A Corpus-Based Study of Edit Categories in Featured and Non-Featured Wikipedia Articles

Johannes Daxenberger[1] *Iryna Gurevych*[1,2]
(1) Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
(2) Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research and Educational Information
www.ukp.tu-darmstadt.de

ABSTRACT
In this paper, we present a study of the collaborative writing process in Wikipedia. Our work is based on a corpus of 1,995 edits obtained from 891 article revisions in the English Wikipedia. We propose a 21-category classification scheme for edits based on Faigley and Witte's (1981) model. Example edit categories include spelling error corrections and vandalism. In a manual multi-label annotation study with 3 annotators, we obtain an inter-annotator agreement of $\alpha = 0.67$. We further analyze the distribution of edit categories for distinct stages in the revision history of 10 featured and 10 non-featured articles. Our results show that the information content in featured articles tends to become more stable after their promotion. On the opposite, this is not true for non-featured articles. We make the resulting corpus and the annotation guidelines freely available.[1]

TITLE AND ABSTRACT IN GERMAN

Eine Korpusbasierte Studie von Änderungstypen in Exzellenten und Nicht-Exzellenten Wikipedia-Artikeln

In dieser Arbeit stellen wir eine Studie über den kollaborativen Schreibprozess in Wikipedia vor. Unsere Studie basiert auf einem Korpus aus 1.995 Änderungen in 891 Artikelrevisionen der englischen Wikipedia. Wir schlagen ein Klassifikationsschema mit 21 Änderungenstypen vor, basierend auf dem Modell von Faigley and Witte (1981). Unter den Änderungenstypen befinden sich beispielsweise Rechtschreibkorrekturen und Vandalismus. In einer manuellen multi-label Annotationsstudie mit 3 Annotatoren erzielen wir eine Interrater-Reliabilität von $\alpha = 0.67$. Wir analysieren außerdem die Verteilung von Änderungstypen zu unterschiedlichen Stadien in der Revisionsgeschichte von 10 exzellenten und 10 nicht-exzellenten Artikeln. Unsere Ergebnisse zeigen, dass der Informationsgehalt in exzellenten Artikeln nach ihrer Auszeichnung tendenziell stabiler wird. Im Gegensatz dazu ist das bei nicht-exzellenten Artikeln nicht der Fall. Das dabei entstandene Korpus und die Annotationsrichtlinien stellen wir zur freien Verfügung.

KEYWORDS: Wikipedia, Revision History, Collaborative Writing, Quality Assessment.

KEYWORDS IN GERMAN: Wikipedia, Revisionsgeschichte, Kollaboratives Schreiben, Qualitätsbewertung.

[1] http://www.ukp.tu-darmstadt.de/data/wiki-edits/

1 Introduction

Team work on a single product is a common process in daily life. Online collaboration software supports project management, version control systems enable the collaborative development of source code, and recent developments in cloud computing have generated new ways of collaborating on single files. A lot of research has been devoted to the development of user-friendly tools and editors for collaborative writing (Noel and Robert, 2004). Free tools include web-based software such as Zoho Writer, Google Drive or Etherpad, as well as Wikis such as Twiki, Foswiki and MediaWiki. Corpora for analyzing the writing process mostly come from the educational domain (Lee and Webster, 2012). An exception is the Digital Variants Archive[2], which contains contemporary texts by Spanish and Italian authors including various revisions of those texts. However, these corpora only consist of textual revisions by one author. Although there are many tools enabling users to collaboratively write texts, little work has been done to analyze the underlying collaboration process of the data that is created with these tools. One possible reason is that few corpora for analyzing collaborative writing are available.

Since their invention in the mid 90s, Wikis have become one of the most important tools for creating and sharing contents. They enable a detailed tracking of changes, as they usually implement a revision control system which saves every change to a page. At the time of writing, the number of revisions in the English online encyclopedia Wikipedia kept growing by 3.2 million revisions each month.[3] Various studies have processed parts of that data for different tasks such as extracting sentence simplification (Woodsend and Lapata, 2011) or spelling error correction (Zesch, 2012).

Whenever an editor of a page in Wikipedia saves changes, a new revision is created. As one revision may contain a set of distinct local changes, we distinguish between revisions and edits. We define an *edit* as a coherent local change, usually perceived by a human reader as one single editing action. For a pair of adjacent revisions, we denote the previous revision with r_{v-1} and the newer revision with r_v. For each (r_{v-1}, r_v)-pair, we calculate a set of n edits $e^k_{v-1,v}$ (where $k = \{0, 1, ...n-1\}$) that have been made to transform r_{v-1} into r_v (see Section 3.2). We label edits with *edit categories*.

Our contribution in this study is three-fold. First, we develop a classification system for edit categories based on established models from research on the writing process (Faigley and Witte, 1981). This addresses the proposal of Ferschke et al. (2012a) to investigate on the classification of textual revisions. The goal is to facilitate data extraction for NLP applications building upon revision history data. Second, we compile and annotate a corpus tailored towards a qualitative analysis of Wikipedia revisions and based on a set of edits and release it for free access to the research community. To the best of our knowledge, such a corpus is not available yet. Third, based on the annotations in our corpus, we analyze differences in the collaborative writing process of featured and non-featured articles. Featured articles are promoted as such after an internal reviewing process which confirms the required quality standards in Wikipedia (cf. Section 3.2). Although it was not possible to identify a relationship between a certain type of collaboration and article quality in terms of featured and non-featured articles, we show that the collaborative behavior among authors significantly changes once an article is awarded featured status.

