Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, pages 73–82
Marseille, 20 June 2022

Beyond Explanation: A Case for Exploratory Text Visualizations of Non-Aggregated, Annotated Datasets

Lucy Havens[†] **Benjamin Bach**[†] **Melissa Terras** [‡] **Beatrice Alex**[†§]
[†]School of Informatics, [‡]College of Arts, Humanities and Social Sciences,
[§]Edinburgh Futures Institute, [§]School of Literatures, Languages and Cultures
University of Edinburgh
lucy.havens@ed.ac.uk, bbach@inf.ed.ac.uk, m.terras@ed.ac.uk, balex@ed.ac.uk

Abstract

This paper presents an overview of text visualization techniques relevant for data perspectivism, aiming to facilitate analysis of annotated datasets for the datasets' creators and stakeholders. Data perspectivism advocates for publishing non-aggregated, annotated text data, recognizing that for highly subjective tasks, such as bias detection and hate speech detection, disagreements among annotators may indicate conflicting yet equally valid interpretations of a text. While the publication of non-aggregated, annotated data makes different interpretations of text corpora available, barriers still exist to investigating patterns and outliers in annotations of the text. Techniques from text visualization can overcome these barriers, facilitating intuitive data analysis for NLP researchers and practitioners, as well as stakeholders in NLP systems, who may not have data science or computing skills. In this paper we discuss challenges with current dataset creation practices and annotation platforms, followed by a discussion of text visualization techniques that enable open-ended, multi-faceted, and iterative analysis of annotated data.

Keywords: annotation, dataset, visualization, data analysis, exploratory search, inter-annotator agreement, perspectivism

1. Introduction

In response to growing evidence of biases in machine learning models, such as classification (Dinan et al., 2020; Diaz et al., 2018), topic modeling (Morstatter et al., 2018), N-grams (Nobata et al., 2016), coreference resolution (Rudinger et al., 2018), machine translation (Nekoto et al., 2020), word embeddings (Caliskan et al., 2017; Bolukbasi et al., 2016), search engines and information retrieval (Noble, 2018; Sweeney, 2013), and computer vision (Prabhu and Birhane, 2021), efforts to uncover the source of such biases have found biased training datasets to be a contributing factor (Prabhu and Birhane, 2021; Cao and Daumé III, 2020; Perez, 2019). The machine learning community has moved towards ever-larger datasets, based on the assumption that more data means more representative datasets (Frické, 2015). In reality, the bigger the dataset, the more difficult it is to ensure the data do not contain harmful representations of people (Bender et al., 2021; Prabhu and Birhane, 2021). The size of a dataset does not correlate to how representative its data are because data must be collected with instruments, and instruments are imperfect (Bender et al., 2021; Welty et al., 2019; Frické, 2015). Choices made regarding what data to collect and how to collect them influence how well a dataset represents the population it is meant to represent (D'Ignazio and Klein, 2020; Perez, 2019; Frické, 2015). Hutchinson et al. (2021), Jo and Gebru (2020), Bender and Friedman (2018), and Gebru et al. (2018) encourage new documentation practices to contextualize datasets and facilitate critical reflection on the implications of their use in models. Documentation alone, however, cannot mitigate datasets' biases and resulting harms.

While documentation of a dataset provides valuable contextual information about why data were collected, how the data are structured, and what the intended use of the data are, documentation cannot replace analysis for understanding the perspectives represented in a dataset. Methods for studying which perspectives are and are not included in a dataset have yet to be established. While methods such as jury learning (Gordon et al., 2022) and perspective-aware modeling (Akhtar et al., 2021) aim to incorporate more than one annotator's perspective in model development, they can only incorporate perspectives that have been represented in the annotation process.

For communities of people not involved in a dataset's creation or annotation, existing approaches to creating and analyzing datasets continue to exclude their perspectives. In this paper, we encourage collaboration across the natural language processing (NLP) and text visualization communities to diversify the perspectives considered during dataset creation. Building on data perspectivism, which advocates for the publication of non-aggregated, annotated datasets (Basile, 2022; Basile et al., 2021), we propose exploratory text visualization techniques as a method for analyzing the different perspectives represented in *and missing from* annotated data.

Though existing text annotation platforms incorporate data visualizations, these platforms assume the aim of the annotation process will be to reconcile disagreements to create a single version of a dataset, or gold standard. This paper presents exploratory text visualization techniques as a complement to data perspectivism, aiming to improve the quality of datasets for model training through analysis of perspectives that are and are not represented in annotations. We begin by defining three key terms used throughout this paper (§2). Next, we summarize current practices for creating annotated datasets and their associated challenges (§3). We then present techniques from the text visualization community that can address these challenges and advance data perspectivism in NLP (§5). Lastly, we conclude with a summary of the paper and envisioned future work for analysis of non-aggregated, annotated text corpora (§6).

2. Definitions

Data Perspectivism We use *data perspectivism* as Basile et al. (2021) define the term: "the adoption of methods that integrate opinions and perspectives of human subjects involved in the knowledge representation step of ML [machine learning] processes" (1). The Perspectivist Data Manifesto expands on this definition with action points for executing data perspectivism in NLP research: publishing non-aggregated, annotated datasets and avoiding training models on aggregated annotated datasets, often referred to as gold standards (Basile, 2022).

Exploratory Search Drawing on information retrieval and human-computer interaction literature, we use the term *exploratory search* to refer to an information seeking process in which the information seeker's task cannot be reduced to a single question and answer. Exploratory search is distinct from look up or querying search tasks (Athukorala et al., 2015; Marchionini, 2006). During exploratory search, the information seeker refines their questions as they become more familiar with a topic. The answer to an initial question often reveals new questions for the seeker to research. Information retrieval and human-computer interaction literature characterizes exploratory search as multi-faceted, iterative, and open-ended (White and Roth, 2009), involving mental processes of synthesis and evaluation to learn something new (Athukorala et al., 2015; Marchionini, 2006).

