The Writing Process in Online Mass Collaboration

NLP-Supported Approaches to Analyzing Collaborative Revision and User Interaction

Vom Fachbereich Informatik
der Technischen Universität Darmstadt
genehmigte

Dissertation

zur Erlangung des akademischen Grades Dr.-Ing.

vorgelegt von
Johannes Daxenberger, M.A.
geboren in Rosenheim

Tag der Einreichung: 28. Mai 2015
Tag der Disputation: 21. Juli 2015

Referenten: Prof. Dr. Iryna Gurevych, Darmstadt
Prof. Dr. Karsten Weihe, Darmstadt
Assoc. Prof. Ofer Arazy, Ph.D., Alberta

Darmstadt 2016
D17

Abstract

In the past 15 years, the rapid development of web technologies has created novel ways of collaborative editing. Open online platforms have attracted millions of users from all over the world. The open encyclopedia Wikipedia, started in 2001, has become a very prominent example of a largely successful platform for collaborative editing and knowledge creation. The wiki model has enabled collaboration at a new scale, with more than 30,000 monthly active users on the English Wikipedia.

Traditional writing research deals with questions concerning revision and the writing process itself. The analysis of collaborative writing additionally raises questions about the interaction of the involved authors. Interaction takes place when authors write on the same document (indirect interaction), or when they coordinate the collaborative writing process by means of communication (direct interaction). The study of collaborative writing in online mass collaboration poses several interesting challenges. First and foremost, the writing process in open online collaboration is typically characterized by a large number of revisions from many different authors. Therefore, it is important to understand the interplay and the sequences of different revision categories. As the quality of documents produced in a collaborative writing process varies greatly, the relationship between collaborative revision and document quality is an important field of study. Furthermore, the impact of direct user interaction through background discussions on the collaborative writing process is largely unknown. In this thesis, we tackle these challenges in the context of online mass collaboration, using one of the largest collaboratively created resources, Wikipedia, as our data source. We will also discuss to which extent our conclusions are valid beyond Wikipedia.

We will be dealing with three aspects of collaborative writing in Wikipedia. First, we carry out a content-oriented analysis of revisions in the Wikipedia revision history. This includes the segmentation of article revisions into human-interpretable edits. We develop a taxonomy of edit categories such as spelling error corrections, vandalism or information adding, and verify our taxonomy in an annotation study on a corpus of edits from the English and German Wikipedia. We use the annotated corpora as training data to create

models which enable the automatic classification of edits. To show that our model is able to generalize beyond our own data, we train and test it on a second corpus of English Wikipedia revisions. We analyze the distribution of edit categories and frequent patterns in edit sequences within a larger set of article revisions. We also assess the relationship between edit categories and article quality, finding that the information content in high-quality articles tends to become more stable after their promotion and that high-quality articles show a higher degree of homogeneity with respect to frequent collaboration patterns as compared to random articles.

Second, we investigate activity-based roles of users in Wikipedia and how they relate to the collaborative writing process. We automatically classify all revisions in a representative sample of Wikipedia articles and cluster users in this sample into seven intuitive roles. The roles are based on the editing behavior of the users. We find roles such as Vandals, Watchdogs, or All-round Contributors. We also analyze the stability of our discovered roles across time and analyze role transitions. The results show that although the nature of roles remains stable across time, more than half of the users in our sample changed their role between two time periods.

Third, we analyze the correspondence between indirect user interaction through collaborative editing and direct user interaction through background discussion. We analyze direct user interaction using the notion of turns, which has been established in previous work. Turns are snippets from Wikipedia discussion pages. We introduce the notion of corresponding edit-turn-pairs. A corresponding edit-turn-pair consists of a turn and an edit from the same Wikipedia article; the turn forms an explicit performative and the edit corresponds to this performative. This happens, for example, when a user complains about a missing reference in the discussion about an article, and another user adds an appropriate reference to the article itself. We identify the distinctive properties of corresponding edit-turn-pairs and use them to create a model for the automatic detection of corresponding and non-corresponding edit-turn-pairs. We show that the percentage of corresponding edit-turn-pairs in a corpus of flawed English Wikipedia articles is typically below 5% and varies considerably across different articles.

The thesis is concluded with a summary of our main contributions and findings. The growing number of collaborative platforms in commercial applications and education, e.g. in massive open online learning courses, demonstrates the need to understand the collaborative writing process and to support collaborating authors. We also discuss several open issues with respect to the questions addressed in the main parts of the thesis and point out possible directions for future work. Many of the experiments we carried out in the course of this thesis rely on supervised text classification. In the appendix, we explain the concepts and technologies underlying these experiments. We also introduce the DKPro TC framework, which was substantially extended as part of this thesis.

Zusammenfassung

Die Weiterentwicklung von Webtechnologien in den vergangenen 15 Jahren hat vollkommen neue Formen gemeinschaftlichen Schreibens im Web hervorgebracht. Open-Access Online-Plattformen haben Millionen Benutzer, die über die gesamte Erde verteilt sind. Die Online-Enzyklopädie Wikipedia, gegründet im Jahr 2001, hat sich zu einer der bekanntesten und erfolgreichsten Plattformen für gemeinschaftliches Schreiben und Wissensgenerierung entwickelt. Das Wiki-Modell macht Zusammenarbeit in einer neuen Dimension möglich, so dass bspw. in der englischen Wikipedia jeden Monat mehr als 30.000 Benutzer aktiv sind.

