# UNIT: One Document, Many Revisions, Too Many Edit Intention Taxonomies

**Fangping Lan[1], Abdullah Aljebreen[2], Eduard C. Dragut[1],**
[1]Temple University [2]Shaqra University
fangping.lan@temple.edu, az.aljebreen@su.edu.sa, edragut@temple.edu

## Abstract

Writing is inherently iterative, each revision enhancing information representation. One revision may contain many edits. Examination of the intentions behind edits provides valuable insights into an editor's expertise, the dynamics of collaborative writing, and the evolution of a document. Current research on edit intentions lacks a comprehensive *edit intention taxonomy* (EIT) that spans multiple application domains. As a result, researchers often create new EITs tailored to specific needs, a process that is both time-consuming and costly. To address this gap, we propose UNIT, a UNIfied edit intention Taxonomy that integrates existing EITs encompassing a wide range of edit intentions. We examine the lineage relationship and the construction of 24 EITs. They together have 232 categories across various domains. During the literature survey and integration process, we identify challenges such as one-to-many category matches, incomplete definitions, and varying hierarchical structures. We propose solutions for resolving these issues. Finally, our evaluation shows that our UNIT achieves higher inter-annotator agreement scores compared to existing EITs and is applicable to a large set of application domains[1].

## 1 Introduction

Text is an important medium for sharing information, whether it is reporting news(Chen et al., 2024; Zhijia Chen and Dragut, 2025), documenting scientific discoveries, or narrating stories. Writing is an iterative process that involves revising text to improve how information is conveyed. One revision is created when an editor saves changes to the current document, potentially including multiple local edits (Yang et al., 2017). A single document may undergo multiple revisions. Studying these revisions and the intentions behind them offers significant value in natural language processing tasks (Wang and Dragut, 2024; Zhang et al., 2024, 2025). For instance, analyzing stylistic edits in writing not only helps identify grammatical errors but also provides insight into the structure of arguments and the overall design of the text (Zhang and Litman, 2016; Shah et al., 2019; Spangher et al., 2022). Revisions may be grouped in patterns and the relationship between the patterns can inform about an editor's level of writing experience (Jones, 2008; Yang et al., 2017). In addition, the analysis of intentions enables the creation of assistive tools, such as recommender systems for the collaborative writing (Yang et al., 2016). For lengthy documents such as laws, user agreements, and software manual, quickly identifying revisions and understanding their purpose can reduce reading time, enabling readers to focus on the most relevant sections without having to review the entire document.

Current studies tend to focus on single domains, such as collaborative authoring (Pfeil et al., 2006), informative writing (Spangher et al., 2022; Guo et al., 2022), journalistic writing (Zhang and Litman, 2015), and source code (Lee et al., 2021). Other works study certain aspects of editing revisions, such as style, syntax (Faigley and Witte, 1981), semantics analysis (Yang et al., 2017), and the edits objects analysis (Yang et al., 2016). A pattern observed in the literature is that when researchers need to analyze revision intentions in a new domain, they develop a new EIT either by building it from scratch or adapting an existing one, incurring time and potential financial cost.

We develop a comprehensive EIT that spans multiple application domains and encompasses all observed edit intentions, reducing the burden to create a new EIT for each application. This can be achieved either by constructing the EIT from scratch or by leveraging existing ones. We opt for the latter because jointly existing EITs cover

[1]Our UNIT is publicly released at https://github.com/lanfangping/UniT-EditIntentionTaxonomy



*Findings of the Association for Computational Linguistics: ACL 2025*, pages 23005–23024
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

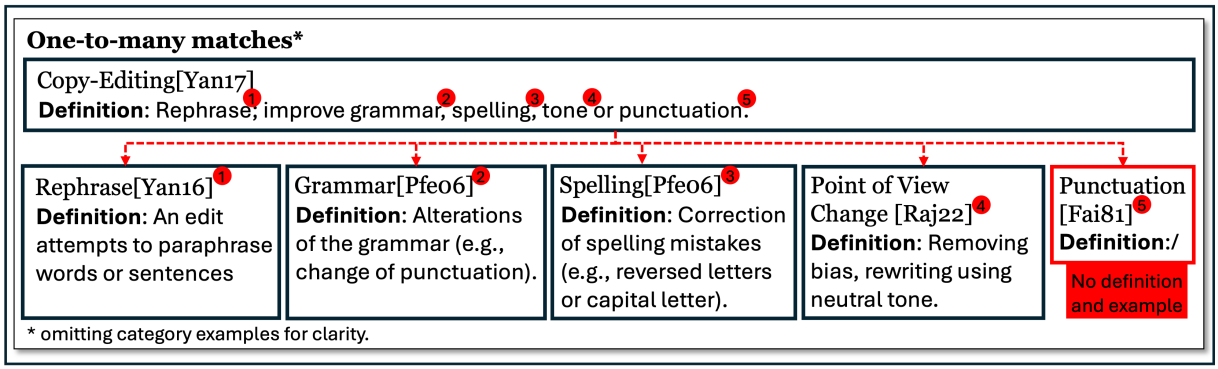



Figure 1: A category in an EIT is overlapping with more categories in other EITs (one-to-many category matches).

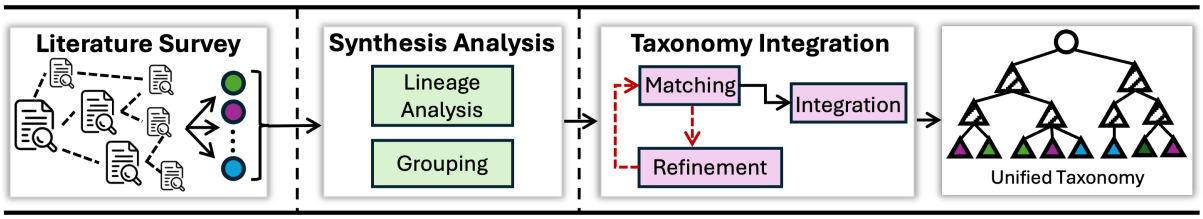



Figure 2: Workflow of EIT Integration

multiple application domains, have a large number of edit (sub)categories, and include many edit intentions. In this paper, we report our work on creating UNIT, a unified and comprehensive EIT, by painstakingly studying and integrating *all*[2] existing EITs. Thus, users aiming to develop a new EIT tailored to their specific task can simply select a subset of edit intentions from our UNIT without the need to search through multiple existing EITs.

### 1.1 Challenges

There are a number of challenges toward the construction of UNIT. First is the literature coverage to find all proposed EITs. Second is the integration procedure. A naive approach is to study EITs pairwise and iteratively merge them (Do and Rahm, 2002; Sundaresan and Hu, 2005; Chiticariu et al., 2007). An alternative approach is to identify groups of related EITs and perform n-way merge within each group (He et al., 2004; Wu et al., 2004; Dragut et al., 2009a,c,b, 2006a, 2022). We apply the n-way merge iteratively for the integrated EITs in each group. Third is edit intention matching. This is challenging because of (1) one-to-many category matches, i.e., a category in an EIT overlapping with two or more categories in another EIT. Figure 1 gives a one-to-many match example, Copy-Editing (Yang et al., 2017) matches Rephrase (Yang et al., 2016), Grammar and Spelling (Pfeil et al., 2006), Point of View Change (Rajagopal et al., 2022) and Punctuation (Faigley and Witte, 1981), if one studies their definitions and revision examples. (2) Incomplete definitions of categories in some EITs, e.g., missing or incomplete definitions and/or revision examples. The category Punctuation (Faigley and Witte, 1981) does not have a definition nor revision examples. We match it to other categories assuming its intended meaning from the name. 58.6% of categories have definitions and at least one revision example, but not all of them have enough revision examples to exemplify their intended edit use cases. (3) EIT hierarchical organization, some EITs are flat while others have multiple levels. Fourth is coping with conflict resolution and missing information (i.e., definition or revision examples) in the integrated EIT. Finally, is the experimental setup. We empirically study if UNIT can be used in multiple domains and if the curation process in constructing UNIT (e.g., improved category definitions and revision examples) leads to better annotations.

### 1.2 Integration Process Overview

Figure 2 depicts our process of building UNIT. We start by searching for publications related to edit intentions. For each relevant paper (i.e., a paper that defines an EIT), we look into its references and the papers that cite it for additional relevant pa-

[2]We acknowledge the possibility of overlooking some relevant works.

pers. We exclude papers that study edit intentions but do not provide an EIT. We iterate until we can not find any relevant papers. We gather 24 papers, summarized in Table 1. Then, we try to group the EITs from the 24 papers in smaller groups. We study the citation network between the papers to find relationships between EITs. That is a useful exercise because EITs in the same group tend to have many one-to-one matches and fewer requirements on solving one-to-many matches. For example, we find EITs that merely adapt existing ones, or authors that refine their EITs over time. This leads to a lineage relationship between subsets of EITs, where an EIT is obtained from another EIT via a number of *edits*, such as adding new nodes, collapsing categories, and renaming categories to denote broader edit intentions. We then integrate the EITs in each group, then integrate the resulting EITs into the final EIT, UNIT. Thus, the integration of EITs is an inherently iterative process. We evaluate our UNIT by evaluating its value to annotators and its versatility across different domains. This is done through the design of a manual annotation task and the measurement of *inter-annotator agreement* (IAA) among the participants. The contributions of this paper are as follows:

- We study 24 EITs and analyze their properties, such as structure, definitions, and revision examples (Section 2). We also collect external resources, i.e., codes and datasets, and report on their reproducibility.
- We study the lineage relationship between EITs. (Section 3.1).
- We build an integrated EIT that spans multiple application domains and includes a large set of edit intentions. (Section 4).
- We compare our UNIT with existing EITs showing that that it achieves higher IAA scores and it is applicable to a larger set of application domains (Section 5).

## 2 Literature Review

This section presents, to our knowledge, all the papers that propose EITs. A criteria for a good EIT is defined as follows: an edit intention taxonomy provides well-structured categories with unambiguous category names, concise definitions, and representative examples. Thus, we organize it along four dimensions: type of the EIT structure, the presence of definitions and revision examples for EIT categories, application domains (covered in Appendix A.1), and availability of EITs (covered in Appendix A.2). Table 1 summarizes our findings.