Stakeholders We refer to *stakeholders* of datasets, and by extension machine learning models trained on those datasets, as people who influence or are influenced by the datasets. Drawing on the definition of stakeholders in NLP research from Havens et al. (2020), we include "(1) the researcher(s), (2) producers of the data, (3) institutions providing access to the data, (4) people represented in the data, and (5) people who use the data" (110) in our use of the term. Furthermore, as Bender and Friedman (2018) note, we emphasize that a dataset's stakeholders may or may not directly interact with the dataset; people may be influenced by a dataset even if they did not participate in its creation. Stakeholders may experience these influences positively, if they are given power or advantage over others, or negatively, if they are oppressed or discriminated against (D'Ignazio and Klein, 2020).[1] For extensive discussion of how stakeholders experience positive and negative impacts from data, please refer to the books by Perez (2019), Noble (2018), and O'Neil (2017).

3. Related Work

Existing annotation platforms assume the aim of the annotation process will be reconciling disagreements to create a single version of a dataset, or gold standard. Many annotation platforms focus on supporting the actual annotation work: loading a text corpus, applying labels, and adding notes explaining the labels (Pérez-Pérez et al., 2015). Among the platforms that allow for annotation workflow management more broadly, such as GATE Teamware (Bontcheva et al., 2013), Argo (Batista-Navarro et al., 2016), and Marky (Pérez-Pérez et al., 2015), the focus is on the management of multiple annotators, facilitating annotation corrections and reconciliation. The underlying assumption of these platforms is that annotator disagreement should be minimized and one version of a dataset will be created. These platforms thus have limited support for a perspectivist approach, where researchers investigate annotators' disagreements and publish non-aggregated, annotated data.

Though existing annotation platforms do provide helpful data visualizations, the visualizations are explanatory rather than exploratory. For example, GATE Teamware uses a flow diagram for visualizing the annotation workflow and a pie chart for visualizing annotation progress (Bontcheva et al., 2013), and Marky uses a bar chart to visualize F scores across rounds of annotation (Pérez-Pérez et al., 2015). As explanatory visualizations, these diagrams and charts are effective in their aim of explaining the annotation workflow, process, and agreement measures. However, they cannot facilitate analysis of patterns and outliers in annotators' distinct approaches to labeling text. Such analysis requires navigation between overviews and detailed views of annotated text corpora, and comparative views of different annotators' labels to the same text. Existing annotation platforms provide no way to study the inconsistencies in an annotator's labels, which could indicate uncertainty in the text that cannot be represented with a single label; nor do they provide a way to study outliers in annotators' labels, which could indicate perspectives that are underrepresented in the data. Instead, existing annotation platforms support practices that minimize inconsistencies and outliers.

For tasks that yield high variability among annotators, there is value in maintaining annotators' disagreeing labels (Davani et al., 2022; Sang and Stanton, 2022; Basile et al., 2021). Basile et al. (2021) propose "data perspectivism" as particularly valuable for these annotation tasks, such as detecting hate speech (Sang and Stanton, 2022), social biases (Sap et al., 2020), or gender biases (Havens et al., 2022). Data perspectivism incorporates multiple perspectives in datasets intended for model training, keeping all annotators' representations of knowledge through the publication of non-aggregated versions of the annotated text data. Representing multiple perspectives in data is important because interpretation of language changes across contexts, such as different geographic locations and cultures (Sambasivan et al., 2021), racialized ethnicities (Crenshaw, 1991), domains (Basta et al., 2020), time periods (Shopland, 2020; Spencer, 2000), and people (Denton et al., 2021).

Data perspectivism aligns with "data feminism" (D'Ignazio and Klein, 2020) which views data as situated and partial. Data feminism draws on feminist theories' rejection of universal knowledge in favor of multiple, different, and equally valid perspectives (Harding, 1995; Haraway, 1988). The process of labeling text and documents, whether

[1]For example, Sweeney (2013) demonstrated how Google Ads discriminated against people whose names are predominant in black communities relative to names predominant in white communities in the United States. This positively impacts job applicants with stereotypically white names and negatively impacts job applicants with stereotypically black names.

with human annotators or classification models, inevitably records, and thus gives power to, particular people's perspectives while excluding others' perspectives (D'Ignazio and Klein, 2020; Bowker and Star, 1999). Though writing for the information sciences, the caution of Bowker and Star (1999) remains relevant to NLP dataset creation and model development: "each category valorizes some point of view and silences another. This is not inherently a bad thing–indeed it is inescapable. But it *is* an ethical choice, and as such it is dangerous" (5). Publishing non-aggregated, annotated data is the first step towards addressing the dangers of classification, making the data available. That being said, the availability of data does not ensure its accessibility (Mons et al., 2017).

4. Why Data Perspectivism Needs Exploratory Visualization

Due to annotated text data's (1) **large size**, (2) **complexity**, and (3) **variability**, publishing non-aggregated, annotated data does not ensure the data's accessibility. Firstly, the amount of annotations and text needed to train NLP models results in annotated datasets of such large size that they cannot be reasonably expected to be manually reviewed. Consequently, analyzing annotated text data requires skills with particular data formats and programming languages, excluding stakeholders in NLP systems who do not have these skills from the data analysis process. Secondly, annotation taxonomies are not standardized. Even for the same task, multiple taxonomies may exist. For example, Dinan et al. (2020) and Hitti et al. (2019) propose two different taxonomies for the same task, classifying gender biased language, that do not have a single category in common.

