An Inclusive Notion of Text

Ilia Kuznetsov and **Iryna Gurevych**

Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt
`ukp.informatik.tu-darmstadt.de`

Abstract

Natural language processing (NLP) researchers develop models of grammar, meaning and communication based on written text. Due to task and data differences, what is considered text can vary substantially across studies. A conceptual framework for systematically capturing these differences is lacking. We argue that clarity on the notion of text is crucial for reproducible and generalizable NLP. Towards that goal, we propose common terminology to discuss the production and transformation of textual data, and introduce a two-tier taxonomy of linguistic and non-linguistic elements that are available in textual sources and can be used in NLP modeling. We apply this taxonomy to survey existing work that extends the notion of text beyond the conservative language-centered view. We outline key desiderata and challenges of the emerging inclusive approach to text in NLP, and suggest community-level reporting as a crucial next step to consolidate the discussion.

1 Introduction

Text is the core object of analysis in NLP. Annotated textual corpora exemplify NLP tasks and serve for training and evaluation of task-specific models, and massive unlabeled collections of texts enable general language model pre-training. To a large extent, natural language processing today is synonymous to text processing.

But what belongs to text? More broadly, what information should be captured in NLP corpora and be available to the models during training and inference? Despite its central role, the notion of text in NLP is vague: while earlier work mostly focused on grammatical phenomena and implicitly limited text to written language, the applied NLP of the past years increasingly takes an *inclusive* approach to text by introducing non-linguistic elements into the analysis. Extensions vary from incorporating emojis to exploiting document structure and cross-document relationships, and apply to all
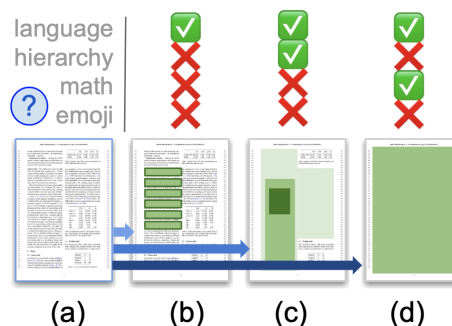

Figure 1: The same textual document (a) can be seen in many ways (b-d) depending on the assumed notion of text: while a syntax researcher might focus on written language (b), a summarization system can use document structure (c), and multimodal applications might use non-linguistic elements like tables (d). Systematically capturing the differences between the assumed notions of text (top) requires a taxonomy of inclusive approaches to text. Such taxonomy is currently lacking.

major components of the modern NLP infrastructure, including unlabeled text collections (Lo et al., 2020), language models (Aghajanyan et al., 2021) and annotated corpora (Kuznetsov et al., 2022). The assumption that text in NLP solely refers to written language no longer holds. Yet, as Figure 1 illustrates, a systematic approach to capturing the differences between the assumed notions of text is lacking.

This is problematic for several reasons. From the **reproducibility** perspective, machine learning assumes similarity between the source and the target distribution – yet lack of consensus on the notion of text might result in undocumented change of the input representation and degraded performance, even if other common variables like domain, language and task remain unchanged. From the **modeling** perspective, the notion of text has major influence on task and model design, as it both determines the tasks NLP aims to tackle, and implies what information should be used to perform those tasks. The final argument for studying the notion of text in

NLP is **conceptual**: the capabilities of strong pre-trained Transformer models (Rogers et al., 2020) and general-purpose NLP frameworks (Gardner et al., 2018; Akbik et al., 2019; Wolf et al., 2020) have led to an explosive growth in NLP beyond traditional, core tasks. The exposure to rich source document types like scientific articles (Lo et al., 2020) and slides (Shirani et al., 2021) and the growing influence of multimodal processing (Xu et al., 2022) motivate the use of additional signals beyond written language in NLP. This leads to a general question on the scope of the field: if written language is no longer the sole object of study, what is, and how can it be formally delineated?

Any empirical discipline relies on *operationalization*, which casts observed phenomena into abstractions, allowing us to formulate claims and perform measurements to evaluate these claims. For example, operationalizing sentiment (phenomenon) as a binary variable (abstraction) allows us to a build a claim (*"this review is positive"*) to be evaluated against the ground truth (review rating), and dictates the downstream NLP task design (binary classification). While widely used, this operationalization is limited: alternative notions of sentiment allow making more nuanced claims, fine-grained measurements and precise models.

The same logic applies to text, which affords a wide range of operationalizations, from a character stream (Akbik et al., 2019) to a rich multimodal graph (Kuznetsov et al., 2022). Yet, the typology for describing text use in NLP is lacking. While concurrent proposals address other key properties of NLP models and corpora (Gavrilidou et al., 2012; Gebru et al., 2018; Bender and Friedman, 2018; Mitchell et al., 2019) like domain, language, demographics, modality and licensing – we lack common terminology and reporting schemata for documenting and formally discussing the assumed notion of text. The growth of the field and the high cost of the retrospective documentation underline the urgent need for a lightweight, semi-structured reporting mechanism to account for text use. To address this need, we contribute the following:

- A common terminology of text use in NLP (Section 2);
- A taxonomy of text extensions beyond the language-focused approach to text (Section 4), based on commonly used sources of NLP data and the current state of the art;
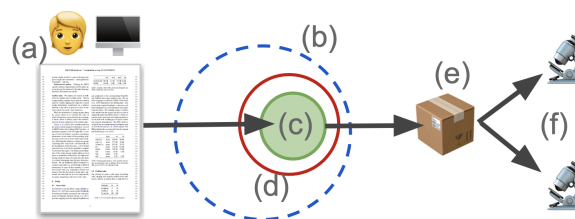- Discussion of the challenges brought by the

Figure 2: A text is produced in an environment (a) and becomes part of the document space (b) that is sampled (c), often based on source (d). The sample is transformed into NLP artifacts (e) that are potentially reused and further refined across multiple studies (f) to produce further artifacts, etc. This process determines the notion of text assumed by the downstream NLP research and the capabilities of the resulting artifacts.

inclusive approach to text (Section 5);
- A new lightweight semi-structured schema for reporting text use in NLP (Section 6).

The notion of text is central to NLP, and we expect our discussion to be broadly relevant, with particular merit for the documentation policy, NLP applications, and basic NLP research. The semi-structured reporting as proposed here is a crucial step towards developing formalized documentation schemata (Gavrilidou et al., 2012) for describing text use and general formats (Hellmann et al., 2013) to encode non-linguistic information in texts. We encourage the community to adopt our reporting schema, and to contribute to the discussion by suggesting new phenomena to be covered by the taxonomy of inclusive approaches to text.

2 Terminology

Textual data available to NLP is a result of multiple processes that determine the composition and properties of texts. To support our discussion, we outline the data journey a typical text undergoes, and introduce common terminology. Figure 2 illustrates our proposed model, and the remainder of this Section provides more details.

Text production. Every text has been produced by a human or an algorithm with a certain communicative purpose. Raw text is rarely exchanged; to avoid ambiguity, we use the term *document* for a unit of information exchange[1]. Documents consist of text along with additional structural and multimodal elements, serialized in a certain *format* and accompanied by metadata. In our definition, tex-

---

[1]There are many other kinds of documents, e.g. images, audio or code; here we focus on "textual" documents.

tual documents cover a broad spectrum ranging from blog posts, Wikipedia articles and Tweets to dialogue turns and search queries. A few widely used formats are plain text, Markdown, PDF.

Document space. All textual documents ever produced make up the abstract *document space*. Document space incorporates both persistent textual documents that are stored (e.g. Wikipedia articles), and transient textual documents that only exist temporarily (e.g. search queries). Despite the apparent abundance of textual documents on the Web, a large (if not major) part of the document space is *not* openly available, or is protected from research use by the copyright, privacy and technical constraints.

Sampling and sources Since capturing the entire document space is not feasible, a *sample* from the subspace of interest is used. Document space can be segmented in a variety of ways, including language, domain or variety (Plank, 2016), creation time, etc. One common way to sample textual documents is based on *source*: documents from the same source often share key characteristics like language variety, text production environment, format and licensing. Some widely used data sources in NLP are Wikipedia, arXiv etc. (Faruqui et al., 2018; Kang et al., 2018).

NLP Artifacts Sampled textual documents are used to create artifacts, including *reference collections* like BooksCorpus (Zhu et al., 2015) and C4 (Raffel et al., 2020), and widely reused general-purpose *language models* like BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020). The notion of text assumed by NLP artifacts is shaped both by the data journey and by the preprocessing decisions during artifact construction. These, in turn, determine how text is operationalized downstream. Due to the differences in how text is produced, sampled and captured, two NLP artifacts might assume very different notions of text. Yet, a framework to systematically capture this difference is lacking.

3 Prior efforts

Our proposal draws inspiration from recent efforts in documenting other common properties of machine learning and NLP artifacts. Model cards (Mitchell et al., 2019) capture core information about machine learning models including technical characteristics, intended and out-of-scope use and preprocessing details. Data sheets (Gebru et al.,

2018) focus on dataset composition, details of the data collection process, preprocessing, distribution and maintenance. In NLP, data statements (Bender and Friedman, 2018) focus on bias mitigation, detailing key aspects of NLP artifact production such as curation strategy, language variety, demographics of speakers and annotators, speech situation, topic and genre. Rogers et al. (2021) propose a formalised checklist documenting risks related to copyright, bias, privacy and confidentiality. Formal proposals are mirrored by community efforts on data repositories like *huggingface datasets* (Lhoest et al., 2021); editorial guidelines[2] encourage the authors to report key parameters of NLP artifacts. Related metadata collection initiatives propose schemata for capturing core information about language resources like language, type, license and provenance (Gavrilidou et al., 2012).

While existing approaches to NLP artifact documentation cover a lot of ground, the requirements for documenting the assumed notion of text remain under-specified. Our work is thus complementary to the prior efforts. Our reporting schema (Section 6) can be seen as specification of the `Speech Situation` and `Text Characteristics` sections of the data statements (Bender and Friedman, 2018), and our taxonomy incorporates some previously proposed documentation dimensions like text creation environment (Gavrilidou et al., 2012) and granularity (Hellmann et al., 2013). Unlike most prior approaches, we deem it desirable to document the assumed notion of text at each step of the NLP data journey, including text production tools, document space samples, as well as NLP models and datasets, with a special focus on widely reused reference corpora and pre-trained language models.