### 2.1 Structure

EITs are tree structures, ranging from flat (i.e., list of categories) to 3-level deep trees. EITs with multiple levels include (sub)categories for edit actions (e.g., add, merge, delete) in addition to the revision purposes (e.g., Update Content).

We present first the 3-level EITs. Their top layer typically indicates whether a revision affects the meaning of text. Faigley and Witte (1981) and Zhang and Litman (2015) use the terms Surface Changes for meaning preserving edits and Text-Base Changes for meaning changing edits. Yang et al. (2016) adopt simpler labels, Meaning-Preserving and Meaning-Changing. Faigley and Witte (1981) subdivide Text-Base Changes based on whether the text's summary is affected, distinguishing Microstructure Changes (summary unchanged) from Macrostructure Changes (summary altered). Yang et al. (2016) refine their Meaning-Changing category by considering the specific objects being edited (e.g., templates, files). The leaf categories represent edit actions (e.g., addition and permutation of text pieces). The top layer of most 2-level EITs (Jones, 2008; Faltings et al., 2021; Du et al., 2022; Daxenberger and Gurevych, 2012, 2013; Zhang et al., 2016, 2017; Kashefi et al., 2022; Laban et al., 2023; Ruan et al., 2024) is similar to that of 3-level EITs. Some 2-level EITs introduce application-specific categories, such as Wikipedia Specific Intention (Jones, 2008; Daxenberger and Gurevych, 2012, 2013). Jiang et al. (2022) mainly focus on meaning-preserving changes. Ruan et al. (2024) categorize edit actions into basic methods(i.e., addition, deletion, modification) and complex operations(i.e., merges, splits and fusions).

Regarding the flat EITs, we note that some works flatten the hierarchical EITs proposed in previous works, e.g., Zhang and Litman (2016) and Rajagopal et al. (2022) flatten those in Zhang and Litman (2015) and Yang et al. (2017), respectively. Spangher et al. (2022) propose the EIT only on edit actions. Rathjens (1985) give fine-grained categories for tracking edits for the clarity of technical writing. Guo et al. (2022) include categories specific to news headline changes. Faruqui et al. (2018) have categories to characterize the types of insertions. Pfeil et al. (2006), Liu and Ram (2011), and Anthonio et al. (2020) define categories for surface and content changes, though they do not

| alias | paper | domain | #le. | has_definition | | | has_examples | | link? |
|---|---|---|---|---|---|---|---|---|---|
| | | | | l1 | l2 | l3 | extents | from | |
| Fai81 | Faigley and Witte, 1981 | Writings | 3 | Yes | Yes | Partial | Partial | Inside | / |
| Zha15 | Zhang and Litman, 2015 | Writings | 3 | Yes | Yes | No | Full | Inside | / |
| Zha16(A) | Zhang et al., 2016 | Writings | 2 | Yes | Yes | | Full | Outside | Active |
| Zha17 | Zhang et al., 2017 | Writings | 2 | Yes | Yes | / | Full | Both | Active |
| Fal21 | Faltings et al., 2021 | Writings, Wikipedia | 2 | Yes | Yes | / | Partial | Inside | Broken |
| Du22* | Du et al., 2022 | Writings, Wikipedia, News | 2 | Yes | Yes | / | Full | Both | Active |
| Jia22 | Jiang et al., 2022 | Writings | 2 | Yes | Yes | / | Full | Both | Active |
| Kas22 | Kashefi et al., 2022 | Writings | 2 | Yes | Yes | / | Full | Both | Active |
| Rua24* | Ruan et al., 2024 | Writings | 2 | Yes | Yes | / | Full | Both | Active |
| Rat85 | Rathjens, 1985 | Writings | 1 | Yes | / | / | Full | Inside | / |
| Zha16(B) | Zhang and Litman, 2016 | Writings | 1 | Yes | / | / | Partial | Inside | / |
| Yan16 | Yang et al., 2016 | Wikipedia | 3 | Yes | Yes | No | No | / | / |
| Jon08 | Jones, 2008 | Wikipedia | 2 | Yes | No | / | No | / | / |
| Dax12 | Daxenberger and Gurevych, 2012 | Wikipedia | 2 | Yes | Yes | / | Full | Inside | Moved |
| Dax13 | Daxenberger and Gurevych, 2013 | Wikipedia | 2 | Yes | Yes | / | No | / | Broken |
| Yan17 | Yang et al., 2017 | Wikipedia | 2 | Yes | Yes | / | Full | Outside | Active |
| Lab23 | Laban et al., 2023 | Wikipedia | 2 | Yes | Yes | / | Full | Both | Active |
| Pfe06 | Pfeil et al., 2006 | Wikipedia | 1 | Yes | / | / | Full | Inside | / |
| Liu11 | Liu and Ram, 2011 | Wikipedia | 1 | Yes | / | / | No | / | / |
| Far18 | Faruqui et al., 2018 | Wikipedia | 1 | Yes | / | / | Full | Both | |
| Raj22 | Rajagopal et al., 2022 | Wikipedia | 1 | Yes | / | / | Full | Both | Active |
| Ant20 | Anthonio et al., 2020 | WikiHow | 1 | No | / | / | Full | Both | Active |
| Spa22 | Spangher et al., 2022 | News | 1 | Yes | / | / | Full | Both | Active |
| Guo22 | Guo et al., 2022 | News | 1 | Yes | / | / | Full | Inside | Active |

Table 1: It shows the application **domain** of the taxonomy in a **paper**; **#le.** is the number of levels and whether it is flat (**#le.**=1) or hierarchical (**#le.**>1); whether the categories in level # (**l#**) has definitions (**has_definition**); whether 'Full', 'Partial' or 'No' leaf categories (**extents**) have concrete revision examples (**has_examples**) and they are **from** 'Inside' of the paper, 'Outside' of the paper through provided link (**link?**) or 'Both'. '/' denotes not applicable. '*' denotes EIT in this work is compared against UNIT in Section 5.

emphasize the meaning changes within revisions.

### 2.2 Definitions and Revision Examples

The ease of use and adoption of a taxonomy depends largely on the clarity of its category definitions and the inclusion of illustrative examples (Dilnutt, 2004). We survey those properties here.

**Definitions** We observe several patterns about category definitions in EITs. While most include definitions for their categories, some may omit them, particularly when a category has a self-explanatory name, such as Spelling/Grammar, or when suggestive examples are provided (Anthonio et al., 2020). There are cases where categories miss both, leading to ambiguity. For example, Style and Readability (Jones, 2008) may refer to the edits that adjust writing style (making the text more technical or narrative) or to changes in the text appearance (e.g., bold and italic formatting). The quality of the definition is also important. For instance, some definitions are overly broad (e.g., Copy-Editing in Figure 1), while others are misaligned with the category name. For example, Adding Support Evidence implies adding information, but its definition includes edits that involve removing or replacing information, creating confusion.

**Revision Examples** Most EITs provide revision examples for leaf categories but not for the internal node ones. Among those that lack revision examples, we notice two cases 1) offer examples only for a subset of the categories (Faigley and Witte, 1981; Zhang and Litman, 2016; Faltings et al., 2021) and 2) no category has examples (Jones, 2008; Liu and Ram, 2011; Yang et al., 2016). In some instances, the absence of examples is mitigated by offering detailed explanations for the categories (Pfeil et al., 2006; Laban et al., 2023).

Definition and revision examples support category matching, while existing EIT structures and their application domains inform the design of top-layer categories, helping annotators identify edit intentions efficiently. Their released datasets are a useful resource to evaluate EITs, including UNIT, and advancing research in this field.

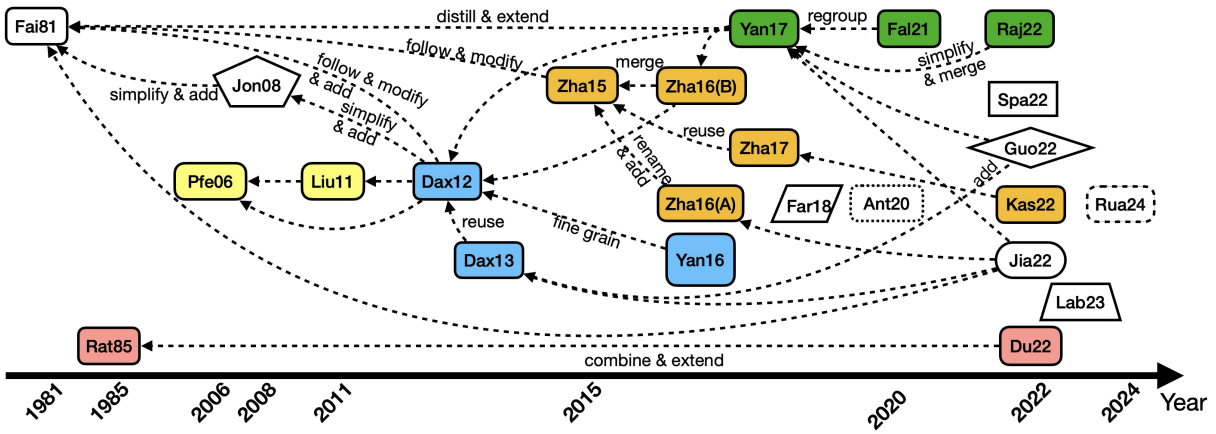



Figure 3: The lineage of the prior EITs. Each block is an EIT in a paper and the dashed arrow lines show the references to other EITs. The link labels indicate how the "child" EIT inherits from the "parent" EITs. The blocks with the same color are grouped based on the lineage analysis, e.g., Pfe06 and Liu11 with yellow color are grouped.

## 3 Grouping EITs based on Lineage

The previous section outlines the contributions of past work on EITs. Here, we examine the relationships between EITs by analyzing the citation network among the reviewed papers, which reveals a lineage network among EITs, e.g., one EIT builds upon another by performing changes, such as category merging or splitting. We describe this network in this section and use it in the next section to group EITs. Then we proceed with integration of EITs by groups. Figure 3 outlines the lineage.

### 3.1 Observations on EIT Lineage

Fai81[3] is the first EIT. Jon08 expanded on it by simplifying the subcategories under Macro/Micro-structure Changes and adding new categories to account for Wikipedia-specific revision behaviors. Dax12 and Zha15 adopted the high-level categories Surface Changes and Text-Base Changes from Fai81 but modified the subcategories to align with their respective domains and targets, focusing on Wikipedia objects (e.g., file, template) and the structural elements of academic writings (e.g., claims/ideas, introductory materials), respectively.