The rest of this paper is structured as follows: Section 2 presents the related work. In Section

[2] `http://www.digitalvariants.org/` (accessed 2012-10-29)

[3] Source: `http://stats.wikimedia.org/EN/TablesDatabaseEdits.htm` (accessed 2012-10-29)

3, we describe our edit classification scheme and the corpus. Furthermore, we explain and evaluate the manual annotation of our corpus. Section 4 discusses the findings of our study with respect to related approaches. Finally, we summarize the main conclusions.

2 Related Work

2.1 Collaborative Writing

Sommers (1980) investigated the connections between writing and quality, particularly with respect to differences in the types of edits performed by experienced and unexperienced writers. Her analysis shows that unexperienced writers tend to revise at the sentence or word level, i.e. to make changes on the surface of the text. On the contrary, experienced writers are rather concerned with the meaning and structure of the entire text, that is, they make changes to the text base. In the later research, there has been a shift from revising one's own work (single-author writing) to collaboratively working on a single document (collaborative writing), cf. Ede and Lunsford (1990). Generally, the importance of collaborative writing has grown over the last decades and receives increased interest due to recent developments in the Web 2.0. Collaboration in Wikipedia has been subject to a series of studies (Liu and Ram, 2011). Wikipedia's revision history reflects a type of distributed collaboration, as the interaction between authors is strictly indirect. The communication between authors takes place via the metadata related to each revision in Wikipedia such as the author comment, the revision timestamp and the author name or IP address.

2.2 Edit Classification Schemes

Faigley and Witte (1981) present the first taxonomy capturing the intentions behind a textual change. Their scheme is designed to analyze the effects of edits on meaning. They define meaning as either inserting new information to the text or deleting old information. Edits which affect meaning are called *Text-Base Changes*; edits which do not affect meaning are called *Surface Changes*. They further divide Surface Changes into *Formal Changes* (mostly copy-edits like spelling corrections etc.) and *Meaning-Preserving Changes* (paraphrases). Text-Base Changes are split into *Microstructure* and *Macrostructure Changes*, where the former describe minor changes and the latter refer to changes that affect the summary or gist of the entire text. Meaning-Preserving Changes, Microstructure Changes and Macrostructure Changes are further divided into Additions, Deletions, Substitutions, Permutations, Distributions and Consolidations. Various studies have classified edits in Wikipedia; we compare them in Table 1.

Pfeil et al. (2006) propose a taxonomy of 13 categories, aiming to compare cultural differences in the writing process of one article in four language versions of Wikipedia (German, Dutch, French and Japanese). Their taxonomy is based on an analysis of the data, not on existing revision theories. Two annotators manually examined and labeled the 500 revision pairs in their corpus. They allowed for multi-labeling and resolved disagreement by discussion. No inter-annotator agreement is reported.

Jones (2008) analyzes differences in the collaborative writing process of featured and non-featured articles in Wikipedia. His taxonomy is based on Faigley and Witte's (1981) distinction between Macrostructure and Microstructure changes. His corpus consists of 10 Wikipedia articles which were nominated to be featured in January 2007, from which 5 were actually promoted and the other 5 were denied the featured article status. For the annotation process, he relies on revision comments that have been generated either by the authors or automatically,

	Pfeil et al. (2006)	**Jones (2008)**	**Liu and Ram (2011)**
Wikipedia Policy	Vandalism	Vandalism	Revert
	Reversion	Revert	
		Disambiguation	
Text-Base	Add Information	Significant addition	Sentence creation
	Delete Information	Significant deletion	Sentence deletion
	Clarify Information	Structural change	Sentence modification[a]
	Add Link	Add image	Link creation
	Delete Link	Fix or delete image	Link deletion
	Fix Link	Add link	Link modification
		Fix or delete link	Reference creation
			Reference deletion
			Reference modification
Surface	Style/Typography	Style or readability	
	Spelling		
	Grammar		
	Format		
	Mark-up Language		

[a] As Liu and Ram (2011) state, this category includes grammar and spelling changes. Hence, it is not entirely a Text-Base category.

Table 1: Three studies classifying revisions in Wikipedia and the categories they use.

not on the actual revision texts. As only one person annotated the corpus, no inter-annotator agreement is reported.

Liu and Ram (2011) study the relationship between collaboration and article quality in Wikipedia, aiming to identify types of authors (e.g. Starter, Copy Editors, All-round contributors). Their taxonomy builds on Pfeil et al. (2006). However, they transformed the taxonomy into higher-level categories, merging clarification and grammar-spelling into sentence modification. By doing this, they are able to automatically identify edit categories; however, because the annotation of edits is not done manually, no inter-annotator agreement can be reported. While the automatic identification of edit categories allows for analyzing a larger corpus, it blurs Faigley and Witte's (1981) distinction between Surface and Text-Base changes. The authors do not report on the quality of the automatic edit category identification. Their corpus consists of 1,600 English Wikipedia articles from March 2010, divided into each 400 articles which have been nominated as either featured, good, B- or C-class according the WikiProject article quality grading scheme[4]. Using these four kinds of Wikipedia-internal evaluated quality grades, Liu and Ram (2011) study the relationship between collaboration and article quality. For their analysis, the authors used only the revisions before the respective nomination of the articles. The novel contribution of their work is that they calculate edits with a higher granularity (sentence level). Each (r_{v-1}, r_v)-pair may be multi-labeled with a list of edit categories to reflect the number of edits.