4 Taxonomy of text extensions

4.1 Preliminaries

We derive our proposal in a bottom-up fashion based on two categories of sources. The text production stage is critical as it determines what information is potentially available to downstream processing; to approximate what information *could be used* by NLP artifacts, we (1) conduct an analysis of four representative document sources widely employed in NLP. On the other side of the data journey are the NLP artifacts, the end-product of NLP

---

[2]https://aclrollingreview.org/responsibleNLPresearch/

preprocessing, modeling and annotation. To approximate what information *is being used* by NLP, we outline the de-facto, conservative approach to text and (2) survey recent efforts that deviate from it towards a more inclusive notion of text.

Sources. Wikipedia[3] (`Wiki`) is a collaborative encyclopedia widely used as a data source for task-specific and general-purpose NLP modeling. BBC News[4] (`BBC`) represents newswire, one of the "canonical" domains characterized by carefully edited written discourse. StackOverflow[5] (`Stack`) is a question-answering platform that represents user-generated technical discourse on social media. Finally, ACL Anthology[6] (`ACL`) is a repository of research papers from the ACL community and represents scientific discourse – a widely studied application domain (Bird et al., 2008; Mohammad, 2020; Lauscher et al., 2022). For our analysis we sampled five documents from each of the data sources (Appendix B): for `Wiki`, we selected featured articles from five distinct portals to ensure variety; from `BBC` we selected top five articles of the day[7]; for `Stack` we used five top-rated question-answer threads; for `ACL`, we picked five papers from the proceedings of ACL-2022 available online. Each document was exported as PDF to accurately reproduce the source, and manually annotated for non-linguistic phenomena by the paper authors, with the annotation refined over multiple iterations.

Baseline: Written language. The conservative, de-facto approach to text in NLP is "text as written language": parts of source documents that contribute to grammatical sentences are the primary modeling target, whereas non-grammatical elements are considered noise and potentially discarded. This tradition is persistent throughout the history of NLP, from classic NLP corpora (Marcus et al., 1993; Pradhan and Xue, 2009) and core NLP research, to modern large-scale unlabeled corpora used for model pre-training (Zhu et al., 2015; Merity et al., 2016; Raffel et al., 2020), language models (Devlin et al., 2019; Brown et al., 2020) and benchmarks (Wang et al., 2018). While focus on text as written language is justified for grammatical and formal semantic analysis, for other use cases it proves limiting. In the following Section we sur-

---

vey the emerging inclusive approaches to text that exploit non-linguistic signals to boost the performance and to enable new applications of NLP.

4.2 Taxonomy overview

Table 1 summarizes our proposed two-tier taxonomy for describing the inclusive approaches to text. It demonstrates the wide variety of signals available and potentially relevant to NLP processing beyond the conservative, language-centric view. The following sections discuss the taxonomy classes in greater detail, and Figure 3 gives examples.

4.3 Body

The first high-level class of our taxonomy encompasses the phenomena related to the main, content-bearing parts of the textual document.

A1: Content. Our source analysis reveals that naturally occurring textual documents systematically make use of signal systems beyond written language. The examples of non-linguistic information in textual documents include, but are not limited to, emojis, math, code, hyperlink-, citation- and footnote anchors, tables and multimedia, as well as arbitrary numerical and categorical information like scores and ratings (e.g. on `STACK`). The stance towards such non-linguistic elements of text ultimately determines whether an NLP artifact can represent them in a satisfactory manner, and recent NLP works successfully use non-linguistic elements to their advantage. Applications in sentiment analysis make use of emoji (Felbo et al., 2017); recent research addresses text generation based on tables (Suadaa et al., 2021); Cohan et al. (2019) use citation anchors for citation intent prediction; Shen et al. (2021), Li et al. (2022) and Aghajanyan et al. (2021) integrate layout information into language model pre-training, resulting in improved performance across a wide range of tasks. The ability to handle non-linguistic signals is key for NLP applications and motivates careful documentation of text content.