Du22 categorized the edits into Meaning-changed and Non-meaning-changed, focusing on the impact of edits on text meaning. Building on the work of Raj85 and Harris (2017), Du22 defined subcategories for Non-meaning-changed. Liu11, in turn, streamlined the detailed categories in Pfe06, arguing that these categories were not feasible for automatic identification by classifiers. In addition to the "link"-related categories in Pfe06, Liu11 added "references"-related categories, arguing that they are important quality indicators.

Zha16(B) built on Zha15, simplifying it by consolidating all subcategories under Surface into a single Surface category, as the focus was on augmentative changes. Zha16(B) also merged Rebuttal and Warrant categories into Warrant based on the observation in Zha15 that rebuttal edits were rare. Yan16 extended Dax12 by further fine-graining Reference. Yan16 used edit intentions to identify editor roles, while Dax12 studied the collaborative behavior among authors.

Yan17 was an EIT for wiki edits by distilling and extending categories in Fai81, Dax12 and Zha16(B). Yan17 organized the top-level categories based on whether the revisions were general or specific to Wikipedia, with most subcategories being newly introduced. Fal21 reorganized the 13 leaf categories from Yan17 into three groups Fluency, Content, and Other, as they distinguished fluency edits, which affect the grammar or structure of a sentence, from content edits, which change the meaning of a sentence. Raj22 further refined Yan17 by collapsing similar categories and merging all Wikipedia-specific categories into Other category.

Guo22 was built using Yan17 and Dax13 as references but tailored to news headline changes, whereas the source EITs focused on edits in Wikipedia articles. Since many categories from the referenced EITs were unsuitable for news headlines, Guo22 added new categories based on their journalism expertise (Blom and Hansen, 2015; Brown and Sinta, 2016). Ant20 was built on their

[3]For compactness, we use aliases for the papers to reference their EITs. See Table 1 for full reference.

experience of manually inspecting revisions in a sample of 100 sentence pairs.

### 3.2 Grouping

We leverage our observations about the lineage relationship between EITs to organize them into groups, such that EITs in the same group have higher overlap between their categories. This approach simplifies category matching between EITs and the integration process. Appendix A.3, Table 6-10 give the resulted groups.

**Reuse-based Grouping** Dax12 and Dax13 go into the same group since Dax13 is builds on the EIT from Dax12 (Table 7). Fal21 and Yan17 are grouped because they share the same leaf categories despite different top layer structures (Table 9). Zha15, Zha16(A), Zha17, and Kas22 are grouped together because the subsequent works reused the EIT in Zha15, making minor adjustments to category names to fit their specific system requirements (Table 8).

**Merge-based Grouping** We group Yan17 and Raj22 as Raj22 consolidates all Wikipedia-related categories into one category and combines similar categories from Yan17 (Table 9). We group Zha15 and Zha16(B) as Zha16(B) mainly merges some of the categories from Zha15 (Table 8). We group Rat85 and Du22 since Du22 consolidates all categories in Rat85 under one category (Table 6).

**Category Refinement Grouping** We group Pfe06 and Liu11 as well as Dax12 and Yan16 because Liu11 simply refines Reference changes from Pfe06, while Yan16 refines Reference categories from Dax12 (Table 7 and Table 10, respectively).

After grouping, we have 5 groups and 9 individual EITs, reducing the total number of EITs to be integrated from 24 to 14. This reduction of nearly 50% significantly reduces the integration workload.

## 4 Taxonomy Integration

We define the EIT integration task as follows: given a set of EITs, generate a single integrated EIT that includes all the categories from the individual EITs, subject to certain modifications. This section discusses the process of EIT integration, including matching (Appendix A.7 gives partial matching results), merging, as well as adding revision examples (naming and adding definitions are covered in Appendix A.4 and A.5).

### 4.1 Matching

Given a group of EITs, the first step is category matching, which identifies semantically equivalent categories across the EITs. We employ the three sources of information: name, definition, and revision examples. The matching process is challenging because of incomplete definitions/revision examples and one-to-many category matches.

We first conduct matching using the category names, e.g., Copy Editing [Yan17] and copy-editing [Fal21]. Next we employ category definitions, which can help in cases such as categories with different names, but equivalent meaning, and categories with broad edit meanings. For example, in Figure 4, Edit [Spa22] and Sentence Split [Lab23] can only be matched because of their definitions. Finally, when neither name nor definition can help match two categories, we study their accompanying revision examples. For example, we are able to match Information Modification/Insertion [Ant20], for which a definition is missing, with Anaphora Resolution [Lab23] (Figure 4). This is achieved by analyzing an example of Information Modification/Insertion where "them" is edited to "your parents". By combining this example with Anaphora Resolution's definition, which states that changing implicit entity mention with a resolved entity mention, we can reliably match them. Note that the above process needs to account for instances when a category has a definition but no examples (e.g., Disambiguation [Fal21]), or has examples but no definition (e.g., Information Modification/Insertion [Ant20]).

**Exceptions** (i) There are categories that lack definitions and revision examples. During the construction process, if we could not reliably match them, we set them aside to revisit after completing the initial draft of the integrated EIT. However, they did not give new categories. (ii) Certain categories share similar names, but different meanings. For example, Clarification [Yan17] and Clarity [Du22] have similar category names but their definitions have different actions, the former defines "*Specify or explain fact or meaning by example or discussion without adding new information*" and the latter defines "*Make the text more formal, concise, readable and understandable*." (iii) Lastly, there are one-to-many category matches, as illustrated in Figure 1. Establishing such matches requires the presence of definitions and revision examples. These matches often indicate categories with a

broad range of edit revisions. We often find the one-side confusing and opt to keep the many-side in the integrated EIT as children of the one-side.

### 4.2 Merging

After the matching step, semantically equivalent categories from the EITs are organized in semantic clusters. There are 31 clusters organizing the 232 categories from the 24 EITs. A cluster can have up to 24 categories, but the average is 8. We organize the clusters, into two groups, Fine-grained Intentions and Edit Intentions (Figure 5 in Appendix A.8). In general, a cluster becomes a node/category in the UNIT. For example, the cluster of Edit [Spa22] and Sentence Split [Lab23] becomes a node Splitting Sentences in UNIT (we discuss category naming in Appendix A.4).

**One-to-Many** Integration of one-to-many cases requires careful consideration. Consider the example of Refactoring [Yan17] in Figure 4. It matches one-to-one with Refactoring [Fal21] and one-to-many with Paraphrase [Dax12], Refactors [Spa22], and Word-smithing [Raj22] (refer to their definitions in the figure). We handle it by creating a subtree with the root Refactoring and Paraphrase, Refactors, and Word-smithing as its children in UNIT. (More precisely the clusters of these categories form the nodes in UNIT.) Note that Refactoring (renaming to Word-Smithing discussed in Appendix A.4) belongs to Edit Intentions, while those of Paraphrase (renaming to Substituting Words/Phrases), Refactors, and Word-smithing (renaming to Permuting Blocks) belong to Fine-grained Intentions (Figure 5). At the end of the merge process, we have 24 categories in Edit Intentions and 13 in Fine-grained Intentions. Next, we discuss the organization of the edit intentions.

**Top Layer Integration** The top-level categories of UNIT are based on two main criteria: (1) whether an edit changes the meaning of text, and (2) whether the edit is domain-specific. We introduce a Top-layer defined as follows. Edit Intention clusters become children of Non-Meaning-Changed and Meaning-Changed if they denote edits that affect the summary of the text; children of Collaborative Authoring if the application domain is an online collaborative writing system, e.g., Wikipedia or Github; children of Informative/Explanatory Writing if the application domain is academic writing, e.g., ArXiv papers, and software manuals; children of Journalistic/Literary Writing if the application domain is News articles or novels; and finally children of Source Code if the application domain is programming languages, like markup languages that are used for text representation, e.g. HTML and XML.

A challenge in top-layer integration is that certain sub-categories can logically fit under multiple parent categories. For instance, Claim/Ideas and Warrant/Reason/Backing may be inserted as children of either Meaning-Changed or Informative/Explanatory Writing. To ensure clarity, we assign each category to a single parent, even though this may not align with a universal consensus. Users can adjust these assignments as needed.

### 4.3 Adding Revision Examples

Our experience during constructing UNIT and then using it in the user studies (Section 5.1) indicate that illustrative examples of edit revisions are critical for understanding the meaning of a category. Thus, we seek to add revision examples for each category in UNIT. We follow two directions. First, we gather all relevant revision examples from the existing EITs. Second, we propose new revision examples. The latter is needed when the provided revision examples do not fully encompass the entire definition of a category (see Table 11, Appendix A.6) or a category does not have an accompanying revision example. The common practice in related work includes one revision example per category, with some cases featuring two or more. We use LLMs to generate revision examples. Specifically, we prompt the LLM by providing the category name and its definition, thereby facilitating the creation of representative revision examples. Due to the limited number of categories and samples that needed to be investigated, we employed consensus discussion to select representative revision examples.

## 5 Evaluation

To evaluate our UNIT, we ask two questions:

RQ1 Does the use of UNIT result in better annotator agreement compared to other EITs?

RQ2 Does UNIT perform consistently across datasets from varied domains?

