentially exciting palette for artists to work with, allowing a novel modality. For example, recent work includes reconsidering storytelling from a generation point of view, which is already leading to new understandings of what constitutes a story (Ammanabrolu et al., 2021). Similarly, artistic ideas for application of NLP may push the boundaries of what is currently computationally possible.

Network. Creative thinking requires a combination of field knowledge and conceptually distant inspiration (Chan et al., 2018). Researchers are usually experts in their own domain, and while we can all benefit from learning new skills, interdisciplinary research is strongest when experts from different fields are brought together. By conducting genuinely interdisciplinary research, one tangible outcome is to discover research avenues outside one's core area, as well as hopefully attending events, workshops such as this one, and conferences where experts from other fields are easy to meet and truly novel collaborations become possible.

4 Barriers and risks

Interdisciplinary research can be challenging and can carry inherent risks which may be mitigated through thorough reflection and preparation. Here, we present some of the primary barriers for combining NLP and HCI research methods, as well as risks for the resulting research output.

4.1 Methods and analyses barriers

4.1.1 Methods barriers, human

Coding skills. One of the main barriers for novices to build, adjust, and integrate NLP systems is programming ability. While the level of proficiency required to access the most advanced systems is lower than previously, implementing these systems using even a well-wrapped library (by today's standards) requires a fairly high level of coding aptitude. This creates a barrier to using NLP technology. Providing simple interfaces to this technology helps reduce this barrier. Various programming toolkits have been developed and flourished over time to reduce this barrier (Maynard et al., 2012; Bird et al., 2009; Wolf et al., 2020), but mostly made NLP more accessible only to those with established programming skills.

Detaching from quantitative methods. Numbers and code offer the illusion of objectivity and detail. This presents barriers to both use and uptake of qualitative methods. While stable, methodical, and

scientific qualitative analysis methods exist, they are not well-established in NLP, creating a risk that NLP researchers and reviewers may have a dimmer view of them due to a lack of familiarity. Establishing excellent qualitative analysis as common practice will require persistence. Frameworks used in HCI applicable to NLP include *thematic analysis*, where themes are identified subjectively by humans based on evidence in results (Castleberry and Nolen, 2018); *in vivo coding*, where attention is placed on participants' use of phrases (Manning, 2017), useful during e.g. annotator debriefing; and *grounded theory method*, that constructs hypotheses through examining data for phenomena where no theory yet exists (Muller, 2014).

Terminology. Developing a lingua franca is a necessity for any two research fields to meet and share knowledge. Each research field has its own terminology and assembly of concepts which mean specific things in specific contexts (CohenMiller and Pate, 2019). An obvious example is the term "coding", which means quite different activities in machine learning and qualitative analysis. While embarking on collaborations between NLP and HCI, it may be worth working on research dictionaries or referenced repositories which define concepts like *toolkit*, *codebook*, and *ontololgy* in easily understandable terms from each field's perspective.

Critical thinking. From an HCI perspective, quantitative results can seem enchanting in their promise of objective truth. Especially when researchers do not have thorough insight into the algorithms that produce numbers (and the biases built into these algorithms), there is a risk of accepting quantitative results as truthful without reflecting critically (O'Neil, 2016). Using NLP methods therefore requires a level of insight into the methods necessary to afford critical examination and questioning of results.

4.1.2 Methods barriers, technical

Limited modalities (speech & text processing). A significant amount of qualitative data from HCI exists first as audio or video recordings. One challenge with such data is that, even if automatic transcription was flawless and available in all languages, NLP focuses on the spoken word, not context or actions. Context, actions, and interactions are often essential information to HCI research. Therefore, valuable NLP transcription/analysis systems might support researchers in annotating actions, interactions, important moments and field notes in relation to spoken and written language.

Access to data. Data is stored, processed, and accessed in different ways in NLP and HCI – because the fields traditionally store different kinds of data. NLP datasets are often stored as large .tsv or JSONL files, while HCI datasets may consist of complex field notes and transcriptions in specialized software like ATLAS.ti or NVIVO. These all come with specific technical requirements for storage and access. Data storage significantly influences the way data is *retrieved*, which is crucial to enabling different kinds of analyses by different researchers.