A2: Decoration. Content is complemented by decoration across all of our sources. Decoration can take the form of font, style, coloring etc. and carries important secondary information, including emphasis, quotation, and signaling Structure (A3). An important function of text decoration is to mark code-switching between different signal systems, from language change to mathematical notation and code, e.g. on `STACK` and `ACL`. Over the
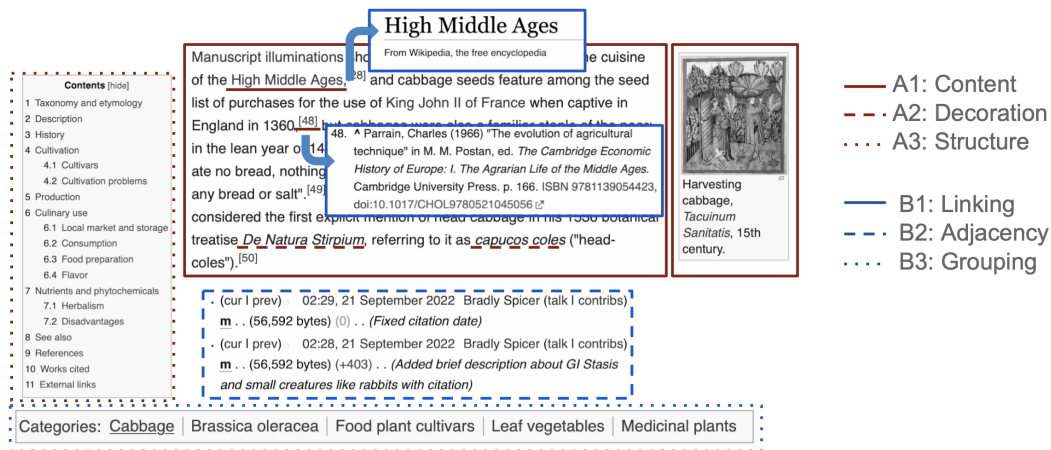
Figure 3: Taxonomy classes in a `WIKI` document. Besides language, content (A1) includes link anchors and images, decoration (A2) marks code-switching to Latin, the article is structured (A3), linked to external sources via hyperlinks and citations (B1), accompanied by an edit history (B2) and contextualized by its Categories (B3).

		examples
A: Body	A1: Content	written language, anchors, math, code, emoji, multimedia
	A2: Decoration	formatting, color
	A3: Structure	document hierarchy, blocks, page and line numbers
B: Context	B1: Linking	implicit links, hyperlinks, citations, footnotes
	B2: Adjacency	comments under post, product and review
	B3: Grouping	tags, document collections, groups

Table 1: Taxonomy of the inclusive notion of text.

past years, decoration received some attention in NLP: Shirani et al. (2019, 2020) explore the task of emphasis modeling in visual media, Shirani et al. (2021) extend it to presentation slides. While humans widely use text decoration, the semantics of decoration are source- and author-dependent and require further systematic investigation.

A3: Structure. Most naturally occurring textual documents are not a flat, linear text as assumed by commonly used reference corpora, from Penn TreeBank (Marcus et al., 1993) to BooksCorpus (Zhu et al., 2015). Instead, the relationships between individual units of content are encoded in document structure. The simplest form of structure is paragraph; longer documents can exhibit a hierarchy of sections; visual proximity is used to include additional content blocks like quotations, definitions, footnotes, or multimedia. In print, textual documents can be organized into pages, columns, lines etc. Explicit document structure is increasingly used in NLP: Cohan et al. (2019) use sections to help citation intent prediction; Ruan et al. (2022) exploit document structure to aid summarization; Sun et al. (2022) use structure to study the capa-

bilities of long-range language models; Kuznetsov et al. (2022) propose Intertextual Graph as a general structure-aware data model for textual documents and use it to support annotation studies and explore how humans use document structure when talking about texts. Document structure is implicitly used in HTML-based pre-training of language models (Aghajanyan et al., 2021), yielding superior performance on a range of tasks, and enabling new pre-training strategies; a separate line of study is dedicated to the analysis of visual document layouts (Shen et al., 2021). The lack of a common approach to formalizing document structure calls for systematic reporting of what structural elements are available in sources, and how document structure is represented and used in NLP.

4.4 Context

The second high-level class of our taxonomy pertains to context. Every text is written and read in the context of other texts, and the ability to capture and use context is a key property of NLP artifacts.

B1: Linking. The first major contextualization mechanism is explicit linking – a marked rela-

tionship between an anchor text and a target text (Kuznetsov et al., 2022). Linking is crucial to many text genres and is found throughout the document sources considered in our analysis. An intra-document link connects two elements within one textual document (e.g. reference to a chapter or footnote), while a cross-document link connects elements in different documents (e.g. hyperlinks and citations). Links differ in granularity of their anchors and targets: the same `Wiki` page can cite its sources on the level of individual sentences (sentence to document) and as a list for further reading (document to document); a research article from `ACL` can reference a particular statement in a cited work (sentence to sentence). A few recent works tap into the narrow context for both task-specific and general-purpose modeling: Bao et al. (2021) investigate the relationships between peer reviewer comments and author rebuttals; Cohan et al. (2020) use information from the citation graph to create better neural representations of scientific documents; Bugert et al. (2021) exploit hyperlinks to generate cross-document event coreference data; Caciularu et al. (2021) show that jointly encoding a document and its near context improves performance on tasks like cross-document coreference resolution and multi-hop question answering; Kuznetsov et al. (2022) and Kennard et al. (2022) jointly model cross-document relations between manuscripts, peer reviews, revisions and author responses. The availability and use of cross-document links are key properties of textual documents and NLP artifacts to be documented.

B2: Adjacency. In addition, textual documents can be related by adjacency; common examples include commentaries attached to the main text, discussion thread replies, copyright notices and prefaces, or peer reviews and the submissions they discuss. Contextualization by adjacency is at play in the NLP study of discussion threads (Jamison and Gurevych, 2013), peer reviews (Gao et al., 2019; Bao et al., 2021; Kennard et al., 2022), etc. Temporal adjacency is a special case where a textual document exists in the context of its previous and future revisions, and is a key feature of document sources like `Wiki`; edit histories have been widely used in NLP as a modeling and annotation target (Zhang et al., 2019; Kuznetsov et al., 2022; Iv et al., 2022; Schick et al., 2022; Spangher et al., 2022). Like linking, adjacency is a rich, naturally occurring type of contextualization.