### 5.1 Evaluation Setups

**Evaluation Task** We evaluate UNIT via an edit intention classification task, where a given text pair (original and revised) and an EIT serve as input to

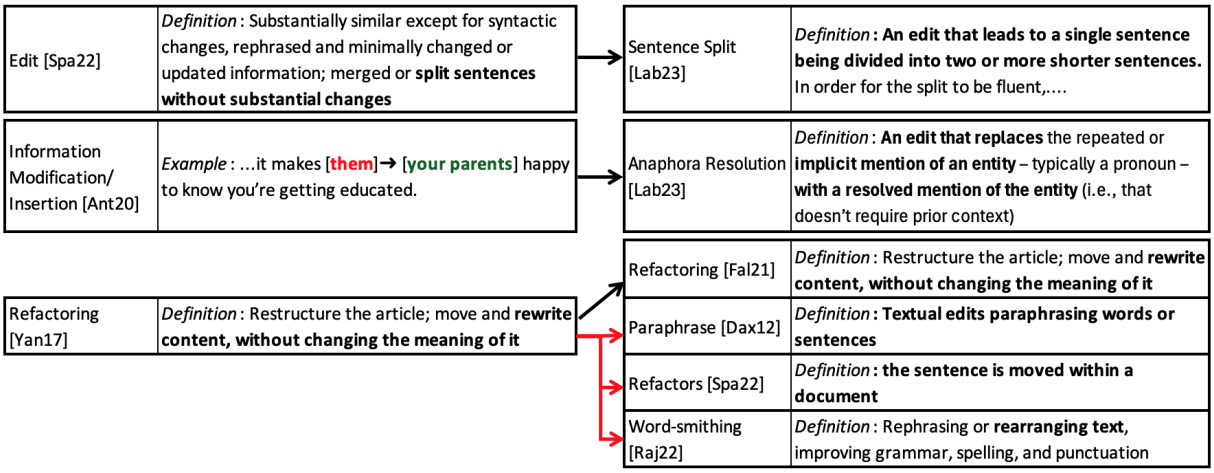



Figure 4: This figure depicts examples of one-to-one, one-to-many cases. We move category examples in Appendix A.6 for compactness. The bold parts of definitions share the same concept between matched categories.

determine the corresponding edit intention. The input granularity varies, and we focus on sentence- and paragraph-level text pairs in this study.

**Datasets** We use two datasets, ITERATER-HUMAN from Du22 and ARXIVEDITS from Jia22. ITERATER-HUMAN covers Wikipedia articles, ArXiv papers, and Wikinews and contains paragraph-level as well as sentence-level edits. ARXIVEDITS is an annotated corpus of papers from ArXiv with different granularities. We randomly sample 10 edits from ITERATER-HUMAN and ARXIVEDITS per category per domain with mixed levels as a training set. This training set is divided into two subsets for two rounds of the training session, 155 and 100 text pairs, respectively. Then, we use the same strategy to sample 1000 text pairs from two datasets, denoted by $D_{Du22}$, and $D_{Jia22}$, for the evaluations.

**Inter-Annotator Agreement** We follow the method of inter-annotator agreement (IAA) in Du22, using the Fleiss' $\kappa$ (Fleiss, 1971). A higher IAA score corresponds to better agreement among annotators, suggesting that the taxonomy is easily understood and used by human annotators.

**Annotators** We follow the qualification criteria from Du22 to select three annotators, who then undergo training for each EIT used in our study (Du22, Rua22, and UNIT). The training covers annotation guidelines, revision examples, live demonstrations, and practice annotations. Appendix A.10 provides the instructions and examples. To ensure a thorough understanding, we conduct two rounds of practice annotations. Table 2 reports the IAA scores on the training set. After the first round, we hold a discussion to clarify annotator confusions, leading to a substantial improvement in IAA in the second round.

| | Dataset | $\kappa_{\text{UNIT}}$ | $\kappa_{Du22}$ | $\kappa_{Rua24}$ |
|---|---|---|---|---|
| Training 1 | $D_{Du22}$ | 0.4744 | 0.4553 | / |
| | $D_{Jia22}$ | 0.5483 | / | 0.3844 |
| Training 2 | $D_{Du22}$ | 0.5422 | 0.4801 | |
| | $D_{Jia22}$ | 0.5590 | / | 0.5107 |

Table 2: Inter-Annotator agreement (Fleiss' $\kappa$) across two rounds of training annotations using the EIT in Du22($\kappa_{Du22}$), Rua24($\kappa_{Rua24}$) and UNIT($\kappa_{\text{UNIT}}$).

**Comparison Groups** We choose EITs in Du22 and Rua24 as our baselines. The two EITs are representative of the set of EITs. Du22 gives a general purpose EIT seeking to cover multiple domains and different edit intentions and provides detailed instruction, revision examples, and labeling interface design. Rua24 is a recent work proposing a holistic framework that includes edit intentions. Appendix A.9 introduces details of these two EITs. We conduct 3 sets of experiments:

CG1 Compare UNIT to Du22 on $D_{Du22}$ (RQ1).
CG2 Compare UNIT to Rua22 on $D_{Jia22}$ (RQ1).
CG3 Use UNIT on multiple datasets (RQ2).

## 5.2 Evaluation Results

Table 3 shows the IAA of manual annotations.

**RQ1** We use the IAA scores for each of the three taxonomies ($\kappa_{\text{UNIT}}$, $\kappa_{Du22}$ and $\kappa_{Rua24}$) to assess UNIT's improvement gain compared to Du22 and Rua24. For CG1, UNIT achieves a higher agreement ($\kappa_{\text{UNIT}}$ = 0.4007) compared to Du22's EIT

| Dataset | $\kappa_{\text{UNIT}}$ | $\kappa_{Du22}$ | $\kappa_{Rua24}$ |
|---|---|---|---|
| $D_{Du22}$ | 0.4007 | 0.3695 | / |
| $D_{Jia22}$ | 0.4139 | / | 0.2885 |

Table 3: The IAA scores are denoted as follows: $\kappa_{Du22}$ for Du22's EIT, $\kappa_{Rua24}$ for Rua24's EIT, and $\kappa_{\text{UNIT}}$ for UNIT. A '/' symbol indicates not applicable.

($\kappa_{Du22}$ = 0.3695). This indicates that UNIT provides clearer definitions or categories that better align with annotators' interpretations compared to the EIT specifically designed for this dataset. For CG2, UNIT outperforms Rua24's EIT, with $\kappa_{\text{UNIT}}$ = 0.4139 versus $\kappa_{Rua24}$ = 0.2885. The performance difference suggests that UNIT better generalizes to human annotation tasks, even on datasets from different sources, as discussed in RQ2.

**RQ2** For CG3, the performance consistency of UNIT is evaluated by analyzing its performance on datasets from diverse and third-party sources $D_{Du22}$ and $D_{Jia22}$, which our work and Rua24 do not release. Consistency is assessed by examining the stability and competitiveness of $\kappa_{\text{UNIT}}$ in comparison to other EITs. UNIT achieves $\kappa_{\text{UNIT}}$ = 0.4007 on $D_{Du22}$ and $\kappa_{\text{UNIT}}$ = 0.4139 on $D_{Jia22}$. On both two datasets, UNIT continues to outperform competing.

## 6 Conclusion

This work addresses the need for a comprehensive and unified EIT across multiple domains by integrating existing EITs into a large, cohesive framework. Our research resolves the inconsistencies and ambiguities inherent in previous works. The proposed EIT, UNIT, not only spans diverse application domains but also considers various analytical aspects, making it highly applicable for human annotators. Evaluation results demonstrate that the integrated taxonomy is more applicable to human annotators by comparing IAA scores among annotators with different EITs. Future research should focus on applying the UNIT to new application domains. We believe its extensibility and adaptability make it a robust tool for future research.

## 7 Limitations

**Comprehensiveness of Evaluation** The evaluation is not comprehensive, as the dataset is limited to already explored domains. We lack edits from new domains, such as software documentation or legal documents, which are necessary to test the generalizability of our UNIT. Future work will involve applying UNIT to new application domains, including the exploration of edit revisions in software documentation, legal documents, and more.

Agreement scores are subject to sampling, data, and user background. In this study, our training and evaluation datasets are sampled from ITERATER-HUMAN and ARXIVEDITS. The slight variation in agreement across training sessions and comparison groups may stem from sampling randomness and dataset size differences.

**Further Refinement and Expansion** We believe UNIT mitigates issues of overlapping definitions by complementing category names and definitions with examples to improve clarity. While we work diligently to construct UNIT, we encourage the community to refine and expand it rather than treating it as a definitive version. UNIT is not a finalized product— taxonomies rarely are.

**Future Work** To ensure consistency, we evaluate UNIT using the same practices reported in the surveyed studies on existing EITs. The study its hierarchical organization is challenging, as it requires masking the intent of evaluating user experiences across taxonomies of different depths. We leave this for future work.

**Ethical Considerations** The data used to evaluate the taxonomy is from publicly available datasets released by Du22 and Jia22. We respect the copyrights of the original document authors and the dataset creators. During the data annotation process, our human annotators are volunteers and are anonymized to protect their privacy rights. We ensure a reasonable annotation time to avoid excessive workload. Our work poses no potential harm to marginalized or vulnerable populations.

## 8 Acknowledgements

This work was supported by the National Science Foundation awards 2026513 and 2230692 (subaward to Temple University). We also thank our reviewers for their feedback and comments.

## References

Talita Anthonio, Irshad Bhat, and Michael Roth. 2020. wikiHowToImprove: A resource and analyses on edits in instructional texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5721–5729, Marseille, France. European Language Resources Association.

Jonas Nygaard Blom and Kenneth Reinecke Hansen. 2015. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76:87–100.

Danielle Brown and Vinicio Sinta. 2016. Six things you didn't know about headline writing: Sensational form in viral news of traditional and digitally native news organizations. *Research Journal of ISOJ*, 6.

Zhijia Chen, Lihong He, Arjun Mukherjee, and Eduard Dragut. 2024. Comquest: Large scale user comment crawling and integration. In *Companion of the 2024 International Conference on Management of Data*, SIGMOD '24, page 432–435, New York, NY, USA. Association for Computing Machinery.

Laura Chiticariu, Mauricio Hernández, Phokion Kolaitis, and Lucian Popa. 2007. Semi-automatic schema integration in clio. pages 1326–1329.

Johannes Daxenberger and Iryna Gurevych. 2012. A corpus-based study of edit categories in featured and non-featured Wikipedia articles. In *Proceedings of COLING 2012*, pages 711–726, Mumbai, India. The COLING 2012 Organizing Committee.

Johannes Daxenberger and Iryna Gurevych. 2013. Automatically classifying edit categories in Wikipedia revisions. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 578–589, Seattle, Washington, USA. Association for Computational Linguistics.

Rod Dilnutt. 2004. The role of taxonomy in knowledge management. *The International Journal of Knowledge, Culture, and Change Management: Annual Review*, 3.

Hong-Hai Do and Erhard Rahm. 2002. Coma: a system for flexible combination of schema matching approaches. In *Proceedings of the 28th International Conference on Very Large Data Bases*, VLDB '02, page 610–621. VLDB Endowment.

E. Dragut, Wensheng Wu, P. Sistla, C. Yu, and Weiyi Meng. 2006a. Merging source query interfaces on-web databases. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 46–46.

E.C. Dragut, W. Meng, and C. Yu. 2022. *Deep Web Query Interface Understanding and Integration*. Synthesis Lectures on Data Management. Springer International Publishing.