Thirdly, data formats for annotated text corpora vary. For example, existing annotation platforms may output annotated data as Plaintext (Stenetorp et al., 2012), CSV (Chew et al., 2019), or JSON (Nakayama et al., 2018). The organization of annotated data within these file formats varies according to each annotation platform's design and each project's annotation taxonomy. As a result, anyone wishing to review an annotated dataset must first learn how the annotations and original text are represented in particular file formats. Though Plaintext, CSV, and JSON are not unusual data formats, for stakeholders of a dataset without data science or computing experience, these file format's organization of data may not be intuitive. Moreover, while documentation such as data statements (Bender and Friedman, 2018) provides valuable overviews of annotated datasets, if a person aims to understand the perspectives that different annotations communicate across a text corpus, data analysis remains necessary. To complement the overview that documentation such as data statements provide, we encourage NLP dataset creators to utilize text visualization techniques when publishing non-aggregated, annotated data.

Within the data visualization community, numerous techniques exist to explore text. For a comprehensive survey of text visualizations, please refer to the survey of surveys by Alharbi and Laramee (2019) or, for an interactive exploration, the Text Visualization Browser of Kucher and Kerren (2015). Focusing specifically on opportunities between text visualization and text mining, Liu et al. (2019) survey 4,609 papers and identify classification as underrepresented in text visualization papers relative to text mining papers. The authors note an opportunity in text visualization research to better study how to interactively visualize the complexities of text classification processes. With many publications of visualization techniques repeatedly using the same selection of datasets that do not reflect the complexities of more widely relevant or important datasets (Kosara, 2018), we identify a mutually beneficial collaboration opportunity between the NLP and data visualization communities. This collaboration could lead to "new techniques for AI explainability" (Basile et al., 2021, 2-3), contributing to the NLP and wider machine learning community, while also beginning to address the "ethical challenge in visualization...to visualize the provenance of data and decision-making" (Correll, 2018, 7), contributing to the data visualization community. The next section describes specific techniques from text visualization relevant to analyzing non-aggregated, annotated data.

5. Exploratory Text Visualization Techniques for Data Perspectivism

Recognizing the need to facilitate exploratory search of non-aggregated, annotated data, we see an opportunity for the NLP community to collaborate with the text visualization community. As defined in §2, *exploratory search* refers to an open-ended information seeking process in which the questions guiding an information seeker are multi-faceted, and the answers to those questions are put together iteratively (Athukorala et al., 2015; White and Roth, 2009; Marchionini, 2006). Exploratory search requires collaboration between computers and humans, as an information retrieval system responds to the information seeker's interactions with it, and the seeker refines and tailors their interactions with the system based on the information presented to them (White and Roth, 2009). Interactive data visualization facilitates such collaboration between computers and humans (Hammer et al., 2013; Keim, 2002).

Interactive data visualization, specifically text visualization for text data, provide techniques for visually representing and interrogating non-aggregated, annotated text data:

(1) **Large Size** Exploring data visually makes use of humans' "perceptual abilities" (Keim, 2002). For a person exploring non-aggregated, annotated text datasets, visual design cues such as color, transparency, and position draw on the strength of human vision (Hutmacher, 2019) to communicate patterns and outliers in the annotations when presented at a high level, or zoomed out, overview.

(2) **Complexity** Providing manual interaction mechanisms facilitates self-guided, iterative analysis. For a person exploring non-aggregated, annotated text data, interactions with high- and low-level views of the data would facilitate learning and comprehension of how different annotators interpreted an annotation taxonomy. Notably, this learning and comprehension would be based on the actual application of the taxonomy's labels to text, rather than an abstract representation of text, for example, in vector space (which Goldfarb-Tarrant et al. (2021) have demonstrated the limitations of regarding offensive language in text).