B3: Grouping Finally, a textual document can be contextualized by the region of the document space that it belongs to: a `Wiki` page exists in the context of other pages belonging to the same portal; a `BBC` article is positioned along the other articles of the same day or topic. Group context both provides the expected common background for text interpretation and sets the standards for the composition of individual documents. Group context plays key role in designing discourse segmentation schemata (Teufel et al., 2009; Hua et al., 2019; Kuznetsov et al., 2022; Kennard et al., 2022), can yield natural labels for text classification, and has been used to augment language models (Caciularu et al., 2021).

4.5 Remarks

Completeness. Our taxonomy serves as the first attempt at capturing the notion of text used in NLP in a structured manner. While we believe that the high-level taxonomy given here is comprehensive, due to our focus on textual documents we do not incorporate further divisions related to multimedia content (e.g. we do not distinguish between images and graphics, although such distinction could be of interest for some applications). As more sources and NLP artifacts are documented, new lower-level taxonomy classes are likely to emerge.

Interactions. The proposed taxonomy dimensions are not orthogonal and do interact: for example, group context (B3) can influence document structure (A2) and decoration standards (A3); in turn, decoration is widely used to signal document structure and linking (B1); the presence of adjacent context (B2) can affect the level of detail in the content (A1). The existence of such inter-dependencies motivates joint documentation and analysis of the different aspects of text even if a conservative notion of text is adopted in the end.

5 Additional considerations

5.1 Interoperability and generalization

A great advantage of the conservative, written-language-only view on text is wide interoperability and generalization: any textual document – from scientific articles to Tweets – can be reduced to written language. This makes it possible to apply a BERT model trained on books to a question-answering prompt and expect non-trivial performance, and enables reuse of text processing frameworks and annotation tools. Yet, such reduction

leads to substantial information loss and bears the danger of confounding due to the interactions between different aspects of text and the text body. While isolated efforts towards inclusive notion of text exist, we are not aware of general approaches that would allow capturing different aspects of text in a systematic manner across domains and document formats. While arriving at a universal, general inclusive notion of text for NLP might not be feasible, we believe that reflecting on the generalization potential of non-linguistic textual elements is the first step in this direction.

5.2 Impact of production environment

Text production environment plays a key role in what information can be captured by the textual document, which, in turn, determines the capabilities of the downstream NLP artifacts. While a sophisticated text editing interface promotes the use of decoration, non-linguistic content, structure and linking, a plain text input field does not. Moreover, the regulating documents and norms that accompany text production have a profound effect on text composition: for example, in addition to common expectations of a scientific publication, `ACL` provides document templates, sets page limits and often enforces obligatory structural elements e.g. reproducibility and limitation sections; `Wiki` is supplied with extensive general and portal-specific guidelines, as well as strict formatting requirements enforced by the community; similar mechanisms are characteristic of most other sources of textual data. Finally, the environment might determine the availability of adjacent and group context during text production. Despite its crucial role, we are not aware of NLP studies that investigate the effect of the production environment on the resulting texts, and believe that our taxonomy can serve as a viable scaffolding for such studies.

5.3 Implications

Efficiency. Computational demands of NLP research are a growing concern (Strubell et al., 2019). It remains unclear how the transition to inclusive treatment of textual documents might affect the efficiency of NLP models. Modeling additional aspects of text might require more parameters and increase the computational demands; yet, the synergies between different aspects of text might allow NLP models to converge faster during training. We are not aware of NLP studies that systematically investigate the effects of inclusive approach to text

on training of NLP models, and believe that this question requires further scrutiny.

Ethics. Recent years are marked by increased attention to the ethics of NLP research, broadly including the issues of privacy, confidentiality, licensing and bias (Bender and Friedman, 2018; Rogers et al., 2021; Dycke et al., 2022). While some types of information beyond written language do not constitute a threat as they are openly accessible in the source textual documents (e.g. textual content A1, decoration A2 and structure A3), others are potentially harmful: precise details of text production might impact privacy, and inclusion of certain contexts (e.g. edit histories, B2) might expose NLP artifacts to false and incomplete information. We are not aware of systematic NLP research into what types of non-linguistic information about textual documents are safe to store and report.

Methodology Current NLP methodology is tailored to a conservative approach to text – from commonly reported dataset statistics (e.g. number of words) to modeling objectives and evaluation metrics. The transition towards an inclusive notion of text calls for a careful revision of the NLP practice. Dataset statistics might include information like the number of figures and tables (A1) or structural information on intra-document (A3) and inter-document (B1-3) level. Pre-trained language models would need to process new types of content, structure and context. Evaluation metrics would need to take into account the new signals. In addition, machine learning models are prone to heuristic behavior (Gururangan et al., 2018) – and besides providing a useful training signal, inclusive notion of text might introduce spurious cues that the models would exploit. Future research must determine the optimal ways to operationalize the inclusive approaches to text in NLP.