4.2 Output risks

Biases in data. All data is biased – by the way it is sampled, by the goal behind gathering and annotating it, by the individual implementing its assembly. These biases may present as: class overrepresentation and underrepresentation; missing phenomena, such as languages, entity names, lexical items, or syntactic structures; skew in treatment of borderline cases; and so on (Søgaard et al., 2014). The important part is to label the biases, so they may be addressed and communicated. While NLP researchers and HCI researchers have the expertise and experiential knowledge to be capable of recognising and documenting the biases within their own field, it is harder to properly understand the biases present in data from another discipline. This is a challenge faced by people on both sides of the HCI/NLP interface. We should be each cognisant of our potential lack of insight into the other field, and aware that data from it might be mis-applied, or that assumptions may not port well from one field to another. An example mitigation is the Data Statement for NLP, itself based on an intersection of these two disciplines, a methodological step that documents intent and factors contributing to bias in order to communicate these to data users (Bender and Friedman, 2018).

5 Outlook

The topics of the interdisciplinary HCI/NLP framework described in this short paper are the result of the authors' initial dialogue and analysis, and they are focused on being broadly relevant and relatable to researchers in both areas. We invite researchers in both NLP and HCI to expand upon the categories, in detail as well in quantity.

References

Prithviraj Ammanabrolu, Wesley Cheung, William Broniec, and Mark O Riedl. 2021. Automated storytelling via causal, commonsense plot ordering. In *Proceedings of AAAI*.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? 🦜. *Proceedings of FAccT*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Jeanette Blomberg, Mark Burrell, and Greg Guest. 2009. An ethnographic approach to design. *Human-Computer Interaction*, pages 71–94.

Tobias Bornakke and Brian L Due. 2018. Big–thick blending: A method for mixing analytical insights from big and thick data sources. *Big Data & Society*, 5(1):2053951718765026.

Ashley Castleberry and Amanda Nolen. 2018. Thematic analysis of qualitative research data: Is it as easy as it sounds? *Currents in Pharmacy Teaching and Learning*, 10(6):807–815.

Joel Chan, Steven P Dow, and Christian D Schunn. 2018. Do the best design ideas (really) come from conceptually distant sources of inspiration? In *Engineering a Better Future*, pages 111–139. Springer, Cham.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions.

AS CohenMiller and P Elizabeth Pate. 2019. A model for developing interdisciplinary research theoretical frameworks. *Qualitative Report*, 24(6).

Kevin Crowston, Eileen E Allen, and Robert Heckman. 2012. Using natural language processing technology for qualitative data analysis. *International Journal of Social Research Methodology*, 15(6):523–543.

Debopam Das, Manfred Stede, and Maite Taboada. 2017. The good, the bad, and the disagreement: Complex ground truth in rhetorical structure analysis. In *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, pages 11–19.

Antonella De Angeli and Daniela Petrelli. 2000. Bridging the gap between nlp and hci: A new synergy in the name of the user. In *Proceedings of the CHI 2000 Workshop on Natural Language Interfaces*, volume 4.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad Twitter Corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179.

Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 chi conference on human factors in computing systems*, pages 278–288.

Abdallah El Ali, Tim C Stratmann, Souneil Park, Johannes Schöning, Wilko Heuten, and Susanne CJ Boll. 2018. Measuring, understanding, and classifying news media sympathy on twitter after crisis events. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Lisa Ferro, L. Gerber, Inderjeet Mani, Beth Sundheim, and Geroge Wilson. 2005. TIDES 2005 standard for the annotation of temporal expressions. Technical report, Mitre Corporation.

Jonas Frich, Nanna Inie, Kim Halskov, and Peter Dalsgaard. 2018. A sentiment analysis of design reflections from design projects. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Graham R Gibbs. 2007. Thematic coding and categorizing. *Analyzing qualitative data*, 703:38–56.

Spence Green, Jeffrey Heer, and Christopher D Manning. 2015. Natural language translation at the intersection of ai and hci. *Communications of the ACM*, 58(9):46–53.

Tad Hirsch, Christina Soma, Kritzia Merced, Patty Kuo, Aaron Dembe, Derek D Caperton, David C Atkins, and Zac E Imel. 2018. " it's hard to argue with a computer" investigating psychotherapists' attitudes towards automated evaluation. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pages 559–571.

Graeme Hirst. 1991. Does conversation analysis have a role in computational linguistics? *Computational Linguistics*, 17(2):211–227.

Victor Hung. 2014. Context and NLP. In *Context in Computing*, pages 143–154. Springer.

Nanna Inie, Jeanette Falk Olesen, and Leon Derczynski. 2020. The Rumour Mill: Making the spread of misinformation explicit and tangible. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–4.