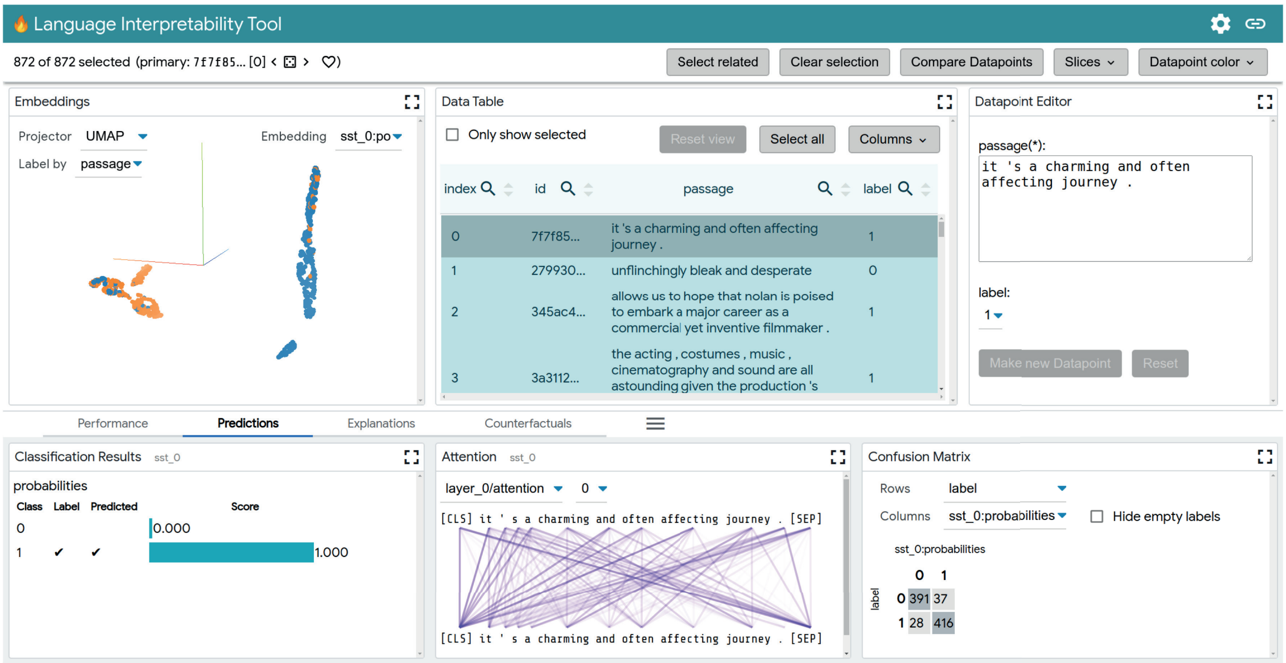

Figure 1: The Language Interpretability Tool by Tenney and Wexler et al. (2020) uses multiple coordinated views for exploratory analysis of a model's performance: "The top half shows a selection toolbar, and, left-to-right: the embedding projector, the data table, and the datapoint editor. Tabs present different modules in the bottom half; the view above shows classifier predictions, an attention visualization, and a confusion matrix." (108). *Figure reproduced with author permission.*

(3) **Variability** Representing annotated text data visually relies on human intuition rather than knowledge of mathematical or statistical concepts, or skills with a particular data format or programming language (Keim, 2002). For a person exploring non-aggregated, annotated text datasets, a text visualization interface would facilitate efficient search and analysis without requiring any prior knowledge or skills in data science or computing. As a result, a more diverse group of stakeholders in an NLP system could participate in the analysis of annotated text corpora.

From among the many text visualization techniques (Cao and Cui, 2016; Puretskiy et al., 2010) that exist, we highlight two techniques particularly relevant to data perspectivism: multiple coordinated views and interconnected terms. Multiple coordinated views combine multiple visual representations of a text corpus at different levels of detail, where interaction with one representation leads to corresponding changes in the other representations (Cao and Cui, 2016). The Language Interpretability Tool (LIT)[2] of Tenney and Wexler et al. (2020), displayed in Figure 1, uses multiple coordinated views to analyze a language model's performance. The visualization supports the three characteristics of exploratory search:

- **Multi-faceted** A person can analyze multiple aspects of a model's performance at multiple levels of detail, including the application of labels in the "Classification Results" view, the attention of the model to specific terms in the "Attention" view, and the data on which to study a model's performance in the "Data Table" and "Datapoint Editor" views.
- **Iterative** A person can iteratively refine their analysis by selecting different subsets of data in the "Data Table" view, or editing datapoints with the "Datapoint Editor" view.
- **Open-Ended** A person is not guided toward a particular answer to a question; rather, a person can ask many questions, each of which can have several answers.

Tenney and Wexler et al. (2020) state that the questions guiding LIT's design was: "What kind of examples does my model perform poorly on? Why did my model make this prediction? And critically, does my model behave consistently if I change things like textual style, verb tense, or pronoun gender?" (107). To answer these questions, people can try numerous approaches, such as applying counterfactual methods to change their data or iterating between analysis tasks at different levels of detail, such as a selected subset of passages in the "Data Table" view or higher-level aggregate views of the model's performance in the "Embeddings" view. The authors' guiding questions are thus exploratory in nature, suitable for an exploratory visualization. For more examples of multiple coordinated views, please refer to the work of Kim et al. (2021), Liu et al. (2015), Isaacs et al. (2014), and Shutt et al. (2009).

Network, or node-link, graphs have been used to visualize interconnected terms. NEREx, displayed in Figure 2, visualizes named entities in a text corpus using a network graph,[3] where nodes represent entities and links represent the distance between the pair of entities they connect (El-

[2] pair-code.github.io/lit/

[3] NEREx also contains five other visualizations.