6 Reporting

An inclusive approach to text is an emerging trend in NLP that demands systematic study. While preparing this work, it became evident that the lack of systematic reporting limits the meta-analysis of text use in NLP. In line with related documentation efforts, here we propose a simple, semi-structured mechanism for reporting text use. In the short term, such reporting would make it easier to gauge the capabilities of data sources and NLP artifacts, increase community awareness on what aspects of

text are represented and used, and allow aggregation of text use information from different studies. In the long term, it would help the community develop standards for applying the inclusive approach to text and formally documenting text use, and allow informed development of general data models and formats (Hellmann et al., 2013) to facilitate interoperability between NLP artifacts that adopt an inclusive approach to text.

6.1 Schema

As our proposed taxonomy is subject to extension, and to keep the reporting effort low, we formulate the proposed reporting schema as a set of open-ended questions guided by examples in Table 1, in the spirit of short-form data statements by Bender and Friedman (2018). We encourage the reporters to complement it with new categories and phenomena if necessary. For each NLP study that uses or creates textual documents or NLP artifacts, we propose to include the following information into the accompanying publication:

- **Body**: Does the source, format, dataset, model or tool incorporate or use any information apart from written language, including non-linguistic content, decoration and structure?
- **Context**: Does the source, format, dataset, model or tool incorporate or make use of additional context beyond single document, including by linking, adjacency or via group context? If yes, what is it and how is it used?

In addition, for text document sources and interactive NLP models we propose to document the **production environment**: How are the documents produced, including guidelines, software and hardware used? Are the documents single-authored or written collaboratively? How can these factors influence text body and context? Optionally, we invite researchers to reflect upon the implications of their approach to text for **generality**, **efficiency**, **ethics** and **methodology**. Is the newly introduced signal widely used across textual documents? Does it add computational overhead or help reduce computational cost? Can new information lead to bias, privacy risks or promote heuristic behavior? Does the selected methodology take the non-linguistic nature of the new information into account?

6.2 Example and Implementation

To illustrate the intended use of the proposed schema, Appendix A provides example documentation for a textual source (StackOverflow). We note that despite the brevity, short form and potential incompleteness, this kind of documentation is highly informative as it both allows to quickly grasp the notion of text assumed by a data source or artifact, and to aggregate this semi-structured information across different kinds of NLP studies in the future.

Unlike prior efforts that focus on documenting datasets and models separately, our schema applies to all stages of the NLP data journey, from data sources to NLP artifacts, including reference corpora, labeled corpora, preprocessing tools, pre-trained and end-task models and applications. The schema can be incorporated into the data statements and editorial guidelines and used to extend prior metadata documentation proposals (Gavrilidou et al., 2012) and data repository submission forms (Lhoest et al., 2021).

We encourage the community to make use of this low-effort mechanism as a step towards better interoperability of NLP artifacts and the systematic study of the inclusive notion of text. We specifically highlight the need for documenting commonly used sources of textual documents; this will provide the NLP community with a better picture of the document space. We deem it equally important to document pre-trained language models and reference corpora, since their capabilities have a major effect on downstream NLP modeling and applications. This would allow us to gauge how far NLP is from accurately modeling the document space, and will highlight the gaps future work would need to address on the way towards a generally applicable inclusive approach to text.

7 Conclusion

Text plays the central role in NLP as a discipline. But what belongs to text? The rise in applications of NLP to non-linguistic tasks motivates an inclusive approach to text beyond written language. Yet, the progress so far has been limited to isolated research efforts. As NLP ventures into new application areas and tackles new tasks, we deem it crucial to document the notion of text assumed by data sources and NLP artifacts. To this end, we have proposed common terminology and a two-tier taxonomy of inclusive approaches to text, complemented by a widely applicable reporting schema. We hope that our contributions and discussion help the community systematically approach the change of NLP scope towards more accurate modeling of text-based communication and interaction.

Limitations

Our proposed taxonomy is subject to extension, and we expect new phenomena to be included into its scope as the field progresses and as more document sources are considered. Using a taxonomy as an organizational basis for the proposed schema is dictated by our aim to keep the schema simple. The design of future, formalized reporting schemata might adopt an ontology-based approach as it affords more flexibility, and take into account interoperability with the existing proposals in the linked open data community (Hellmann et al., 2013).

While source analysis is only one of our contributions and is thus limited in scope, we have observed that increasing the number of documents from the *same* source yields diminishing value: if a source uses a certain non-linguistic textual element, it does so consistently. This suggests that the future qualitative studies of document sources used in NLP should be conducted in a breadth-first fashion, with few documents samples from many sources, unless quantitative measurement is desired (e.g. *"how often do Wikipedia authors use text formatting"*) or unless a source is known to accommodate a wide variety of document types with different publication and formatting standards.

We do not provide specific details on documenting the text production environment, which represents a promising future research avenue. The study of how the texts in NLP are created is a critical research direction: due to the increased applied use of pre-trained generative language models, documenting the text form and origin is a pressing need.