Eduard Dragut, Fang Fang, Prasad Sistla, Clement Yu, and Weiyi Meng. 2009a. Stop word and related problems in web interface integration. *Proc. VLDB Endow.*, 2(1):349–360.

Eduard C. Dragut, Fang Fang, Clement Yu, and Weiyi Meng. 2009b. Deriving customized integrated web query interfaces. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 685–688.

Eduard C. Dragut, Thomas Kabisch, Clement Yu, and Ulf Leser. 2009c. A hierarchical approach to model web query interfaces for web source integration. *Proc. VLDB Endow.*, 2(1):325–336.

Eduard C. Dragut, Clement Yu, and Weiyi Meng. 2006b. Meaningful labeling of integrated query interfaces. In *Proceedings of the 32nd International Conference on Very Large Data Bases*, VLDB '06, page 679–690. VLDB Endowment.

Wanyu Du, Vipul Raheja, Dhruv Kumar, Zae Myung Kim, Melissa Lopez, and Dongyeop Kang. 2022. Understanding iterative revision from human-written text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3573–3590, Dublin, Ireland. Association for Computational Linguistics.

Lester Faigley and Stephen Witte. 1981. Analyzing revision. *College Composition and Communication*, 32(4):400–414.

Felix Faltings, Michel Galley, Gerold Hintz, Chris Brockett, Chris Quirk, Jianfeng Gao, and Bill Dolan. 2021. Text editing by command. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5259–5274, Online. 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.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.*, 76(5):378–382.

Xingzhi Guo, Brian Kondracki, Nick Nikiforakis, and Steven Skiena. 2022. Verba volant, scripta volant: Understanding post-publication title changes in news outlets. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 588–598, New York, NY, USA. Association for Computing Machinery.

Robert A Harris. 2017. *Writing with clarity and style*, 2 edition. Routledge, Second edition. | New York, NY : Routledge, [2018] | Previous edition published in 2003.

Bin He, Kevin Chen-Chuan Chang, and Jiawei Han. 2004. Discovering complex matchings across web query interfaces: a correlation mining approach. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 148–157, New York, NY, USA. Association for Computing Machinery.

Chao Jiang, Wei Xu, and Samuel Stevens. 2022. arXivEdits: Understanding the human revision process

in scientific writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9420–9435, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

John Jones. 2008. Patterns of revision in online writing: A study of wikipedia's featured articles. *Written Communication*, 25(2):262–289.

Omid Kashefi, Tazin Afrin, Meghan Dale, Christopher Olshefski, Amanda Godley, Diane Litman, and Rebecca Hwa. 2022. ArgRewrite v.2: an annotated argumentative revisions corpus. *Lang. Resour. Eval.*, 56(3):881–915.

Philippe Laban, Jesse Vig, Wojciech Kryscinski, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. SWiPE: A dataset for document-level simplification of Wikipedia pages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10674–10695, Toronto, Canada. Association for Computational Linguistics.

Seonah Lee, Rongxin Wu, Shing-Chi Cheung, and Sungwon Kang. 2021. Automatic detection and update suggestion for outdated api names in documentation. *IEEE Transactions on Software Engineering*, 47(4):653–675.

Xin Li, Jia Zhou, Honglian Xiang, and Jingjing Cao and. 2024. Attention grabbing through forward reference: An erp study on clickbait and top news stories. *International Journal of Human–Computer Interaction*, 40(11):3014–3029.

Jun Liu and Sudha Ram. 2011. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Trans. Manage. Inf. Syst.*, 2(2).

Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. 2006. Cultural Differences in Collaborative Authoring of Wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113.

Dheeraj Rajagopal, Xuchao Zhang, Michael Gamon, Sujay Kumar Jauhar, Diyi Yang, and Eduard Hovy. 2022. One document, many revisions: A dataset for classification and description of edit intents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5517–5524, Marseille, France. European Language Resources Association.

Dietrich Rathjens. 1985. The seven components of clarity in technical writing. *IEEE Transactions on Professional Communication*, PC-28(4):42–46.

Qian Ruan, Ilia Kuznetsov, and Iryna Gurevych. 2024. Re3: A holistic framework and dataset for modeling collaborative document revision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4635–4655, Bangkok, Thailand. Association for Computational Linguistics.

Darsh J. Shah, Tal Schuster, and Regina Barzilay. 2019. Automatic fact-guided sentence modification. In *AAAI Conference on Artificial Intelligence*.

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.

Srikrishnan Sundaresan and G. Hu. 2005. Schema integration of distributed databases using hyper-graph data model. In *IRI -2005 IEEE International Conference on Information Reuse and Integration, Conf, 2005.*, pages 548–553.

Lei Wang and Eduard Dragut. 2024. The overlooked repetitive lengthening form in sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16225–16238, Miami, Florida, USA. Association for Computational Linguistics.

Wensheng Wu, Clement Yu, AnHai Doan, and Weiyi Meng. 2004. An interactive clustering-based approach to integrating source query interfaces on the deep web. In *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, SIGMOD '04, page 95–106, New York, NY, USA. Association for Computing Machinery.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2016. Who did what: Editor role identification in wikipedia. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):446–455.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in Wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010, Copenhagen, Denmark. Association for Computational Linguistics.

Fan Zhang, Homa B. Hashemi, Rebecca Hwa, and Diane Litman. 2017. A corpus of annotated revisions for studying argumentative writing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1568–1578, Vancouver, Canada. Association for Computational Linguistics.

Fan Zhang, Rebecca Hwa, Diane Litman, and Homa B. Hashemi. 2016. ArgRewrite: A web-based revision assistant for argumentative writings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 37–41, San Diego, California. Association for Computational Linguistics.

Fan Zhang and Diane Litman. 2015. Annotation and classification of argumentative writing revisions. In

*Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 133–143, Denver, Colorado. Association for Computational Linguistics.

Fan Zhang and Diane Litman. 2016. Using context to predict the purpose of argumentative writing revisions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1424–1430, San Diego, California. Association for Computational Linguistics.

Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Dragut. 2024. Scier: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. *Preprint*, arXiv:2410.21155.

Qi Zhang, Huitong Pan, Zhijia Chen, Longin Jan Latecki, Cornelia Caragea, and Eduard Dragut. 2025. Dynclean: Training dynamics-based label cleaning for distantly-supervised named entity recognition. *Preprint*, arXiv:2504.04616.

Weiyi Meng Zhijia Chen and Eduard Dragut. 2025. Comcrawler: General crawling solution for article comments. https://openproceedings.org/2025/conf/edbt/paper-297.pdf. [Accessed 27-05-2025].

# A Appendix

## A.1 Application Domains

Existing research investigates edit intentions across various application domains, such as article writing, news reporting, Wikipedia, and scientific papers.

**Wikipedia** Pfeil et al. (2006) studies the impact of national cultures on Wikipedia by exploring the patterns in editors' contributions. Jones (2008) examines revision patterns in the collaborative writing process of Wikipedia to explore the editors' experience. Liu and Ram (2011) studies the relationship between collaboration and article quality in Wikipedia, identifying the roles of editors such starters, copy editors, and all-round contributors. Daxenberger and Gurevych (2012, 2013) compile a corpus of English Wikipedia edits, train a classifier to identify the edit intentions and study the collaborative behaviors among editors. Yang et al. (2016) explores the roles of editors in Wikipedia article revisions and examines how these roles influence article quality. Yang et al. (2017) focuses on capturing semantic edit intentions underlying syntactic edit operations in Wikipedia articles. (Faltings et al., 2021) analyzes the revisions with comments to develop an interactive text editing system that assists users in editing text according to their goals. (Rajagopal et al., 2022) models the relationship between edits and comments, offering insights into document evolution through an analysis of revision intentions. Similarly, Anthonio et al. (2020) investigates edits in instructional text on wikiHow to determine if the changes improve instructions in terms of style or clarity.

**Writings** Faigley and Witte (1981) examines the intentions behind revisions and the impact of edits on meaning by analyzing both student and expert writings. Rathjens (1985) provides an exploration of the concept of 'clarity' in technical writing. Zhang and Litman (2015, 2016) identify the revision purposes in student writings to predict potential grade improvements after revisions. Zhang et al. (2017); Kashefi et al. (2022) provide revision corpora to support natural language applications. Jiang et al. (2022) unveils the common strategies practiced by researchers for revising their papers.

**News** Spangher et al. (2022) analyzes the revision histories to explore the narrative and factual evolution of News articles, as well as the predictability of edit actions. Guo et al. (2022) investigates the motivation and frequency of post-publication changes to news headlines, particularly those made after the

initial release. Du et al. (2022) focuses on defining broader edit intentions across multiple domains, including news.

**Others** Anthonio et al. (2020) explored the edits of instructional text in wikiHow to identify whether the edits improve instructions in terms of style and correctness or necessary information for clarity.

### A.2 Data Sources and Links

We survey the availability of publicly released data such EITs, annotated text, and guidelines in this section. We gather the sources of data that researchers analyze to explore edit revisions, as detailed in Table 4. Additionally, we collect the links of source code for re-generating data, re-training classifiers, and datasets with annotated labels, contributing to the analysis of edit intentions, as presented in Table 5.

Daxenberger and Gurevych (2012) released their annotated data with a complete annotation guideline, but the link in the follow-up work Daxenberger and Gurevych (2013) is no longer valid. Zhang et al. (2016) deployed a rewrite assistant website and released a revision component in Java. Yang et al. (2017) released the code for re-extracting the data from Wikipedia article revisions, along with a multi-label classifier to predict edit intentions. Zhang et al. (2017) released a small revision corpus with sentence-level edits in student papers and manually developed annotations. Faruqui et al. (2018) released revision corpus focusing on insertions and deletions. Kashefi et al. (2022) provides a larger revision corpus of student papers with sentential and subsentential-level edits. Anthonio et al. (2020) provided the pre-trained BiLSTM model to identify the edit intentions for text pairs, the code for re-training the classifier from scratch, and the crawler to fetch the instructional text revisions from the wikiHow website. Rajagopal et al. (2022) released data annotations along with editor comments. Spangher et al. (2022) collected news edits from 20 news agencies, processing revision from 17 of them and releasing the data publicly. Jiang et al. (2022) released datasets with different granularities of edits and provided a pipeline to do sentence-level alignment and extract fine-grained edits by inputting aligned paragraph pairs. Du et al. (2022) released both human-annotated and automatic-annotated datasets, as well as the code for crawling and preprocessing data for model training and testing. Guo et al. (2022) released their news headline dataset, though the data lacks the annotation with their taxonomy. Laban et al. (2023) released document-level revisions for the natural language simplification task and provided fine-trained language models to automatically label edits. Ruan et al. (2024) released corpus including review, revision and response data.