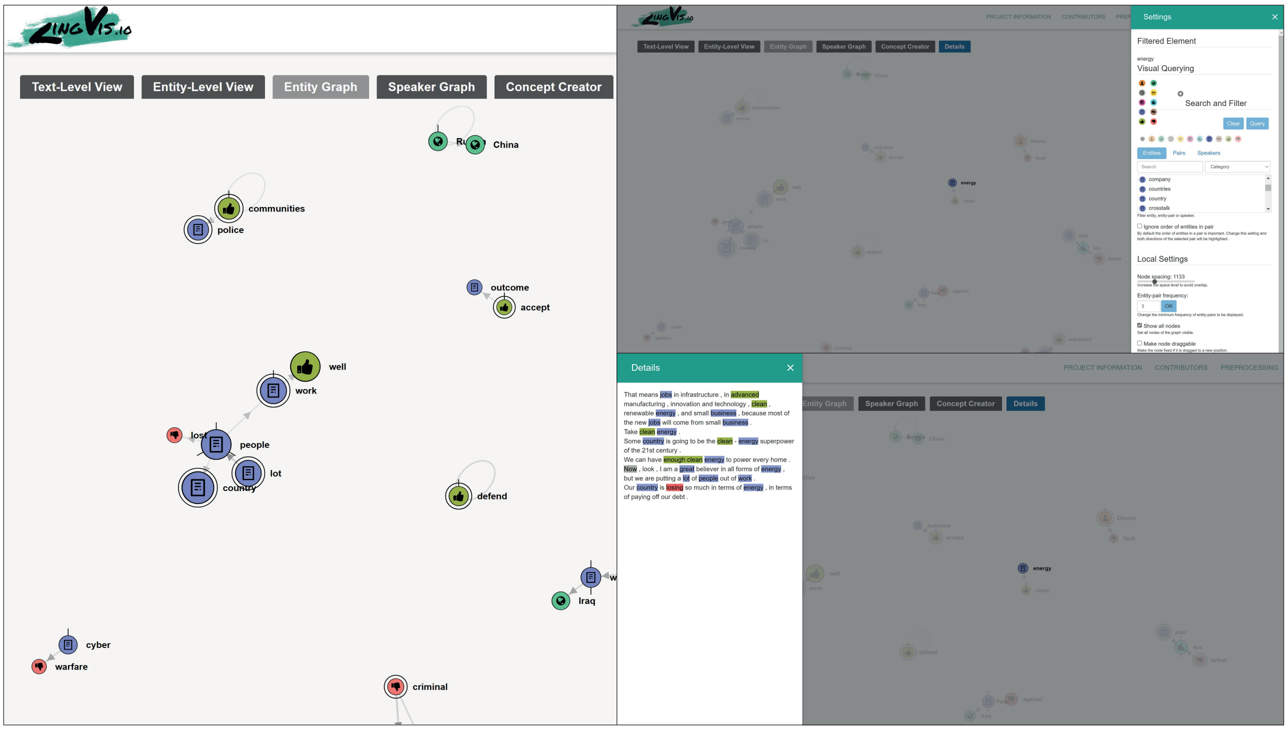

Figure 2: NEREx (El-Assady et al., 2017) uses a network graph, called the "Entity Graph," to provide an overview of text data in a corpus, displaying named identities as nodes and their relationships as links. People can interact with the graph and adjust the data it displays using the "Settings" pane displayed in the top right image. People can view the original text with the "Detail" pane displayed in the bottom left image, which overlays color-coded highlights onto the text to indicate associated words in the text and nodes in the graph. *Figure reproduced with author permission.*

Assady et al., 2017). The network of named entities supports the three characteristics of exploratory search:

- **Multi-faceted** A person can choose to study multiple types of named entities, represented in Figure 2 by color and icon, and study the relationship between the entities, represented by the links between the nodes. Longer links indicate greater distance between the entities as they appear in the text corpus.

- **Iterative** A person can refine their search by filtering the visualization using the "Settings" pane (Figure 2, top right), selecting particular entities, entity pairs, or speakers; or a person can iterate between a detailed view of the text using the "Detail" pane (Figure 2, bottom right) and a distant view of the text as the network graph (Figure 2, left).

- **Open-Ended** A person is not directed toward a particular question and answer. Instead, a person can ask many questions and obtain many answers, such as getting an overview of relationships between named entities, studying the influence of particular people, and analyzing the frequency of and relationships between topics.

For analyzing non-aggregated, annotated text data, connections between terms in network graphs could be based on labels annotators applied to the terms, where a link's length corresponds to the distance between two annotated terms.

Location clouds and lattice graphs also provide approaches to visualizing interconnected terms in text visualization. The Trading Consequences platform of Hinrichs et al. (2015) includes a location cloud (Figure 3) to display relationships between commodities and country names over time. Adapting this visualization to exploring non-aggregated, annotated data, an annotation label could be searched instead of a commodity, and the decade columns could be replaced with columns for each annotator of a corpus, displaying the text spans to which each annotator applied the searched label. The lattice graph proposed for machine translation and automated speech recognition systems (Figure 4) by Collins et al. (2007) provides another example of visualizing interconnected terms. Adapting this visualization to exploring non-aggregated, annotated data, different annotators' labels of a particular sentence or document could be displayed above the sentence or document running along the bottom of the visualization, instead of alternative translations.

Though collaboration with the text visualization community in support of data perspectivism may be new, examples of other interdisciplinary collaborations with the visualization community exist as guides. `Lingvis.io` contains a repository of projects focused on data visualization for linguistics and machine learning. Interdisciplinary work between the humanities and visualization communities demonstrates the value of collaboratively creating visualizations, in addition to using the visualizations for analysis (Hinrichs et al., 2018; Jänicke et al., 2017). That

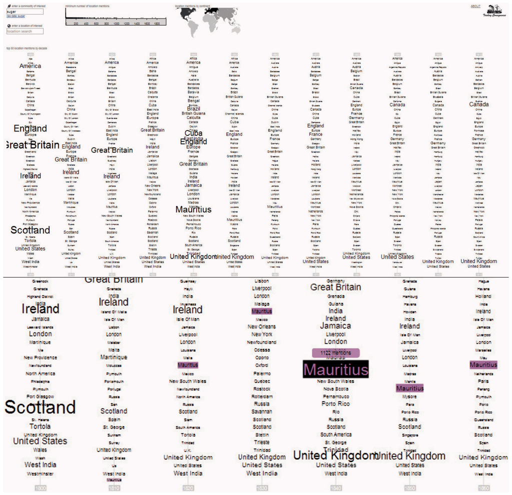

Figure 3: In the location cloud by Hinrichs et al. (2015), country names that appear in the text corpus in relation to a commodity ("sugar" above) are visualized. The size of a country's name indicates the frequency of that country in documents from the corresponding column's decade. When a country name is hovered over ("Mauritius" above), that name is highlighted where it appears in all decades, facilitating easy comparison of how frequently it is mentioned in relation to the searched commodity in the corpus. *Figure reproduced with author permission.*

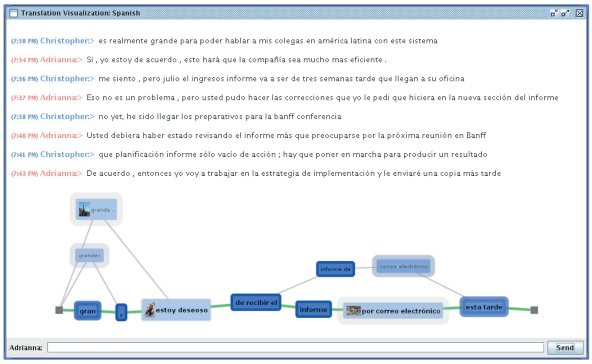

Figure 4: Collins et al. (2007) demonstrate how visualizing a machine translation system's output as a lattice graph facilitates communication between people who speak different languages. The transparency of each word's rectangle and the position of the words correspond to a model's score of its likelihood of being an accurate translation. *Figure reproduced with author permission.*

being said, interdisciplinary collaboration presents challenges due to different vocabularies, working practices, and project timelines across disciplines. Hinrichs et al. (2017) encourage critical reflection on the process of collaboration when undertaking interdisciplinary projects, providing questions that can serve as a guide for such reflection to support effective communication between people in different disciplines. For a broader summary of the benefits and challenges to working across disciplines to collaboratively create data visualizations, please refer to the survey of Jänicke et al. (2017).