Our discussion stresses the overall need for more careful handling of terminology in NLP. In this work we chose the term "text" to refer to the object of study in NLP – hence an approach that incorporates non-linguistic elements into text is considered "inclusive". We note that "text" itself is an overloaded term associated with writing on one hand, and text as a format on the other hand; from a cross-disciplinary perspective, e.g. in semiotics, a musical piece or an advertisement would be termed "text" as well. An alternative terminology would use "document" instead of "text" – however, we have opted against this choice, as document can be non-textual (e.g. images, spreadsheets), carries certain implications on length, structure and standalone nature ("document-level NLP"), and comes with its own cross-disciplinary connotations.

As NLP progresses methodologically and interacts with other disciplines, we deem it plausible that a more precise terminology will emerge.

References

Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. 2021. HTLM: Hyper-text pre-training and prompting of language models. *arXiv:2107.06955*.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Jianzhu Bao, Bin Liang, Jingyi Sun, Yice Zhang, Min Yang, and Ruifeng Xu. 2021. Argument pair extraction with mutual guidance and inter-sentence relation graph. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3923–3934, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

---

[8] https://intertext.ukp-lab.de

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Michael Bugert, Nils Reimers, and Iryna Gurevych. 2021. Generalizing cross-document event coreference resolution across multiple corpora. *Computational Linguistics*, 47(3):575–614.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. 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 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. Yes-yes-yes: Proactive data collection for ACL rolling review and beyond. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 300–318, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Manaal Faruqui, Ellie Pavlick, Ian Tenney, and Dipanjan Das. 2018. WikiAtomicEdits: A multilingual corpus of Wikipedia edits for modeling language and discourse. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 305–315, Brussels, Belgium. Association for Computational Linguistics.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. 2019. Does my rebuttal matter? insights from a major NLP conference. 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 1274–1290, Minneapolis, Minnesota. Association for Computational Linguistics.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Maria Gavrilidou, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Haris Papageorgiou, Monica Monachini, Francesca Frontini, Thierry Declerck, Gil Francopoulo, Victoria Arranz, and Valerie Mapelli. 2012. The META-SHARE metadata schema for the description of language resources. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1090–1097, Istanbul, Turkey. European Language Resources Association (ELRA).

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv:1803.09010*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating NLP using Linked Data. In *12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia*, pages 98–113.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. 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 2131–2137, Minneapolis, Minnesota. Association for Computational Linguistics.

Robert Iv, Alexandre Passos, Sameer Singh, and Ming-Wei Chang. 2022. FRUIT: Faithfully reflecting updated information in text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3670–3686, Seattle, United States. Association for Computational Linguistics.

Emily Jamison and Iryna Gurevych. 2013. Headerless, quoteless, but not hopeless? using pairwise email classification to disentangle email threads. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 327–335, Hissar, Bulgaria.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Neha Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. 2022. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1234–1249, Seattle, United States. Association for Computational Linguistics.

Ilia Kuznetsov, Jan Buchmann, Max Eichler, and Iryna Gurevych. 2022. Revise and Resubmit: An Intertextual Model of Text-based Collaboration in Peer Review. *Computational Linguistics*, 48(4):1–38.

Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, Arman Cohan, David Jurgens, and Kyle Lo. 2022. MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1875–1889, Seattle, United States. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junlong Li, Yiheng Xu, Lei Cui, and Furu Wei. 2022. MarkupLM: Pre-training of text and markup language for visually rich document understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6078–6087, Dublin, Ireland. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv:1609.07843*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA. Association for Computing Machinery.

Saif M. Mohammad. 2020. NLP scholar: An interactive visual explorer for natural language processing literature. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 232–255, Online. Association for Computational Linguistics.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. *arXiv:1608.07836*.

Sameer S. Pradhan and Nianwen Xue. 2009. OntoNotes: The 90% solution. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Tutorial Abstracts*, pages 11–12, Boulder, Colorado. Association for Computational Linguistics.

11321

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'just what do you think you're doing, dave?' a checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Qian Ruan, Malte Ostendorff, and Georg Rehm. 2022. HiStruct+: Improving extractive text summarization with hierarchical structure information. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1292–1308, Dublin, Ireland. Association for Computational Linguistics.

Timo Schick, Jane Dwivedi-Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2022. PEER: A collaborative language model. *arXiv:2208.11663*.

Zejiang Shen, Kyle Lo, Lucy Lu Wang, Bailey Kuehl, Daniel S Weld, and Doug Downey. 2021. VILA: Improving structured content extraction from scientific PDFs using visual layout groups. *arXiv:2106.00676*.

Amirreza Shirani, Franck Dernoncourt, Paul Asente, Nedim Lipka, Seokhwan Kim, Jose Echevarria, and Thamar Solorio. 2019. Learning emphasis selection for written text in visual media from crowd-sourced label distributions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1167–1172, Florence, Italy. Association for Computational Linguistics.

Amirreza Shirani, Franck Dernoncourt, Nedim Lipka, Paul Asente, Jose Echevarria, and Thamar Solorio. 2020. SemEval-2020 task 10: Emphasis selection for written text in visual media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1360–1370, Barcelona (online). International Committee for Computational Linguistics.