### A.3 Grouping Results

Tables 6 to 10 illustrate the results of grouped EITs after lineage analysis. Each column, except for the leftmost one, represents an existing EIT. Each row presents matched categories, conveying a similar edit intention. Entries in a row may contain one or more categories. A single category may appear in different entries due to one-to-many category matches. The leftmost column of each table is a node proposed to tentatively represent the grouped categories for the following integration process.

### A.4 Category Name Assignment

Another task in the EIT integration is to assign names to the categories/nodes of UNIT. Good category name helps user to understand the edit intention and also for their applications (Dragut et al., 2006b). We follow three strategies, selecting an existing category name, slightly modifying an existing category name, or proposing a new name. For example, in the cluster Refactoring, Word-Smithing, Paraphrase, and Word-usage/Clarity, we designate Word-Smithing as the cluster name. In general, we prioritize longer and more descriptive names. However, when a name is consistently used across EITs and conveys a precise, narrow meaning, we opt for shorter, single-word names. In some cases, we make slight edits to candidate names, such as modifying Syntactic Generic to Syntactic Changes. Finally, we introduce new names, particularly for top-layer nodes, which lack corresponding categories in the original EITs. Examples of new names include Source Code and Collaborative Authoring. Figure 5, which illustrates UNIT, uses a coding scheme to indicate the procedures followed for naming.

### A.5 Category Definition

We follow a similar approach for assigning definitions to categories. Our first attempt is find a representative definition from the existing ones for a category. In some clusters, we create a definition by modifying an existing one or combining text from multiple definitions. For example, the

| paper | code? | data? | sources | link? |
|---|---|---|---|---|
| Fai81 | No | No | (In)experienced student and expert revisions | No |
| Zha15 | No | No | Student written papers | No |
| Zha16(A) | No | No | Student writings | Yes |
| Zha17 | No | Ye | Student writings | Yes |
| Fal21 | Yes | Ye | Wikipedia revision histories | Yes |
| Du22 | Yes | Ye | Formally human-written text(Wikipedia, ArXiv, Wikinews) | Yes |
| Jia22 | Yes | Ye | ArXiv Papers Revisions | Yes |
| Kas22 | / | Ye | argumentative writing essays | Yes |
| Rua24 | Yes | Ye | F1000RD dataset in Kuznetsov et al., 2022, ARR-22 subset of the NLPeer corpus in Dycke et al., 2023 | Yes |
| Rat85 | / | / | / | No |
| Zha16(B) | No | No | High school student written papers | No |
| Yan16 | No | No | Three datasets from English edition of Wikipedian | No |
| Jon08 | No | No | (Non-)featured Wikipedia articles revision histories | No |
| Dax12 | No | Ye | (Non-)featured Wikipedia articles revision histories | Yes |
| Dax13 | No | Ye | Wikipedia revision histories | Yes |
| Yan17 | Yes | Ye | Wikipedia revision histories | Yes |
| Lab23 | Yes | Ye | Wikipedia's revision history | Yes |
| Pfe06 | No | No | Wikipedia page revision histories | No |
| Liu11 | No | No | Different quality-level Wikipedia articles based on quality | No |
| Far18 | No | Yes | revision histories | Yes |
| Raj22 | No | Yes | Wikipedia revision histories | Yes |
| Ant20 | Yes | Yes | wikiHow (non-)featured articles revision histories | Yes |
| Spa22 | Yes | Yes | News revision histories | Yes |
| Guo22 | No | Yes | News headlines from major US news agencies | Yes |

Table 4: This table summarizes whether the literature provided their code and data. **code?** indicates if the code for preprocessing or regenerating data was released. **data?** shows if the data was provided along with their taxonomy. **sources** specifies the sources of the revisions their taxonomy and analysis are based on. **link?** indicates if the links to their code or data were provided. '/' denotes not applicable.

definition of Forward Reference (Figure 5) is modified from "*a reference to upcoming discourse or to a word or a phrase later in the text*" (Li et al., 2024). When the existing definitions are specific to the methodology/application domain of a particular paper, we generalize them to ensure broader applicability. We create new definitions to a few categories, such as "*Edits for all online collaborative writing: Wikipedia, Github, online forums, etc.*" for the top-layer category Collaborative Authoring. We use the same coloring scheme as in Figure 5 to indicate the definition assignment procedure.

### A.6 Revision Examples

Table 11 illustrates the revision examples of categories in Figure 4.

### A.7 Existing EITs Matching and Comparison

Table 12 presents a structured comparison of edit intention categories across different studies, including their definitions and representative revision examples. Due to space constraints, we illustrate only a subset of this comparison—specifically, the categories related to *Information Modification* and *Reference Changes*. The full comparison table is available at the provided link.

**Information Modification**. The edit intentions grouped under Information Modification share a common purpose: to alter or update informational content. This includes changes to facts, theorems, or supporting evidence (Zha15/Zha16/Kas22), as well as factual data updates (Fal21/Raj22). Dax12 introduces a broader category for content-affecting edits, while Du22 emphasizes the addition of new information, bridging factual revision and meaning shifts. The bolded phrases in their definitions highlight a shared goal of modifying factual or evidential content.

**Reference Changes**. The intentions categorized under Reference Changes center around modifying references or citations—whether by adding, removing, or verifying them. While Zha15/Zha16/Kas22 emphasize citation as a form of evidential support, Yan17 and Fal21 focus more specifically on source verifiability. Despite overlapping definitions, their underlying purposes differ: one supports claims, the other ensures reliability.

It is worth noting that the Evidence category (Zha15/Zha16/Kas22) spans both Information Modification and Reference Changes due to its broad definition. However, the revision examples associated with it in the Reference Changes context do not directly align with the bolded definition. This indicates the need for more representative revi-

| alias | Link | Work? |
|---|---|---|
| Zha16(A) | https://people.cs.pitt.edu/~zhangfan/argrewrite/ | Active |
| | https://www.cs.pitt.edu/~zhangfan/revisionCorrection.jar | Active |
| Zha17 | https://argrewrite.cs.pitt.edu/ | Active |
| Fal21 | http://microsoft.com/research/project/interactive-document-generation | broken |
| Du22 | https://github.com/vipulraheja/IteraTeR | Active |
| Jia22 | https://github.com/chaojiang06/arXivEdits | Active |
| Kas22 | https://argrewrite.cs.pitt.edu/ | Active |
| Rua24 | Code: https://github.com/UKPLab/re3 | Active |
| | Dataset:https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/4300 | Active |
| Dax12 | https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2354 | Moved |
| Dax13 | http://www.ukp.tu-darmstadt.de/data/edit-classification | Broken |
| Yan17 | https://github.com/diyiy/Wiki_Semantic_Intention | Active |
| Lab23 | https://github.com/Salesforce/simplification | Active |
| Far18 | https://github.com/google-research-datasets/wiki-atomic-edits?tab=readme-ov-file | Active |
| Raj22 | https://tinyurl.com/editsumm | Active |
| Ant20 | https://github.com/irshadbhat/wikiHowToImprove | Active |
| Spa22 | https://github.com/isi-nlp/NewsEdits.git | Active |
| Guo22 | https://scripta-volant.github.io/ | Active |

Table 5: It collects the links from the existing works which release their code or data. **work?** shows whether the provided link is still working or not; 'Moved' means the source link has changed but is still accessible. 'Broken' indicates the provided link is not accessible and the source cannot be found via searches.

| Combined(DuRa) | Rat85 | Du22 |
|---|---|---|
| **Fluency** | | Fluency |
| **Coherence** | | Coherence |
| **Clarity** | Brevity, Order, Accuracy, Completeness, Emphasis, Consistency, Objectivity | Clarity |
| **Style** | | Style |
| **Meaning-Changed** | | Meaning-Changed |

Table 6: The result of grouping Rat85 and Du22.

| Combined (DaYang) | Dax12/Dax13 | Yan16 |
|---|---|---|
| **Vandalism** | Vandalism | |
| **Revert** | Revert | |
| **Paragraph** | Paragraph | Rephrase |
| **Spelling /Grammar** | Spelling /Grammar | Grammar |
| **Relocation** | Relocation | Relocation |
| **Markup-IDM** | Markup-IDM | Markp-IDM |
| **Info.-IDM** | Info.-IDM | Info.-IDM |
| **File-IDM** | File-IDM | File-IDM |
| **Ref.-IDM** | Ref.-IDM | External Link-IDM, Ref.-IDM, Wikilink-IDM |
| **Template-IDM** | Template-IDM | Template-IDM |

Table 7: The result of grouping Dax12, Dax13, and Yan16. Dax13 reused EIT in Dax12, so we combine two EITs into one column. Inf.-IDM and Ref.-IDM are abbreviated forms of Information-IDM and Reference-IDM, respectively.

sion examples—particularly ones that align better with the reference-focused intentions seen in Verification (Yan17) and Verifiability (Fal21).

### A.8 Taxonomy Integration Result

Figure 5 presents our UNIT: the unified and comprehensive EIT. UNIT has three layers including Top-layer (helping annotators quickly locate edit intentions), Edit Intentions (the categories used in the evaluation), Fine-grained Intentions (using as needed). Categories in each layer have clear definitions and illustrative revision examples (omit in Figure 5 due to space limitations). Blocks surrounded by a solid line are categories and by a dashed line are definitions of categories. The blocks in grey blocks mean that categories/definitions are taken from the literature. Those in blue mean that categories are proposed by the authors based on matched categories, addressing cases where definitions may be missing or not general for all edit intentions. Those in purple mean that categories/definitions are modified from the literature.

**Application Configuration** Our UNIT can be applied across various domains. For instance, if our research focuses on analyzing the edit intentions behind PostgreSQL documentation, which falls under informative/explanatory writings, we need to filter relevant categories from our UNIT. In this scenario, we select categories under Informative/Explanatory Writing, Source Code, Non-Meaning-Changed, and Meaning-Changed. We include categories under Source Code categories because PostgreSQL documentation also incorporates SQL code to demonstrate SQL usage.