Other approaches propose special purpose classification systems for Wikipedia edits. Among the latter, Chin et al. (2010) focus on vandalism classification. Their top-level categories are Revert, Delete, Insert and Change; their system cannot easily be compared to the aforementioned systems, which distinguish between Text-Base and Surface changes. They introduce a basic

[4] `http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment` (accessed 2012-10-29)

distinction between content and format changes. Content includes text, links and images, format refers to HTML/CSS and templates. Fong and Biuk-Aghai (2010) present a system to automatically calculate and categorize edits. Similar to our system, their system computes a list of basic edit actions on the unparsed source text (i.e. including markup). Edits are calculated with granularity at the sentence and token level. The authors suggest a set of rules and categories to label the basic edit actions calculated before. Examples of their categories are (De)Wikify, Content Modification or Spelling Correction. The implementation and evaluation of their system is rather preliminary. Bronner and Monz (2012) distinguish between Factual and Fluency edits. They segment adjacent revisions into edits and classify them in a supervised machine learning system. Their system successfully classifies edits into Factual and Fluency edits with a maximum accuracy of 0.88.

Except for Bronner and Monz (2012), all of the above presented annotation studies label pairs of adjacent revisions, not edits. Hence, even if multi-labeling is applied, it is not possible to reassign each local edit $e^k_{v-1,v}$ with a category from the set or list of categories assigned to the (r_{v-1}, r_v)-pair. No manual annotation study which explicitly analyzes the agreement between raters has been carried out so far. Hence, the reliability of previous annotations is unclear. We address these issues as we annotate a set of edits rather than revisions. Furthermore, we evaluated our annotation study with a detailed inter-annotator agreement and error analysis.

3 Proposed Edit Classification

3.1 Classification Scheme

Our approach of classifying edits in Wikipedia builds upon previous work on document revision classification (Faigley and Witte, 1981) and studies about edits in Wikipedia (Pfeil et al., 2006; Jones, 2008; Liu and Ram, 2011). We follow Faigley and Witte (1981) and define the top level layers *Surface* and *Text-Base*, which differentiate between meaning-preserving and meaning-changing edits. However, contrary to Faigley and Witte (1981), we do consider all deletions and insertions of text as Text-Base changes. The only categories for textual edits in the Surface layer are PARAPHRASE and RELOCATION, cf. Table 2. To keep the taxonomy manageable, we do not follow Faigley and Witte (1981) in their fine-grained distinction of textual edits in Additions, Deletions, Substitutions, Permutations, Distributions and Consolidations. Our taxonomy is hierarchical with the three top layers *Wikipedia Policy*, *Surface* and *Text-Base*. Table 2 presents a short explanation and example for each category.

VANDALISM and REVERT are edit categories related to **Wikipedia Policies**. We define VANDALISM as an edit deliberately compromising Wikipedia's integrity (Adler et al., 2011). A REVERT undoes past edits by restoring previous revisions or parts of them (Flöck et al., 2012). As for the **Surface** layer, we include changes to the markup, as well as relocations, spelling and grammar corrections and paraphrases. We define all elements related to the Wiki markup language (see the examples in Table 2) as MARKUP. This includes HTML code, which can also be used in Wikipedia to render the layout of a page. The RELOCATION category is assigned to edits which move entire lines (copy-paste). We use the SPELLING/GRAMMAR category to label corrections of spelling or grammatical errors. Edits which rephrase or paraphrase words or sentences without altering their meaning, are labeled with the PARAPHRASE category. In the **Text-Base** layer, we define the INFORMATION category which labels meaning-changing edits to the text itself. We use the FILE category to label edits related to media types like images, videos or audio files. The REFERENCES category is assigned to edits affecting internal and external links as well as

Category	**Description**	r_{v-1}	r_v
Wikipedia Policy *Invalid edits as defined by internal Wikipedia Policies and respective defense mechanisms*			
Vandalism	Edits deliberately compromising Wikipedia's integrity	Einstein's **key** insight was	Einstein's **cheese master** insight was
Revert	Edits restoring a previous state of a page	**Hahahahhahahahahahahaha:)**	
Surface *Edits not affecting the meaning of the text*			
Paraphrase	Textual edits paraphrasing words or sentences	denominations **like** the	denominations **such as** the
Spelling/Grammar	Edits correcting spelling or grammatical errors	in **the** Ireland	in Ireland
Relocation	Edits moving entire lines[a]	**... -->** {{ChineseText}} **[[Category ...**	**...Dynasty}}** {{ChineseText}} ***[[Chinese...**
Markup-Insert	Edits affecting markup segments	an infant Parasaurolophus	an infant "Parasaurolophus"
Markup-Delete		'"AcDec"'	"AcDec"
Markup-Modify		=== The geometry of gravitation ===	== The geometry of gravitation ==
Text-Base *Edits affecting the meaning of the text*			
Information-Insert	Textual edits affecting information content	effects were tested by	effects were **both** tested by
Information-Delete		it is **not** a sacrament	it is a sacrament
Information-Modify		open steppes of Kashmir and **Siberia**.	open steppes of Kashmir and **Manchuria**.
File-Insert	Edits affecting files (media content)		**[[File:Dholeskull.jpg|...]]**
File-Delete		**[[Image:Victoria_Cross_bar.JPG|...]]**	
File-Modify		{{Infobox...|image=UMCLogo.**svg**... }}	{{Infobox...|image=UMCLogo.**xml**... }}
Reference-Insert	Edits affecting links, interwiki/language links or bibliographical references and citations		**[[sl:Erlang]]**
Reference-Delete		**[[**spondee**]]**	spondee
Reference-Modify		[[**molar**]]	**[[molar (tooth)**|molar]]
Template-Insert	Edits affecting templates (for including text from other pages, automatic text generation etc.)		**{{cite book|... }}**
Template-Delete		**{{clear}}**	
Template-Modify		{{**Unit** m|2|1}}	{{**Convert**|2|m|**ft**|1}}
Other	*Segmentation Errors*	[[pirate**s]]**	[[pirate**]]s**

[a] To increase readability, line breaks are omitted in the examples.