Amirreza Shirani, Giai Tran, Hieu Trinh, Franck Dernoncourt, Nedim Lipka, Jose Echevarria, Thamar Solorio, and Paul Asente. 2021. PSED: A dataset for selecting emphasis in presentation slides. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4314–4320, Online. Association for Computational Linguistics.

Alexander Spangher, Xiang Ren, Jonathan May, and Nanyun Peng. 2022. NewsEdits: A news article revision dataset and a novel document-level reasoning challenge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 127–157, Seattle, United States. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards table-to-text generation with numerical reasoning. 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 1451–1465, Online. Association for Computational Linguistics.

Simeng Sun, Katherine Thai, and Mohit Iyyer. 2022. ChapterBreak: A challenge dataset for long-range language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3704–3714, Seattle, United States. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Peng Xu, Xiatian Zhu, and David A Clifton. 2022. Multimodal learning with transformers: A survey. *arXiv:2206.06488*.

Xuchao Zhang, Dheeraj Rajagopal, Michael Gamon, Sujay Kumar Jauhar, and ChangTien Lu. 2019. Modeling the relationship between user comments and edits in document revision. 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 5002–5011, Hong Kong, China. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

A Documentation example: text source

StackOverflow hosts three main types of textual documents: questions, answers and commentaries. (A) Body: documents are richly formatted, include multiple content types (text, code, math, images) and decoration (emphasis, code-switching, links). Documents are associated with additional metadata, author and creation/edit time; questions and answers are assigned a rating (number of votes), questions are tagged. Basic structure is present: questions and answers can be logically structured; questions are titled; yet, commentaries are usually short and not structured. (B) Context: linking is used throughout, mostly via hyperlinks, both to the documents on the platform and to external documents; questions, answers and commentaries are related by adjacency; revision histories are available for questions and answers; questions are grouped via tags, and answers and commentaries are grouped by question. Production environment: questions and answers are entered via a UI based on Markdown[9], that supports formatting, structuring, lists, links, code and block inserts, and table formatting. The question submission form additionally includes a title and a tag field. While posting the answer, the user has direct access to the question, previous answers and commentaries. Guidelines for asking and answering questions are available[10] and enforced both by explicit moderation and by the community.

B Source documents

Table 2 summarizes our source analysis. Note that it serves an illustrative purpose and should be used neither as a comprehensive list of non-linguistic phenomena (see Section 4 instead), nor as a comprehensive documentation of the data sources: if substantially more documents were considered, mathematical notation would be eventually found in `STACK`, a `WIKI` article would eventually feature a code snippet, and an eventual `ACL` paper would be accompanied by an adjacent erratum or a peer review. The list below enumerates the documents used in our study, retrieved on October 4th, 2022.

WIKI
- https://en.wikipedia.org/wiki/Euclidean_algorithm

	WIKI	BBC	STACK	ACL
A Body				
A1 Content				
- math	yes	no	no	yes
- code	no	no	yes	yes
- hyperlinks	yes	yes	yes	yes
- citations	yes	no	no	yes
- footnotes	yes	no	no	yes
- images	yes	yes	yes	yes
A2 Decoration				
- font	yes	no	yes	yes
- style	yes	yes	yes	yes
A3 Structure				
- paragraphs	yes	yes	yes	yes
- sections	yes	yes	yes	yes
- blocks	yes	yes	no	yes
- pages	no	no	no	yes
- columns	no	no	no	yes
B Context				
B1 Linking	yes	yes	yes	yes
B2 Adjacency	yes	yes	yes	no
B3 Grouping	yes	yes	yes	yes

Table 2: Non-linguistic elements of text by data source, "yes" – encountered in at least one document from the study sample.

- https://en.wikipedia.org/wiki/Cabbage
- https://en.wikipedia.org/wiki/1689_Boston_revolt
- https://en.wikipedia.org/wiki/Abdication_of_Edward_VIII
- https://en.wikipedia.org/wiki/243_Ida

STACK
- https://stackoverflow.com/questions/477816/which-json-content-type-do-i-use
- https://stackoverflow.com/questions/5767325/how-can-i-remove-a-specific-item-from-an-array
- https://stackoverflow.com/questions/6591213/how-do-i-rename-a-local-git-branch
- https://stackoverflow.com/questions/348170/how-do-i-undo-git-add-before-commit
- https://stackoverflow.com/questions/1642028/what-is-the-ope

rator-in-c

BBC

- https://www.bbc.com/news/busin
  ess-63126558
- https://www.bbc.com/news/world
  -latin-america-63126159
- https://www.bbc.com/news/world
  -europe-63119180
- https://www.bbc.com/news/world
  -australia-63126430
- https://www.bbc.com/news/world
  -asia-india-63127202

ACL

- https://aclanthology.org/2022.
  acl-long.6.pdf
- https://aclanthology.org/2022.
  acl-long.7.pdf
- https://aclanthology.org/2022.
  acl-long.8.pdf
- https://aclanthology.org/2022.
  acl-long.9.pdf
- https://aclanthology.org/2022.
  acl-long.10.pdf

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes, in the Limitations section

☑ A2. Did you discuss any potential risks of your work?
Yes, Implications (5.3)

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.