Our UNIT does not explicitly specify the unit of edit within categories, aligning with conventions in related work. Instead, we allow users to determine the input granularity.

| Combined(zhaka) | Zha15 | Zha16(B) | Zha16(A) | Zha17 | Kas22 |
|---|---|---|---|---|---|
| **Organization** | Organization | Organization | Reordering (Organization) | Organization (Reordering) | Organization (ORG) |
| **Conventions /Grammar /Spelling** | Conventions /Grammar /Spelling | Conventions | Errors (Conventions /Grammar /Spelling) | Conventions /Grammar /Spelling (Errors) | Spelling and Gramma (SPL) |
| **Word-usage /Clarity** | Word-usage /Clarity | Clarity | Fluency (Word-usage /Clarity) | Word Usage /Clarity (Fluency) | Word Usage (WRD) |
| **Claim/Ideas** | Claim/Ideas | Claim/Ideas | Thesis/Ideas (Claim) | Claims/Ideas (Ideas) | Claim(CLM) |
| **Warrant /Reasoning /Backing** | Warrant /Reasoning /Backing | Warrant /Reasoning /Backing | Reasoning (Warrant) | Warrant /Reasoning /Backing (Reasoning) | Reasoning (RSN) |
| **Rebuttal /Reservation** | Rebuttal /Reservation | | Rebuttal | Rebuttal /Reservation (Rebuttal) | Rebuttal (RBL) |
| **General Content** | General Content | General Content | Other content changes | General Content (Other) | General Content Development(GCD) |
| **Evidence** | Evidence | Evidence | Evidence | Evidence | Evidence (EVD) |
| **Precision** | | | Precision | | |
| **Unknown** | | | Unknown | | |

Table 8: The result of grouping Zha15, Zha16(B), Zha16(A), Zha17 and Kas22

| Combined(YFR) | Yan17 | Fal21 | Raj22 |
|---|---|---|---|
| **Clarification** | Clarification | Clarification | Adding Support Evidence |
| **Copy-editing** | Copy Editing | Copy-editing | Word-smithing |
| **Counter-vandalism** | Counter Vandalism | Counter-vandalism | Wikipedia-specific edits |
| **Disambiguation** | Disambiguation | Disambiguation | Wikipedia-specific edits |
| **Elaboration** | Elaboration | Elaboration | Adding New Information |
| **Fact Update** | Fact Update | Fact Update | Fact Update |
| **Point of View** | Point of View | Point of View | Point of View Change |
| **Process** | Process | Process | Wikipedia-specific edits |
| **Refactoring** | Refactoring | Refactoring | Word-smithing |
| **Simplification** | Simplification | Simplification | Remove existing information |
| **Vandalism** | Vandalism | Vandalism | Wikipedia-specific edits |
| **Verification** | Verification | Verifiability | Adding Support Evidence |
| **Wikification** | Wikification | Wikification | Wikipedia-specific edits |
| **Unlabeled** | | Unlabeled | |

Table 9: The result of grouping Yan17, Fal21 and Raj22.

| Combined(PL) | Pfe06 | Liu11 |
|---|---|---|
| **Add Information** | Add Information | Sentence Insertion |
| **Reference Insertion** | Add link | Link Insertion, Reference Insertion |
| **Modify Information** | Clarify Information | Sentence Modification |
| **Delete Information** | Delete Information | Sentence Deletion |
| **Reference Deletion** | Delete Link | Link Deletion, Reference Deletioin |
| **Reference Modification** | Fix Link | Link Modification, Reference Modification |
| **Format** | Format | |
| **Grammar** | Grammar | Sentence Modification |
| **Mark-up Language** | Mark-up Language | |
| **Reversion** | Reversion | Revert |
| **Spelling** | Spelling | Sentence Modification |
| **Style/Typography** | Style/Typography | |
| **Vandalism** | Vandalism | |

Table 10: The result of grouping Pfe06 and Liu11.

| Category | Definition | Example |
|---|---|---|
| Edit[Spa22] | Substantially similar except for syntactic changes, rephrased and minimally changed or updated information; merged or **split sentences without substantial changes** | ...he regrets exposing her to the deadly virus [and had] → [. Had] he known he was carrying Ebola, ... |
| Sentence Split[Lab23] | An edit that leads to a single sentence being divided into two or more shorter sentences. In order for the split to be fluent, words are typically removed and inserted at the sentence boundary. If no non-connector content is added, then it is not only a sentence split. | / |
| Anaphora Resolution[Lab23] | An edit that replaces the repeated or implicit mention of an entity – typically a pronoun – with a resolved mention of the entity (i.e., that doesn't require prior context). | / |
| Refactoring[Yan17] | Restructure the article; move and **rewrite content, without changing the meaning of it** | / |
| Refactoring[Fal21] | Restructure the article; move and **rewrite content, without changing the meaning of it** | / |
| Paraphrase[Dax12] | **Textual edits paraphrasing words or sentences** | denominations [like] → [such as] the |
| Refactors[Spa22] | **the sentence is moved within a document** | "[The mother, this was the first time seeing her son since he got to the States.]" ... → "[The mother, this was the first time seeing her son since he got to the States.]" |
| Word-smithing[Raj22] | Rephrasing or rearranging text, **improving grammar, spelling, and punctuation** | ...as well as [accomodation] → [accommodation] for visitors... |

Table 11: The categories in this table are the actual category names, definitions, and revision examples in existing Edit Intention Taxonomies (EITs). The revision examples of categories only cover the **bold** part of the definitions.

### A.9 Compared EITs in Evaluation

Du22 introduces a comprehensive, multi-domain taxonomy to model iterative text revision processes in formal writing. Its edit intention taxonomy is divided into two broad categories:

*Non-meaning-changing edits:*

- Fluency: Fixing grammar and syntax issues.
- Clarity: Improving formality, conciseness, or readability.
- Coherence: Enhancing logical flow and consistency.
- Style: Reflecting personal writing preferences (tone, emotion).

*Meaning-changing edits:*

- Meaning-Changed: Adding or modifying factual content or evidence.

*Other:* Edits not fitting into the above categories.

This taxonomy supports annotations at sentence and paragraph levels across domains like scientific abstracts, Wikipedia, and news. The goal is to better understand how edits improve writing and guide models in generating high-quality revisions.

Rua24 proposes the Re3 framework, which focuses on the collaborative revision process, particularly in academic writing, integrating reviews, revisions, and author responses. It offers a full-scope annotation taxonomy along three dimensions:

*Granularity:* Edits are labeled at the section, paragraph, sentence, and subsentence levels.

*Action:* Includes Add, Delete, Modify, Merge, Split, and Fusion.

*Intent:*

- Surface-level: Grammar and Clarity.
- Semantic-level: Claims and Factual/Evidence changes.
- Other: Edits that don't fall into the above.

This taxonomy enables deep analysis of scholarly editing behaviors and supports novel tasks like edit intent classification and document edit summarization using large language models.

### A.10 Human Annotation Instructions and Interface

To guide human annotators in making accurate edit-intention annotations, we provide a brief task instruction in Figure A.10, followed by concrete revision examples. Due to space limitations, these revision examples are shown in the shared link. We highlight the edits (i.e., the differences between text pairs) within sentence-level revisions and pose questions to the annotators to accurately determine the edit intention, as illustrated in Figure 8. Figure 8 displays only partial questions for obtaining the edit intention; the complete set of questions is available in the shared link. Additionally, we provide

| UNIT Label | Category | Definition | Revision Examples |
|---|---|---|---|
| **Information Modification** | Information-Modify[Dax12] | **Textual edits affecting information content** | open steppes of Kashmir and {Siberia → Manchuria}. |
| | Evidence[Zha15] | **change of facts, theorems** or citations **for supporting claims/ideas** | In the circle I would place Fidel, He was annoyed with the existence of the United States and used his army to force them out of his country, Although Fidel claimed that this is for his peoples' interest, it could not change the fact that he is a wrathful person. |
| | Evidence[Zha16(B)] | **change of facts, theorems** or citations **for supporting claims/ideas** | / |
| | Evidence[Kas22] | **change of facts, theorems** or citations **for supporting claims/ideas** | An example for the case where the electronic communication is limited would be {China→ North Korea}. |
| | Fact Update[Fal21] | **Update numbers, dates, scores, episodes, status, etc. based on newly available information** | He married Margaret Frances Prowse Shaw in Sydney in {1874 → 1871}. |
| | Fact Update[Raj22] | **Updating facts in document** | / |
| | Meaning-Changed[Du22] | **Update or add new information to the text** | This method improves the model accuracy from 64% to {78→ 83}%. |
| **Reference Changes** | Evidence[Zha15] | **change of** facts, theorems or **citations for supporting claims/ideas** | ~~In the circle I would place Fidel, He was annoyed with the existence of the United States and used his army to force them out of his country, Although Fidel claimed that this is for his peoples' interest, it could not change the fact that he is a wrathful person.~~ |
| | Evidence[Zha16(B)] | **change of** facts, theorems or **citations for supporting claims/ideas** | / |
| | Evidence[Kas22] | **change of** facts, theorems or **citations for supporting claims/ideas** | ~~An example for the case where the electronic communication is limited would be {China → North Korea}.~~ |
| | Verification[Yan17] | **Add/modify references/citations**; remove unverified text | / |
| | Verifiability[Fal21] | **Add/modify references/citations**; remove unverified text | / |

Table 12: Partial comparison of labels related to Information Modification and Reference Changes across different EITs. Bolded phrases in the definitions highlight semantically aligned concepts across studies. The table also includes example revisions and comments on subtle differences in scope or intent.

the context of the edit (the document revision) to better ascertain the edit intention, as some edits require contextual understanding.