Table 2: Classification of Wikipedia edits with truncated examples from our corpus. Where necessary, we added the context, while the actual edit is bold-faced in r_{v-1} and/or r_v.

bibliographical citations. Different from Liu and Ram (2011), we do not distinguish between links and citations, as these edits refer to the same action in the sense of referencing something. Finally, the TEMPLATE category labels all edits related to templates. In Wikipedia, templates are indicated by double curly brackets and are used for including text from other pages, creating standardized messages or other automated text generation tasks.

All Text-Base edits and those in the MARKUP category are further divided into Insertions (I), Deletions (D) and Modifications (M). Insertions apply when new content or markup is added to the article, i.e. if the content or markup of $e^k_{v-1,v}$ has not been present in r_{v-1} but is present in r_v. Correspondingly, deletions remove the content or markup of $e^k_{v-1,v}$, so that the text that has been present in r_{v-1} is not present in r_v. Modifications apply to content and markup belonging to the same segment which has been changed from r_{v-1} to r_v. Here, we define a *segment* as the source element which is affected by the category of the respective edit, e.g. for modifications of the MARKUP, a markup element must be changed, for FILE edits, the embedded file must be changed etc. Correspondingly, a TEMPLATE-M edit must change the type of the template (i.e. its name) and not just a parameter of the template, as indicated in the respective example in Table 2.

We classify changes to the source text of a Wiki page, as opposed to the visual changes on the pages surface, i.e. the translated HTML which is displayed in the browser. We believe this yields a more accurate analysis of the writing process itself. Our taxonomy is geared toward edits in Wikis; however, it is fully language independent.

3.2 Corpus Construction

To draw conclusions about the relationship between the writing process and article quality, we determine distinguished articles based on the *featured* label[5] as defined by the Wikipedia community (Stvilia et al., 2008). Wikipedia has an internal review system to label articles that meet certain predefined quality criteria, e.g. they should be comprehensive, contain images where appropriate etc. The highest status an article can achieve is the featured status. Kittur and Kraut (2008) validated a set of articles with ratings from external users and found that the agreement between the external ratings and the internal ratings according to the WikiProject article quality grading scheme is substantial. For each featured articles (FA) in the English Wikipedia, we selected a non-featured article (NFA) with equal character length. From these article pairs, we randomly selected 10 pairs with equal or almost equal edit frequency (i.e. number of revisions per day) from different size ranges (see Table 3). Although we can assume that the FAs in our corpus have high quality, the NFAs show a broad quality spectrum according to the ratings by the WikiProjects' quality assessment teams, ranging from Start- to Good-class articles. However, none of the NFAs have been rated with the highest quality scores, namely featured or A-class. The selected articles cover a range of topics on historical, scientific and political issues. The youngest article is almost 6 years, the oldest is more than 9. We call the result *Wikipedia Quality Assessment Corpus* (WPQAC).

***Pre* and *Post* Revision Groups** From these article pairs, we selected 891 revisions containing 1,995 edits for the annotation study. From the FAs, we took the revision at the time of promotion to featured status (referred to as r_{prom}) specified on the respective Talk page as the reference and divided the article history into a *pre* and a *post* stage. *Pre* denotes all revisions made

[5] http://en.wikipedia.org/wiki/Wikipedia:Featured_articles (accessed 2012-10-29)

FA	NFA	Size	Freq.
1941 Atlantic hurricane season	Dactylic hexameter	18	0.1
William de Corbeil	European Liberal Democrat and Reform Party	26	0.1
Victoria Cross (Canada)	Erlang (programming language)	27	0.2
Deinosuchus	Intel 8086	32	0.2
Winfield Scott Hancock	Dhole	44	0.2
Laplace-Runge-Lenz vector	United Nations Relief and Works Agency	63	0.2
Introduction to general relativity	Subwoofer	70	0.4
United States Academic Decathlon	John Cage	78	0.5
Song Dynasty	Haile Selassie I	106	1.1
Euclidean algorithm	United Methodist Church	109	0.5

Table 3: The size of the latest revision (in 1,000 characters including Wiki markup) and edit frequency (average number of revisions per day) in WPQAC are equal for each FA-NFA pair.

Group	N_e	N_r	N_e/N_r
pre-FA	515	234	2.2
post-FA	485	144	3.4
pre-NFA	496	256	1.9
post-NFA	499	257	1.9
all	1995	891	2.2

Table 4: Revision groups in the annotated part of WPQAC with absolute numbers of edits and revisions.

previously to r_{prom} and *post* all revisions made after r_{prom}. Then, for each of the ten article pairs, we selected approximately 200 edits, namely each 50 edits from (r_{v-1}, r_v)-pairs

- in the second quarter of the pre stage of the FA article history (**pre-FA**),
- in the second half of the post stage of the FA article history (**post-FA**),
- in a pre-FA parallel stage in the NFA article history (**pre-NFA**),
- in a post-FA parallel stage in the NFA article history (**post-NFA**).

This way, we ensure that pre and post stage are comparable for all article pairs in our corpus with respect to the date of promotion of the FA. The annotated corpus is therefore split into four groups, with about 500 edits each, see Table 4. Slight differences in the sizes of the groups result from the fact that we had to choose adjacent revisions for each article and stage. These revisions contain diverging numbers of edits which did not always sum up to precisely 50.

The corpus has been selected to reflect the entire range of possible edits in Wikipedia, including bot edits, vandalism and reverts. Hence, no further filtering is done.

Edit segmentation The raw data for our corpus is extracted from the English Wikipedia Revision History, from the dump as of April 2011. We process the revision content (text with markup) using the Wikipedia Revision Toolkit (Ferschke et al., 2011). We do not parse the revision text, as we want to include both edits affecting the content and edits affecting the layout into one taxonomy. For each (r_{v-1}, r_v)-pair, we calculate all of the n changes $e^k_{v-1,v}$ that have been made to the current revision via an adapted version of the diff comparison algorithm by Heckel (1978). The algorithm splits each revision into its lines and numbers them. Then, it compares each line in r_{v-1} with each line in r_v to find differences in terms of inserted, deleted, modified and relocated lines. Although we only work with data from the English Wikipedia in this study, the segmentation process is fully language independent.