Nikiforos Karamanis, Anne Schneider, Ielka Van Der Sluis, Stephan Schlogl, Gavin Doherty, and Saturnino Luz. 2009. Do hci and nlp interact? In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 4333–4338. ACM.

Andrew J Ko and Brad A Myers. 2005. A framework and methodology for studying the causes of software errors in programming systems. *Journal of Visual Languages & Computing*, 16(1-2):41–84.

Nikhil Krishnaswamy and James Pustejovsky. 2016. Voxsim: A visual platform for modeling motion language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 54–58.

Andrés Lucero. 2015. Using affinity diagrams to evaluate interactive prototypes. In *IFIP Conference on Human-Computer Interaction*, pages 231–248. Springer.

Neil Maiden, Konstantinos Zachos, Amanda Brown, George Brock, Lars Nyre, Aleksander Nygård Tonheim, Dimitris Apsotolou, and Jeremy Evans. 2018. Making the news: Digital creativity support for journalists. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11.

Jimmie Manning. 2017. In vivo coding. *The international encyclopedia of communication research methods*, pages 1–2.

Megh Marathe and Kentaro Toyama. 2018. Semi-automated coding for qualitative research: A user-centered inquiry and initial prototypes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, Thomas Heitz, Leon Derczynski, et al. 2012. Developing Language Processing Components with GATE Version 7 (a User Guide).

Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23.

Christopher D Mellinger. 2017. Translators and machine translation: knowledge and skills gaps in translator pedagogy. *The Interpreter and Translator Trainer*, 11(4):280–293.

Michael Muller. 2014. Curiosity, creativity, and surprise as analytic tools: Grounded theory method. In *Ways of Knowing in HCI*, pages 25–48. Springer.

Cosmin Munteanu, Matt Jones, Sharon Oviatt, Stephen Brewster, Gerald Penn, Steve Whittaker, Nitendra Rajput, and Amit Nanavati. 2013. We need to talk: HCI and the delicate topic of spoken language interaction. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 2459–2464. ACM.

Michael A. Norman and Peter J. Thomas. 1991. Informing hci design through conversation analysis. *International journal of man-machine studies*, 35(2):235–250.

Judith Reitman Olson and Gary M Olson. 1995. The growth of cognitive modeling in human-computer interaction since goms. In *Readings in Human–Computer Interaction*, pages 603–625. Elsevier.

Judith S Olson and Wendy A Kellogg. 2014. *Ways of Knowing in HCI*, volume 2. Springer.

Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Nadine Ozkan and Cécile Paris. 2002. Cross-fertilization between human computer interaction and natural language processing: Why and how. *International Journal of Speech Technology*, 5(2):135–146.

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.

James Pustejovsky and Nikhil Krishnaswamy. 2020. Embodied human-computer interactions through situated grounding. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–3.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly Media, Inc.

Dennis Reidsma and Jean Carletta. 2008. Reliability measurement without limits. *Computational Linguistics*, 34(3):319–326.

Rubin. 2002. Listening post (artwork). https://en.wikipedia.org/wiki/Listening_Post_(artwork).

Ingo Siegert, Ronald Böck, and Andreas Wendemuth. 2014. Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements. *Journal on Multimodal User Interfaces*, 8(1):17–28.

Anders Søgaard, Barbara Plank, and Dirk Hovy. 2014. Selection bias, label bias, and bias in ground truth. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts*, pages 11–13.

Pia Sommerauer, Antske Fokkens, and Piek Vossen. 2020. Would you describe a leopard as yellow? evaluating crowd-annotations with justified and informative disagreement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4798–4809.

Anselm Strauss and Juliet M Corbin. 1997. *Grounded theory in practice*. Sage.

Lucy A Suchman. 1987. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.

Peter Tolmie, Rob Procter, David William Randall, Mark Rouncefield, Christian Burger, Geraldine Wong Sak Hoi, Arkaitz Zubiaga, and Maria Liakata. 2017. Supporting the use of user generated content in journalistic practice. In *Proceedings of the 2017 chi conference on human factors in computing systems*, pages 3632–3644.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS ONE*, 15(12):e0243300.

Tricia Wang. 2013. Big data needs thick data. *Ethnography Matters*, 13.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the EMNLP demos*, pages 38–45.

Qian Yang. 2017. The role of design in creating machine-learning-enhanced user experience. In *2017 AAAI Spring Symposium Series*.

John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 493–502.