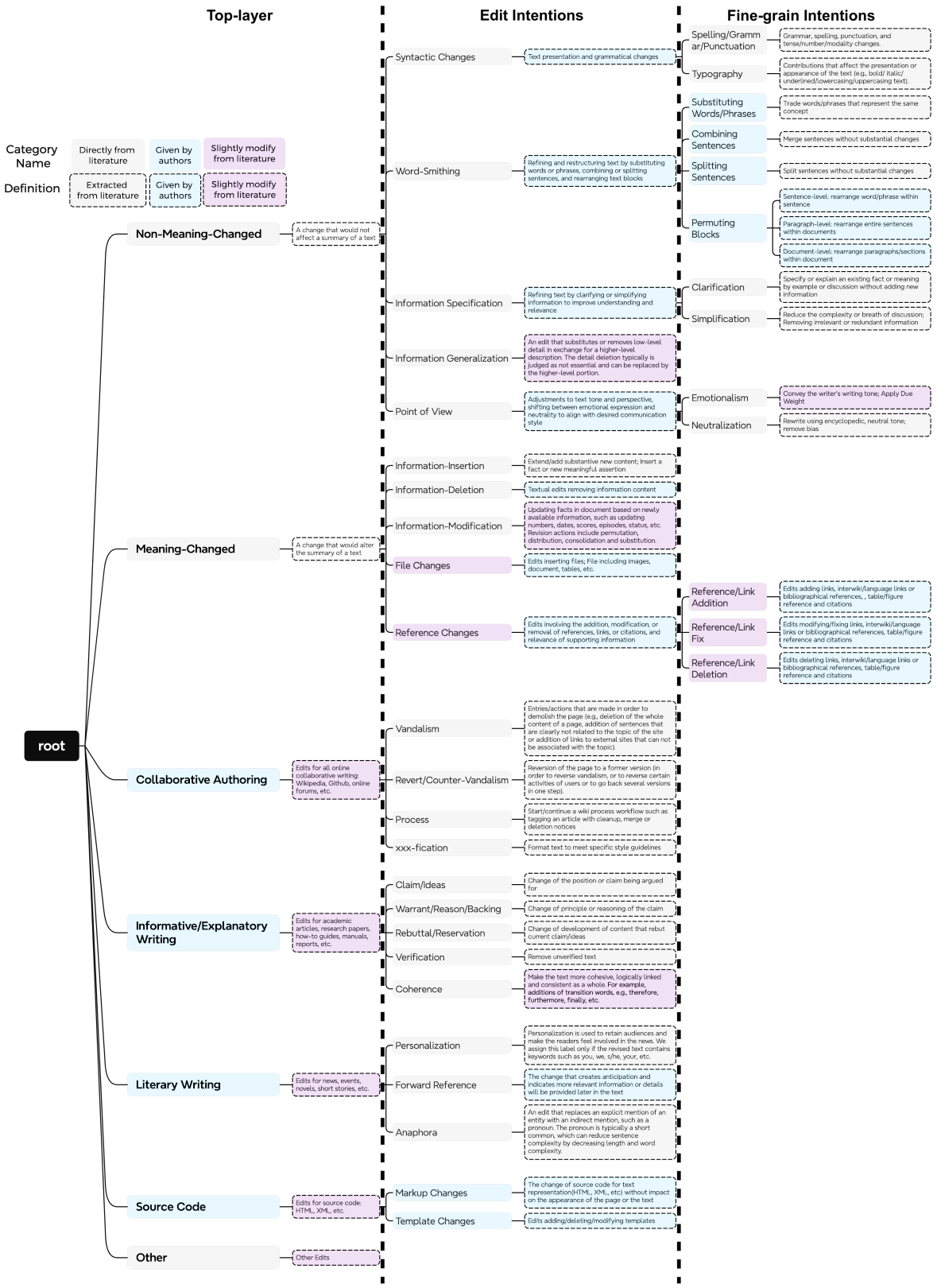
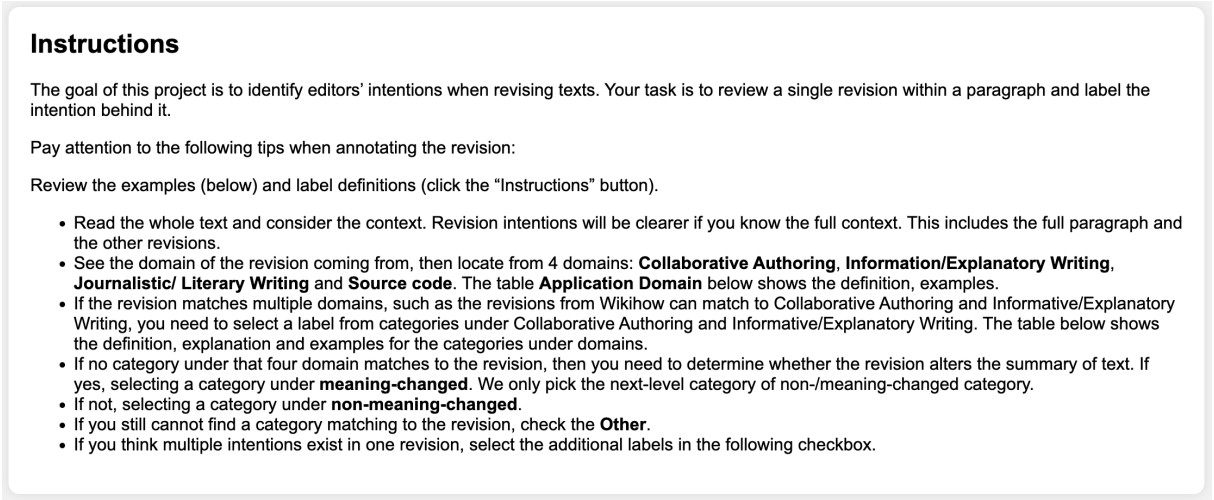



Figure 5: UNIT, a unified, comprehensive, large, extensible edit-intention taxonomy by integrating all EITs.



## Instructions

The goal of this project is to identify editors' intentions when revising texts. Your task is to review a single revision within a paragraph and label the intention behind it.

Pay attention to the following tips when annotating the revision:

Review the examples (below) and label definitions (click the "Instructions" button).

- Read the whole text and consider the context. Revision intentions will be clearer if you know the full context. This includes the full paragraph and the other revisions.
- See the domain of the revision coming from, then locate from 4 domains: **Collaborative Authoring**, **Information/Explanatory Writing**, **Journalistic/ Literary Writing** and **Source code**. The table **Application Domain** below shows the definition, examples.
- If the revision matches multiple domains, such as the revisions from Wikihow can match to Collaborative Authoring and Informative/Explanatory Writing, you need to select a label from categories under Collaborative Authoring and Informative/Explanatory Writing. The table below shows the definition, explanation and examples for the categories under domains.
- If no category under that four domain matches to the revision, then you need to determine whether the revision alters the summary of text. If yes, selecting a category under **meaning-changed**. We only pick the next-level category of non-/meaning-changed category.
- If not, selecting a category under **non-meaning-changed**.
- If you still cannot find a category matching to the revision, check the **Other**.
- If you think multiple intentions exist in one revision, select the additional labels in the following checkbox.

Figure 6: A screenshot of the annotation instruction for human annotators.

**Application Domains**

| Domain | Definition | Examples |
|---|---|---|
| **Collaborative Authoring** | Edits for all online collaborative writing: Wikipedia, Github, online forums, etc. | Wikipedia, Wikihow, etc. |
| **Informative/Explanatory Writing** | Edits for academic articles, research papers, how-to guides, manuals, reports, etc. | Wikihow, ArXiv papers, Wikipedia articles, etc. |
| **Journalistic/Literary Writing** | Edits for news, events, novels, short stories, etc. | News, WikiNews, etc. |
| **Source Code** | Edits for source code: HTML, XML, etc. | HTML, Wikipedia markup languages, etc. |

Figure 7: A screenshot of the **Application Domain** table in the instruction.

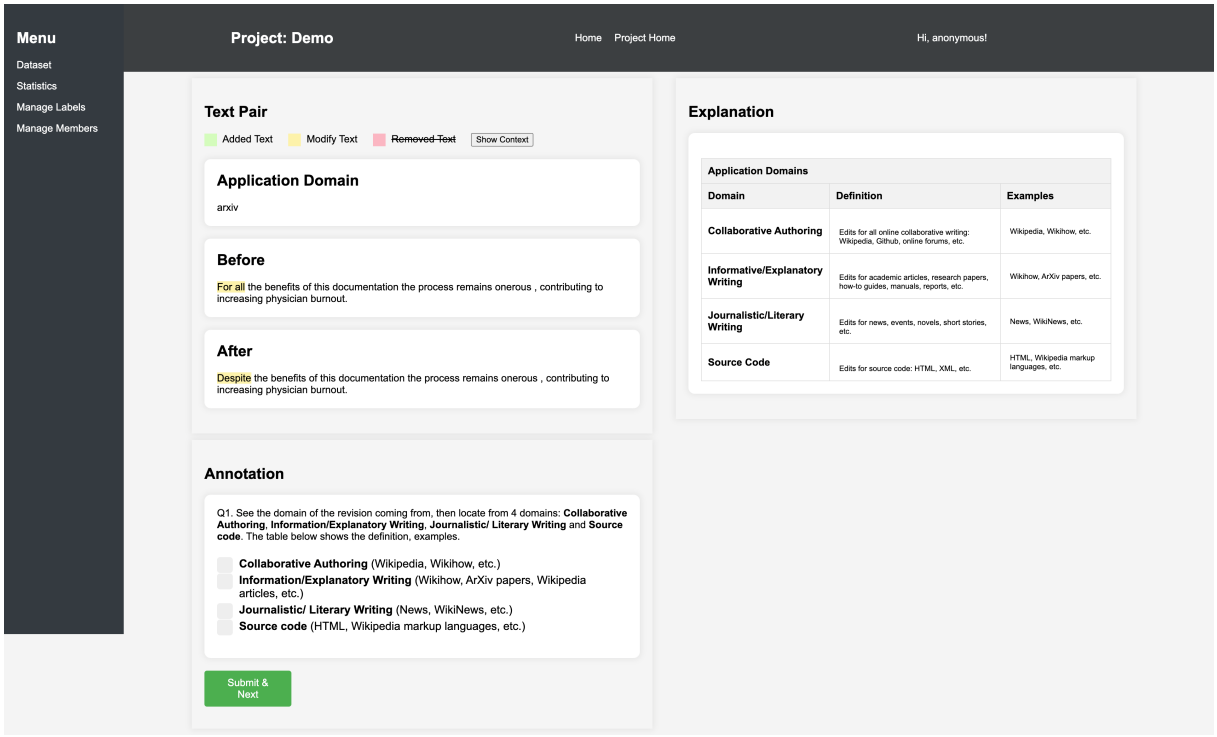



Figure 8: A screenshot of the annotation interface with partial questions for human annotators.