Inside modified lines, we additionally detect and mark changes (i.e. deletions, insertions and modifications) in situ using Neil Fraser's google-diff-match-patch library[6]. The last step is only done where the ratio of the number of overall changes in that line to the number of tokens in that line does not exceed a certain threshold. The latter serves to avoid splitting heavily edited lines into a very high number of counterintuitive edits. If, for example, stopwords like "the" or "a" are the only unchanged segments inside a modified line, we want the entire line to be marked as modified. We do further post-processing to recognize and merge associated edits, e.g. when adding a link (to merge `[[` and `]]`). This may yield errors as Wiki markup is a context-sensitive language and hence difficult to parse. In the manual annotation study, we annotate segmentation errors due to associated edits which have not been detected and merged by our algorithm with the OTHER category (cf. Table 2).

Our annotation study is carried out on edits as calculated by the segmentation algorithm explained above. The basic types of edits which the algorithm detects are insertions, deletions, modifications and relocations. Correspondingly, each (r_{v-1}, r_v)-pair can create more than one object to classify, depending on the number of edits it contains. Our annotated corpus consists of $N_r = 891$ revisions containing $N_e = 1,995$ edits. The median of edits per revision is 1, the standard deviation is 14.5 with a minimum of 1 and a maximum of 55 edits per revision. That is, most of the changes in our corpus modify articles in only one particular place.

3.3 Annotation Study

We employed three non-native speakers with working knowledge of the Wikipedia policies and markup to label the corpus based on written annotation guidelines. We define the annotation task as a multi-label classification, i.e. each $e^k_{v-1,v}$ calculated from a (r_{v-1}, r_v)-pair is assigned a set of categories $Y \subset L$, where L is the set of categories as defined in Table 2 (hence $|L| = 21$ and $|Y| \geq 1$). If, for example, an entire sentence is rewritten, this might not only affect the words but also the markup (e.g. when a bold-faced word is deleted) or references (e.g. when a link is added). Such an edit would be multi-labeled with INFORMATION-M and MARKUP-D or REFERENCE-I respectively. Further guidelines include the following:

- Edits labeled as VANDALISM, REVERT, RELOCATION or OTHER cannot be multi-labeled.
- If $e^k_{v-1,v}$ is labeled as VANDALISM, all $e^0_{v-1,v}, e^1_{v-1,v}, ...e^n_{v-1,v}$ must be labeled as VANDALISM, since all of those edits have the same author (with bad intentions).
- Edits removing or inserting white spaces or line breaks are labeled as MARKUP.

For the annotation of edits, we used the Apache UIMA[7] Cas Editor. That way, we were able to directly annotate on the source files which are produced by the UIMA pipeline we use to extract the raw text for each revision and to segment each (r_{v-1}, r_v)-pair into a list of edits. The annotators had access to all metadata information (author name, comment etc.) and the entire text of r_{v-1} and r_v.

We derive the gold standard annotations by means of a majority vote for each category. That means, for each $e^k_{v-1,v}$ which has been labeled with $l \in L$ by at least 2 annotators, we assign the category l in the gold standard. If all 3 annotators disagreed, i.e. if an edit was labeled with none of the categories at least 2 times, it is assigned the OTHER category in the gold standard. For example, one edit changed "...algorithm *will* not terminate..." to "...algorithm *does* not

[6] `http://code.google.com/p/google-diff-match-patch/` (accessed 2012-10-29)

[7] Unstructured Information Management System, `http://uima.apache.org/` (accessed 2012-10-29)

terminate...”. One annotator labeled this edit as PARAPHRASE, the other one as INFORMATION-M and the third one as SPELLING/GRAMMAR. We observed this kind of total disagreement in 5.7% of all edits. The gold standard annotations have not been manually corrected subsequently.

Inter-annotator Agreement To estimate the reliability of the annotations, we compute the inter-annotator agreement per category using the multi-rater Kappa κ measure (Fleiss, 1971), see Table 5. For each edit, the proportion of agreeing votes (i.e. judgment pairs) out of the total number of pairs is calculated. With regard to the overall agreement, we need an appropriate agreement measure for multiple raters and multi-labeled edits. We employ Krippendorff’s Alpha (Krippendorff, 1980) with a set-valued distance function, MASI (Passonneau, 2006). For each edit, we have a set of categories and consider the possibly partial agreement in the assigned category sets. The overall agreement in terms of Krippendorff’s Alpha is $\alpha = 0.67$. This is at the lower boundary of what is usually considered to allow for drawing tentative conclusions (Krippendorff, 1980). To the best of our knowledge, no annotation study based on edit categories in Wikipedia has been carried out, hence, this value is hard to judge as we cannot compare it to other studies. We discuss the κ values across categories below (cf. Error Analysis).

Edit- vs. Revision-based Category Distribution To measure the absolute number of revisions labeled with a certain category C_r, we built the set of edit categories over all $e^k_{v-1,v}$ in each (r_{v-1}, r_v)-pair. When comparing the absolute number of edits labeled with a certain category C_e to C_r in Table 5, we observe that the MARKUP-D, SPELLING/GRAMMAR and PARAPHRASE categories have on average the highest number of edits per revision (more than two). All of them belong to the Surface layer, whereas many of the Text-Base edits (e.g. FILE, REFERENCE) show a lower ratio of edits per revision. This might be due to the fact that authors carrying out copy-edit changes have a focus on the entire article and change the text in various places which results in a higher number of edits. To the contrary, Text-Base edits may have a focus on a limited part of the article and hence edit in only one place. Furthermore, we could conclude that authors changing the article’s text base save their edits more often, as this creates a higher number of revisions.