6. Conclusion and Future Work

We have described how collaboration between the NLP and visualization communities could facilitate exploratory analysis of non-aggregated, annotated datasets. Exploratory analysis of these datasets would lead to better understandings of the perspectives they represent, improving the transparency of datasets' documentation. Furthermore, by using exploratory analysis to identify perspectives that are not represented in an annotated dataset, along with the perspectives that are represented, dataset creators will be able to determine how to collect additional data that make their dataset more representative of its stakeholders. Due to the underlying motivation of existing annotation platforms (to support the development of one aggregated dataset), the platforms do not provide the exploratory search capabilities necessary for such analysis.

The process of creating a dataset for NLP models inevitably involves curation (Rogers, 2021). We encourage the NLP community's collaboration with the text visualization community to facilitate critical analysis of who and what are included and excluded during dataset creation in support of data perspectivism (Basile et al., 2021), as well as data-centric AI (Press, 2021) and data feminism (D'Ignazio and Klein, 2020). As approaches to incorporating more diverse perspectives in datasets develop, the NLP community could look beyond the text visualization community for collaboration opportunities. Jo and Gebru (2020) recommend looking towards the archival sciences for guidance on data collection and curation. More broadly, the gallery, library, archive, and museum (GLAM) sector has extensive experience creating datasets and enabling their interoperability across systems with metadata standards and supporting infrastructures (RDA Steering Committee, 2022; Library of Congress, 2021; Dunsire and Willer, 2014). Interdisciplinary collaboration would lend value to datasets published under the data perspectivism paradigm, facilitating access to data for stakeholders outside the NLP and wider machine learning communities.

We encourage the development of new platforms with interactive, exploratory text visualizations, in which data analysis becomes an intuitive process relying on human vision, rather than a person's data science or computing skills. Such platforms could lead to new insights about annotations and empower of a more diverse group of stakeholders to participate in data analysis. In future work we will create an exploratory visualization for data published under the data perspectivist paradigm, providing a use case for multi-faceted, iterative, and open-ended analysis of non-aggregated, annotated text data.

7. Acknowledgements

Thank you to the reviewers for their thoughtful comments on this paper. Additional thanks go to the School of Informatics Graduate School and Edinburgh Futures Institute at the University of Edinburgh, and the UK's Engineering and Physical Sciences Research Council for funding our research. We are also grateful to the text visualization creators who permitted us to include images of their work in this paper, Mennatallah El-Assady, Rita Sevastjanova, James Wexler, Uta Hinrichs, and Christopher Collins.

8. Bibliographical References

Akhtar, S., Basile, V., and Patti, V. (2021). Whose Opinions Matter? Perspective-aware Models to Identify Opinions of Hate Speech Victims in Abusive Language Detection. *CoRR*, abs/2106.15896.

Alharbi, M. and Laramee, R. S. (2019). SoS TextVis: An Extended Survey of Surveys on Text Visualization. *Computers*, 8(1).

Athukorala, K., Glowacka, D., Jacucci, G., Oulasvirta, A., and Vreeken, J. (2015). Is exploratory search different? A comparison of information search behavior for exploratory and lookup tasks. *Journal of the Association for Information Science and Technology*, 67(11):2635–2651, October. DOI: 10.1002/asi.23617.

Basile, V., Cabitza, F., Campagner, A., and Fell, M. (2021). Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *CoRR*, abs/2109.04270.

Basile, V. (2022). The Perspectivist Data Manifesto. [Online; accessed March 21, 2022].

Basta, C., Costa-jussà, M. R., and Casas, N. (2020). Extensive Study on the Underlying Gender Bias in Contextualized Word Embeddings. *Neural Computing & Applications*, 33(8):3371–3384.

Bender, E. M. and Friedman, B. (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, December. DOI: 10.1162/tacl_a_00041.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3442188.3445922.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4356–4364. DOI: 10.18653/v1/2020.acl-main.485.

Bowker, G. C. and Star, S. L. (1999). *Sorting Things Out: Classification and Its Consequences*. Inside technology. MIT Press, Cambridge, USA.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April. DOI: 10.1126/science.aal4230.

Cao, N. and Cui, W. (2016). *Overview Text Visualization Techniques*. Atlantis Briefs in Artificial Intelligence ; 1. Atlantis Press, Paris, 1st ed. 2016. edition.

Cao, Y. T. and Daumé III, H. (2020). Toward Gender-Inclusive Coreference Resolution. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. DOI: 10.18653/v1/2020.acl-main.418.

Collins, C., Carpendale, S., and Penn, G. (2007). Visualization of Uncertainty in Lattices to Support Decision-Making. In *Proceedings of Eurographics/IEEE VGTC Symposium on Visualization (EuroVis)*, pages 51–58, Norrköping, Sweden, May. Eurographics. DOI: 10.2312/VisSym/EuroVis07/051-058.

Correll, M. (2018). Ethical Dimensions of Visualization Research. *CoRR*, abs/1811.07271.

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, 43(6):1241–1299. DOI: 10.2307/1229039.