Single- vs. Multi-label Annotation Almost 15% of the edits are multi-labeled, and more than 30% of all revisions are multi-labeled. This shows that a lot of information would be lost if we opted against a multi-label annotation. The label cardinality, i.e. the average number of assigned categories per edit, cf. Tsoumakas et al. (2010), is $LC = \frac{1}{|D|}\sum_{i=1}^{|D|}|Y_i| = 1.2$ and the label density, i.e. the average fraction of assigned categories per edit, is $LD = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i|}{|L|} = 0.06$, where D denotes our data set.

Error Analysis We turned the multi-labeled data into single-labeled data by transforming each unique category set which has been assigned to one of the edits into a new category $t \in T$. In our corpus, $|T| = 90$. Tsoumakas et al. (2010) refer to this transformation method as *Label Powerset*, as $T \subseteq P(L)$. We created and analyzed confusion matrices over the unique category sets for each annotator with respect to the gold standard. About 25% of all disagreement in terms of confused categories is due to edits which are labeled with the OTHER category in the gold standard. This is partly related to the fact that we labeled edits where all 3 annotators disagreed with the OTHER category in the gold standard. Furthermore, this category is not

Label	κ	P_O	**Edits**		**Revisions**		C_e			
			C_e	%	C_r	%	**pre-FA**	**post-FA**	**pre-NFA**	**post-NFA**
Information-I	0.64	0.91	280	11.67	200	13.11	71	59	81	69
Reference-I	0.79	0.95	262	10.92	209	13.70	59	37	87	79
Revert	0.83	0.96	254	10.59	128	8.39	66	55	50	83
Information-M	0.58	0.90	237	9.88	145	9.50	62	40	72	63
Markup-I	0.61	0.92	223	9.30	133	8.72	50	54	80	39
Vandalism	0.69	0.95	163	6.79	98	6.42	50	28	43	42
Spelling/Grammar	0.73	0.96	161	6.71	80	5.24	32	75	30	24
Information-D	0.55	0.93	139	5.79	80	5.24	54	32	22	31
Other[a]	0.18	0.97	139	5.79	86	5.64	42	36	26	35
Markup-D	0.58	0.95	131	5.46	59	3.87	22	60	23	26
Reference-D	0.68	0.97	88	3.67	66	4.33	35	6	24	23
Reference-M	0.54	0.96	88	3.67	78	5.11	24	8	30	26
Template-I	0.78	0.99	72	3.00	62	4.06	27	20	5	20
Paraphrase	0.31	0.96	54	2.25	24	1.57	6	12	7	29
Relocation	0.71	0.99	29	1.21	17	1.11	6	2	17	4
Template-D	0.66	0.99	26	1.08	20	1.31	13	5	1	7
Markup-M	0.25	0.97	17	0.71	13	0.85	8	2	6	1
Template-M	0.73	0.99	17	0.71	9	0.59	9	3	0	5
File-I	0.78	0.997	13	0.54	13	0.85	5	3	4	1
File-D	0.72	0.998	5	0.21	5	0.33	2	1	2	0
File-M	0.25	0.999	1	0.04	1	0.07	0	0	0	1
Text-Base	0.66	0.83	1228	51.19	888	58.19	361	214	328	325
Surface	0.61	0.83	615	25.64	326	21.36	124	205	163	123
Wikipedia Policy	0.79	0.93	417	17.38	226	14.81	116	83	93	125
All	—	—	2399	100	1526	100	643	538	610	608

[a] Excluded from top level categories. For that reason, percentages in the bottom rows do not sum up to 100%.

Table 5: Inter-annotator agreement, where κ is Fleiss' Kappa per category/layer and P_O the observed agreement per category/layer. C_e resp. C_r and % are the absolute numbers and percentages of edits resp. revisions labeled with a certain category in the gold standard.

well-defined. Further categories with low agreement are Paraphrase, File-M and Markup-M (cf. Table 5). File-M occurred only once in the gold standard. More than 40% of cases of disagreement involving Markup-M are labeled as Other in the gold standard, either due to segmentation errors (cf. Section 3.2), or because of general disagreement between all annotators. The Paraphrase category was not used consistently among the annotators and frequently confused with Information-M and Spelling/Grammar. Hence, the distinction between Paraphrase (non-meaning change) and Information-M (meaning change) has not been clear in many cases. For example, one edit replaced "several" with "many". Two annotators annotated this edit as Paraphrase, one as Information-M. A common problem in each of the categories was the distinction between insertions, modifications and deletions, particularly in the Information category. The annotators did not consequently adhere to the annotation guidelines (cf. Section 3.1) in some cases. If, for example, an edit *deletes* the word "not" in a phrase like "it is not a sacrament" (cf. Table 5), this edit also *changes* the meaning, which complicates the annotation of such edits.

One annotator labeled many instances of Markup-D as Information-D (9% of all cases of disagreement with respect to the gold standard annotations). Furthermore, one annotator frequently (8%) forgot to multi-label MARKUP-I when larger portions of text were inserted (e.g. INFORMATION-I, REFERENCE-I instead of INFORMATION-I, REFERENCE-I, MARKUP-I).

For future work, we recommend to ignore edits labeled with the OTHER category. Categories with low agreement such as PARAPHRASE and MARKUP-M should be used with a grain of salt.