Davani, A. M., Díaz, M., and Prabhakaran, V. (2022). Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Denton, E., Díaz, M., Kivlichan, I., Prabhakaran, V., and Rosen, R. (2021). Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation. *CoRR*, abs/2112.04554.

Diaz, M., Johnson, I., Lazar, A., Piper, A. M., and Gergle, D. (2018). Addressing Age-Related Bias in Sentiment Analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, pages 1–14, Montréal, CA. ACM Press. DOI: 10.1145/3173574.3173986.

D'Ignazio, C. and Klein, L. F. (2020). *Data Feminism*. Strong ideas series. The MIT Press, Cambridge, MA, USA.

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., and Williams, A. (2020). Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online, November. Association for Computational Linguistics.

Dunsire, G. and Willer, M. (2014). The local in the global: universal bibliographic control from the bottom up. In *80th IFLA General Conference And Assembly*, Lyon, FR. International Federation of Library Associations.

El-Assady, M., Sevastjanova, R., Gipp, B., Keim, D., and Collins, C. (2017). NEREx: Named-Entity Relationship Exploration in Multi-Party Conversations. *Computer Graphics Forum*, 36(3):213 – 225.

Frické, M. (2015). Big data and its epistemology. *Journal of the Association for Information Science and Technology*, 66(4):651–661. DOI: 10.1002/asi.23212.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H. M., III, H. D., and Crawford, K. (2018). Datasheets for Datasets. *Computing Research Repository*, arXiv:1803.09010.

Goldfarb-Tarrant, S., Marchant, R., Muñoz Sánchez, R., Pandya, M., and Lopez, A. (2021). Intrinsic Bias Metrics Do Not Correlate with Application Bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Vol-*

ume 1: Long Papers), pages 1926–1940, Online, August. Association for Computational Linguistics. DOI: 10.18653/v1/2021.acl-long.150.

Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J. T., Hashimoto, T., and Bernstein, M. S. (2022). Jury Learning: Integrating Dissenting Voices into Machine Learning Models. *CoRR*, abs/2202.02950.

Hammer, B., Keim, D., Lawrence, N., and Lebanon, G. (2013). Preface: Intelligent interactive data visualization. *Data Mining and Knowledge Discovery*, pages 1–3. DOI: 10.1007/s10618-013-0309-y.

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3):575. DOI: 10.2307/3178066.

Harding, S. (1995). "Strong objectivity": A response to the new objectivity question. *Synthese*, 104(3), September. DOI: 10.1007/BF01064504.

Havens, L., Terras, M., Bach, B., and Alex, B. (2020). Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 107–124, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Havens, L., Terras, M., Bach, B., and Alex, B. (2022). Uncertainty and Inclusivity in Gender Bias Annotation: An Annotation Taxonomy and Annotated Datasets of British English Text. In *Proceedings of the Fourth Workshop on Gender Bias in Natural Language Processing*, Seattle, WA, USA, July. Association for Computational Linguistics. [Forthcoming].

Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E., and Coates, C. M. (2015). Trading Consequences: A Case Study of Combining Text Mining and Visualization to Facilitate Document Exploration. *Digital Scholarship in the Humanities*, 30(Supplement 1):i50–i75, 10. DOI: 10.1093/llc/fqv046.

Hinrichs, U., El-Assady, M., Bradely, A. J., Forlini, S., and Collins, C. (2017). Risk the drift! Stretching disciplinary boundaries through critical collaborations between the humanities and visualization. *Second Workshop on Visualization for the Digital Humanities*.

Hinrichs, U., Forlini, S., and Moynihan, B. (2018). In defense of sandcastles: Research thinking through visualization in digital humanities. *Digital Scholarship in the Humanities*, 34(Supplement 1):i80–i99, 10. DOI: 10.1093/llc/fqy051.

Hitti, Y., Jang, E., Moreno, I., and Pelletier, C. (2019). Proposed Taxonomy for Gender Bias in Text; A Filtering Methodology for the Gender Generalization Subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, IT. Association for Computational Linguistics. DOI: 10.18653/v1/W19-3802.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., and Mitchell, M. (2021). Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560–575, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3442188.3445918.

Hutmacher, F. (2019). Why Is There So Much More Research on Vision Than on Any Other Sensory Modality? *Frontiers in Psychology*, 10(2246). DOI: 10.3389/fpsyg.2019.02246.

Isaacs, E., Damico, K., Ahern, S., Bart, E., and Singhal, M. (2014). Footprints: A Visual Search Tool that Supports Discovery and Coverage Tracking. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1793–1802. DOI: 10.1109/TVCG.2014.2346743.

Jänicke, S., Franzini, G., Cheema, M. F., and Scheuermann, G. (2017). Visual Text Analysis in Digital Humanities. *Computer Graphics Forum*, 36(6):226–250.

Jo, E. S. and Gebru, T. (2020). Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 306–316, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3351095.3372829.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8. DOI: 10.1109/2945.981847.

Kim, C., Lin, X., Collins, C., Taylor, G. W., and Amer, M. R. (2021). Learn, Generate, Rank, Explain: A Case Study of Visual Explanation by Generative Machine Learning. *ACM Trans. Interact. Intell. Syst.*, 11(3–4), August. DOI: 10.1145/3465407.

Kosara, R. (2018). How to Get Excited About Standard Datasets.

Kucher, K. and Kerren, A. (2015). Text visualization techniques: Taxonomy, visual survey, and community insights. In *2015 IEEE Pacific Visualization Symposium (PacificVis)*, pages 117–121. DOI: 10.1109/PACIFICVIS.2015.7156366.

Library of Congress. (2021). Library of Congress Subject Headings PDF Files.

Liu, S., Chen, Y., Wei, H., Yang, J., Zhou, K., and Drucker, S. M. (2015). Exploring Topical Lead-Lag across Corpora. *IEEE Transactions on Knowledge and Data Engineering*, 27(1):115–129. DOI: 10.1109/TKDE.2014.2324581.