4 Discussion

4.1 Edit Category Distribution

The category distribution in our corpus partly corresponds with that in Pfeil et al. (2006) for the French, German, Japanese and Dutch Wikipedia, cf. Table 1. Additions of INFORMATION and REFERENCES are the most frequent categories.[8] VANDALISM in our corpus accounts for about 7% of all edits, which confirms the findings of Potthast (2010). Insertions clearly outnumber modifications and deletions, consistent with the studies of Jones (2008) and Pfeil et al. (2006). These findings confirm that our annotated corpus is a representative sample with regard to the collaborative writing process in Wikipedia.

Jones (2008) quotes only around 3% of edits in his *Add link* category, as compared to 28% in Pfeil et al. (2006) and 11% in the REFERENCE-I category in our corpus. Despite the fact that the categories might not fully overlap in their definitions, the low number in Jones's (2008) study could be an indicator that his approach to label edit categories based on the authors' comments does not fully capture the extent of certain edits.

The high deviation of absolute numbers of VANDALISM edits and REVERTS in our corpus is surprising. Manual inspection of the data shows that there are some Reverts of REVERTS (so called edit wars). Also, when comparing C_r to C_e for REVERT and VANDALISM in Table 5, apparently the number of edits per revision is much higher for REVERTS than for VANDALISM. This might be a particularity in our corpus, but we could also assume that vandals usually change a small portion of text, e.g. by inserting a swear word. On the other hand, authors applying a REVERT might not only revert vandalism but also undo legitimate edits which do not conform with their point of view.

4.2 Collaborative Writing and Quality

We designed WPQAC as a corpus to study differences in the quality of FAs and NFAs. To gain insights into the writing process, we analyzed the category distributions for different revision groups (cf. Table 4). Table 6 shows the Pearson correlations over category distributions between relevant groups. These calculations are based on the category frequencies of multi-labeled edits (Table 5, column C_e) for the revision groups.

Over all categories, we can see significant ($p < 0.01$, using Student's t-test) correlations between all of the groups, i.e. the frequencies of types of edits do not show significant differences among the revision groups. Generally, FAs and NFAs show a relatively high correlation. However, the correlation for pre-FA and post-FA revisions is clearly lower, as compared to pre-NFA and post-NFA. To reduce possible noise, we excluded the smaller categories from the groups and calculated the same correlations only for categories used to label at least 20 edits, i.e. with $C_e \geq 20$. As indicated in Table 6, the correlations between the pre-FA and post-FA as well as

[8] Ignoring Pfeil's (2006) *Format* category, which has partial overlap with our MARKUP category.

Group	r (all)	r (Top-16)	r (Jones, 2008)	Correlation criteria
All	0.87*	0.80*	0.91*	FA/NFA
All	0.90*	0.84*	—	pre/post
FA	0.72*	0.57	0.68	pre/post
NFA	0.87*	0.81*	—	pre/post
pre	0.86*	0.80*	—	FA/NFA
post	0.68*	0.52	—	FA/NFA

Table 6: Pearson correlation r between frequency distributions of edit categories by revision group for all and for the 16 largest categories. For comparison, we added the corresponding numbers for Jones's (2008) study. Values marked with * are statistically significant for $p < 0.01$.

post-FA and post-NFA are not statistically significant when calculated for the top 16 categories, i.e. we can assume that the two distributions come from different samples.

For the SPELLING/GRAMMAR and REFERENCE categories, deviances between the absolute number of edits in FAs and NFAs are particularly high (see Table 5). This is mainly because post-FA revisions show a higher number of SPELLING/GRAMMAR corrections and a lower number of REFERENCE edits as compared to pre-FA and NFAs. Improvements of style and grammar or spelling corrections are essential edits to produce thorough and high-quality content, hence, the higher number of this type of edits in post-FA revisions might be the result of the increased attention by experienced Wikipedia authors (Liu and Ram, 2011). The lower number of REFERENCE edits in post-FA revisions is not very surprising, as FAs need to be "well-researched", i.e. "verifiable against [...] reliable sources" according to Wikipedia's FA criteria[9] and we assume that this is the case for post-FA revisions. The high number of MARKUP-D edits in the post-FA revision group is due to one particular (r_{v-1}, r_v)-pair which deleted 42 markup tags in various places across the entire revision text.

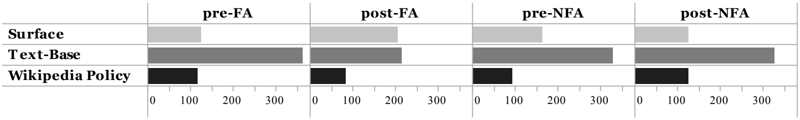

Figure 1: Absolute number of edits C_e for layers in revision groups.

It is not possible to verify the distinction between experienced and unexperienced authors as explained by Sommers (1980) for the collaborative writing process in Wikipedia. As can be seen in Table 5, the number of Surface respective Text-Base edits is higher respective lower for FAs compared to NFAs. This might be due to the fact that not only experienced authors work on FAs and vice versa.

The relationship between the distribution of edit types and quality has earlier been addressed by Jones (2008), who included in his corpus all FA revisions before and after their promotion. Like ours, his analysis shows a high correlation between FAs and NFAs, while pre-FA and post-FA differ significantly, cf. Table 6. Although it is hard to explain the reasons for this difference with his data, our corpus shows a clear difference in the ratio of Surface to Text-Base edits when

[9] http://en.wikipedia.org/w/index.php?title=Wikipedia:Featured_article_criteria&oldid=506642325

comparing post-FA revisions to pre-FA, pre-NFA and post-NFA revisions, cf. Table 5. Hence, even if we cannot find significant differences in the editing history of FAs and NFAs, there is a deviation in the collaborative writing process (in terms of editing behavior) before and after the promotion of FAs. The distinctive behavior of the post-FA revision group as compared to pre-FA and NFA revisions suggests that the nomination and promotion as FA triggers a distinguished type of collaboration. The collaborative writing process in post-FA revisions can be characterized through a relatively high number of surface edits (in particular, Spelling/Grammar corrections) and a low number of changes to the Text-Base. Figure 1 highlights the distinction between different revision groups. The lower number of Text-Base edits and the higher number of copy-edits in post-FA revisions can be interpreted as a sign of stability which FAs show after their promotion.