Liu, S., Wang, X., Collins, C., Dou, W., Ouyang, F., El-Assady, M., Jiang, L., and Keim, D. A. (2019). Bridging Text Visualization and Mining: A Task-Driven Survey. *IEEE Transactions on Visualization and Computer Graphics*, 25(7):2482–2504. DOI: 10.1109/TVCG.2018.2834341.

Marchionini, G. (2006). Exploratory Search: From Finding to Understanding. *Commun. ACM*, 49(4):41–46, apr. DOI: 10.1145/1121949.1121979.

Mons, B., Neylon, C., Velter, J., Dumontier, M., da Silva Santos, L. O. B., and Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR

Data guiding principles for the European Open Science Cloud. *Information Services & Use*, 37(1):49–56. DOI: 10.3233/ISU-170824.

Morstatter, F., Wu, L., Yavanoglu, U., Corman, S. R., and Liu, H. (2018). Identifying Framing Bias in Online News. *ACM Transactions on Social Computing*, 1(2):1–18, 6. DOI: 10.1145/3204948.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S. H., Kabongo, S., Osei, S., Freshia, S., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Meressa, M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L. J., Tajudeen, K., Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J. T., Abbott, J., Orife, I., Ezeani, I., Dangana, I. A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Murhabazi, E., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C., Dossou, B., Sibanda, B., Bassey, B. I., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., and Bashir, A. (2020). Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, pages 145–153, Montréal, QC, CA. ACM Press. DOI: 10.1145/2872427.2883062.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, New York, USA.

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, New York, NY, USA, first paperback edition. edition.

Perez, C. C. (2019). *Invisible Women: Exposing Data Bias in a World Designed for Men*. Vintage, London, GB.

Prabhu, V. U. and Birhane, A. (2021). Large image datasets: A pyrrhic win for computer vision? *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546.

Press, G. (2021). Andrew Ng Launches A Campaign For Data-Centric AI. *Forbes*.

Puretskiy, A. A., Shutt, G. L., and Berry, M. W., (2010). *Survey of Text Visualization Techniques*, chapter 6, pages 105–127. John Wiley & Sons, Ltd. DOI: 10.1002/9780470689646.ch6.

RDA Steering Committee. (2022). About RDA.

Rogers, A. (2021). Changing the World by Changing the Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2182–2194, Online, August. Association for Computational Linguistics. DOI: 10.18653/v1/2021.acl-long.170.

Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). Gender Bias in Coreference Resolution. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. DOI: 10.18653/v1/N18-2002.

Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., and Prabhakaran, V. (2021). Re-Imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 315–328, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3442188.3445896.

Sang, Y. and Stanton, J. (2022). The Origin and Value of Disagreement Among Data Labelers: A Case Study of Individual Differences in Hate Speech Annotation. In *Information for a Better World: Shaping the Global Future*, Lecture Notes in Computer Science, pages 425–444. Springer International Publishing, Cham.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., and Choi, Y. (2020). Social Bias Frames: Reasoning about Social and Power Implications of Language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.486.

Shopland, N. (2020). *A Practical Guide to Searching LGBTQIA Historical Records*. Taylor & Francis Group, Milton. DOI: 10.4324/9781003006787.

Shutt, G. L., Puretskiy, A. A., and Berry, M. W. (2009). FutureLens: Software for Text Visualization and Tracking. In *Proceedings of the Ninth SIAM International Conference on Data Mining*, Sparks, NV, USA.

Spencer, D. (2000). Language and reality: Who made the world? (1980). In Lucy Burke, et al., editors, *The Routledge language and cultural theory reader*. Routledge, London, UK.

Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, May. DOI: 10.1145/2447976.2447990.

Tenney, I., Wexler, J., Bastings, J., Bolukbasi, T., Coenen, A., Gehrmann, S., Jiang, E., Pushkarna, M., Radebaugh, C., Reif, E., and Yuan, A. (2020). The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, October.

Welty, C., Paritosh, P. K., and Aroyo, L. (2019). Metrology for AI: From Benchmarks to Instruments. *CoRR*, abs/1911.01875.

White, R. W. and Roth, R. A. (2009). *Exploratory search: beyond the query-response paradigm*. Synthesis lectures on information concepts, retrieval, and services ; # 3. Morgan & Claypool Publishers, San Rafael, CA, USA.

9. Language Resource References

Batista-Navarro, R., Carter, J., and Ananiadou, S. (2016). Argo: enabling the development of bespoke workflows and services for disease annotation. *Database*, 2016, 05.

Kalina Bontcheva and Hamish Cunningham and Ian Roberts and Angus Roberts and Valentin Tablan and

Niraj Aswani and Genevieve Gorrell. (2013). *GATE Teamware: a web-based, collaborative text annotation framework*. Springer.

Chew, Rob and Wenger, Michael and Kery, Caroline and Nance, Jason and Richards, Keith and Hadley, Emily and Baumgartner, Peter. (2019). *SMART: An Open Source Data Labeling Platform for Supervised Learning*. JMLR.org.

Hiroki Nakayama and Takahiro Kubo and Junya Kamura and Yasufumi Taniguchi and Xu Liang. (2018). *doccano: Text Annotation Tool for Human*.

Martín Pérez-Pérez and Daniel Glez-Peña and Florentino Fdez-Riverola and Anália Lourenço. (2015). *Marky: A tool supporting annotation consistency in multi-user and iterative document annotation projects*.

Pontus Stenetorp and Sampo Pyysalo and Goran Topić and Tomoko Ohta, Sophia Ananiadou and Jun'ichi Tsujii. (2012). Proceedings of the Demonstrations Session at EACL 2012, v1.3 Crunchy Frog (2012-11-08).