Conclusion

As explained in the above, there is a need for corpora to analyze the collaborative writing process. To address this problem, we introduced a classification scheme of edits established on previous work of the writing research. We applied this scheme to annotate a sample from the revision history of the English Wikipedia. To verify the reliability of our annotations, we measured and analyzed the inter-annotator agreement across categories. We published our corpus, providing free access to the research community. Furthermore, we compared the edit category distribution in featured and non-featured article revisions. Our findings show that featured articles differ from non-featured articles mainly because of a distinguished process of collaboration after an article achieved featured status. This collaboration process includes a higher number of surface changes and on the opposite a lower number of edits changing the meaning.

Further work should incorporate a deeper analysis of article quality and quality flaws in Wikipedia (Ferschke et al., 2012b). Since revisions in Wikipedia are accompanied by metadata and in particular, user comments, an analysis of the metadata based on edit categories might yield interesting results. Although we analyzed edit categories in the English Wikipedia, our approach (i.e. the classification scheme and the edit segmentation) can be applied to any language version of Wikipedia. Given that other language versions might use existing information in the English Wikipedia and translate it rather than creating completely new content (e.g. to keep the language versions with a smaller set of authors up-to-date), our taxonomy can also be used to distinguish between surface edits and edits which add new information, similar to the approach of Bronner and Monz (2012).

Finally, there remains a need for more data. Our assumptions have to be confirmed on a larger corpus. We will address this issue by augmenting the labeled corpus with an automated approach using Machine Learning on the annotated data.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the Hessian research excellence program “Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz” (LOEWE) as part of the research center “Digital Humanities”. We thank the anonymous reviewers for their helpful remarks.

References

Adler, B. T., Alfaro, L., Mola-Velasco, S. M., Rosso, P., and West, A. G. (2011). Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 277–288. Springer.

Bronner, A. and Monz, C. (2012). User Edits Classification Using Document Revision Histories. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 356–366, Avignon, France. Association for Computational Linguistics.

Chin, S.-C., Street, W. N., Srinivasan, P., and Eichmann, D. (2010). Detecting Wikipedia vandalism with active learning and statistical language models. In *Proceedings of the 4th workshop on Information credibility*, pages 3–10, Raleigh, NC, USA.

Ede, L. and Lunsford, A. (1990). *Singular Text/Plural Authors: Perspectives on Collaborative Writing*. Southern Illinois University Press.

Faigley, L. and Witte, S. (1981). Analyzing Revision. *College Composition and Communication*, 32(4):400.

Ferschke, O., Daxenberger, J., and Gurevych, I. (2012a). A Survey of NLP Methods and Resources for Analyzing the Collaborative Writing Process in Wikipedia. In Gurevych, I. and Kim, J., editors, *The People's Web Meets NLP: Collaboratively Constructed Language Resources*, Theory and Applications of Natural Language Processing, chapter 5. Springer, Heidelberg.

Ferschke, O., Gurevych, I., and Rittberger, M. (2012b). FlawFinder: A Modular System for Predicting Quality Flaws in Wikipedia. In *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers*, Rome, Italy.

Ferschke, O., Zesch, T., and Gurevych, I. (2011). Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia's Edit History. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. System Demonstrations*, pages 97–102, Portland, OR, USA.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Flöck, F., Vrandečić, D., and Simperl, E. (2012). Revisiting reverts: Accurate revert detection in Wikipedia. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 3–12, Milwaukee, WI, USA.

Fong, P. K.-F. and Biuk-Aghai, R. P. (2010). What did they do? Deriving high-level edit histories in Wikis. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, WikiSym '10, Gdansk, Poland.

Heckel, P. (1978). A technique for isolating differences between files. *Communications of the ACM*, 21(4):264–268.

Jones, J. (2008). Patterns of Revision in Online Writing: A Study of Wikipedia's Featured Articles. *Written Communication*, 25(2):262–289.

Kittur, A. and Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, CSCW '08, pages 37–46, San Diego, CA, USA.

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

Lee, J. and Webster, J. (2012). A Corpus of Textual Revisions in Second Language Writing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 248–252, Jeju Island, Republic of Korea.

Liu, J. and Ram, S. (2011). Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Transactions on Management Information Systems*, 2(2).

Noel, S. and Robert, J.-M. (2004). Empirical Study on Collaborative Writing: What Do Co-authors Do, Use, and Like? *Computer Supported Cooperative Work*, 13(1):63–89.

Passonneau, R. (2006). Measuring Agreement on Set-valued Items (MASI) for Semantic and Pragmatic Annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

Pfeil, U., Zaphiris, P., and Ang, C. S. (2006). Cultural Differences in Collaborative Authoring of Wikipedia. *Journal of Computer-Mediated Communication*, 12(1):88–113.

Potthast, M. (2010). Crowdsourcing a Wikipedia Vandalism Corpus. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research & Development on Information Retrieval*, pages 789–790, Geneva, Switzerland.

Sommers, N. (1980). Revision strategies of student writers and experienced adult writers. *College composition and communication*, 31(4):378–388.

Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. (2008). Information quality work organization in wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001.

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2010). Mining Multi-label Data. In Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*, chapter 34, pages 667–685. Springer.

Woodsend, K. and Lapata, M. (2011). Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK.

Zesch, T. (2012). Measuring Contextual Fitness Using Error Contexts Extracted from the Wikipedia Revision History. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 529–538, Avignon, France.