Die traditionelle Schreibforschung setzt sich mit Fragen über Revision und den Schreibprozess auseinander. Die Analyse gemeinschaftlichen Schreibens interessiert sich darüber hinaus für die Interaktion der beteiligten Benutzer. Solche Interaktion findet statt wenn Autoren am selben Dokument schreiben (indirekte Interaktion), oder wenn Autoren den gemeinschaftlichen Schreibprozess mittels mündlicher oder schriftlicher Kommunikation koordinieren (direkte Interaktion). Die Erforschung gemeinschaftlichen Schreibens unter massiver Zusammenarbeit auf Online-Plattformen beinhaltet mehrere interessante Herausforderungen.

Der gemeinschaftliche Schreibprozess im Web ist gekennzeichnet durch eine typischerweise sehr hohe Zahl von Änderungen, die von vielen verschiedenen Autoren stammen. Dementsprechend ist es unverzichtbar, den Zusammenhang und die Abfolge unterschiedlicher Revisionstypen zu verstehen. Da die inhaltliche Qualität der Dokumente, die unter Zusammenarbeit erstellt werden, sehr unterschiedlich ist, ist außerdem die Erforschung der Korrelation zwischen gemeinschaftlichen Änderungen und Dokumentqualität ein wichtiges Feld. Desweiteren ist der Einfluss direkter Benutzerinteraktion mittels Diskussionen im Hintergrund auf den gemeinschaftlichen Schreibprozess größtenteils unbekannt. In der vorliegenden Arbeit setzen wir uns mit diesen Herausforderungen im Kontext massiver Zusammenarbeit auf Online-Plattformen auseinander. Dabei verwenden wir Wikipedia, eine der größten gemeinschaftlich erstellen Online-Ressourcen, als Datengrundlage. Wir werden auch diskutieren, inwiefern unsere Erkenntnisse über Wikipedia hinaus Gültigkeit besitzen.

Drei Hauptaspekte gemeinschaftlichen Schreibens in Wikipedia stellen das Grundgerüst dieser Arbeit dar. Als erstes führen wir eine inhaltliche Analyse von Revisionstypen in der Wikipedia Versionsgeschichte durch, wozu auch die Segmentierung von Artikelrevisionen in kleinere Edits, die einfacher zu interpretieren sind, zählt. Wir entwickeln eine Taxonomie für Edittypen, die bspw. Rechtschreibkorrekturen, Vandalismus oder Ergänzungen von Information beinhaltet. Die Taxonomie wird getestet in einer Annotationsstudie auf Edits aus der englischen und der deutschen Wikipedia. Wir verwenden die annotierten Korpora als Trainingsdaten zum Erstellen eines Modells für die automatischen Klassifikation von Edits. Um zu zeigen, dass dieses Modell auch in der Lage ist, über unsere eigenen Daten hinaus zu generalisieren, trainieren und testen wir es zusätzlich auf einem zweiten Korpus, das englische Wikipedia Revisionen annotiert. Wir analysieren die Verteilung der Edittypen sowie häufig auftretende Muster in Editsequenzen auf einer größeren Menge von Artikelrevisionen. Außerdem untersuchen wir den Zusammenhang zwischen Edittypen und Artikelqualität. Das Ergebnis zeigt, dass der Informationsgehalt in hochqualitativen Wikipedia Artikeln tendentiell stabiler wird sobald die Artikel ausgezeichnet werden. Ebenfalls zeigen hoch-qualitative Artikel im Vergleich zu zufällig gewählten Artikeln eine gesteigerte Homogenität mit Bezug auf häufig auftretende Editsequenzen.

Als zweites untersuchen wir auf Benutzeraktivität basierende Rollen und deren Zusammenhang mit dem gemeinschaftlichen Schreibprozess in Wikipedia. Dazu klassifizieren wir sämtliche Revisionen auf einem repräsentativen Teil der englischen Wikipedia und clustern deren Autoren in sieben interpretierbare Rollen, die das Editierverhalten der Autoren widerspiegeln. Wir identifizieren bspw. die Rollen von Vandalen, All-round Contributors oder Watchdogs. Außerdem untersuchen wir die Stabilität der Rollen über Zeiträume hinweg und analysieren Übergänge einzelner Benutzer in andere Rollen. Die Ergebnisse zeigen, dass die Beschaffenheit der Rollen über zwei Zeiträume hinweg stabil ist, allerdings wechseln im Laufe der Zeit mehr als die Hälfte der Benutzer ihre Rolle.

Als drittes untersuchen wir den Zusammenhang zwischen direkter Benutzerinteraktion mittels gemeinschaftlichem Editieren und indirekter Benutzerinteraktion mittels Diskussion im Hintergrund. Dabei analysieren wir direkte Interaktion mit Hilfe des Konzepts von Turns, welches aus Vorarbeiten stammt. Turns sind kurze Ausschnitte aus Wikipedia Diskussionsseiten, auf denen basierend wir sogenannte übereinstimmende Edit-Turn-Paare definieren. Ein übereinstimmendes Edit-Turn-Paar beinhaltet einen Turn und einen Edit von derselben Wikipedia Seite, dabei stellt der Turn einen expliziten Performativ dar und der Edit führt diesen Performativ aus. Das passiert bspw. wenn sich ein Benutzer in der Diskussion eines Artikels über eine fehlende Referenz beschwert und ein weiterer Benutzer die entsprechende Referenz zum Artikel selbst hinzufügt. Wir identifizieren distinktive Merkmale übereinstimmender Edit-Turn-Paare und verwenden diese um ein Modell zum automatischen Auffinden von (nicht-)übereinstimmenden Edit-Turn-Paaren zu entwickeln. Dabei zeigen wir, dass der Prozentsatz übereinstimmender Paare in einem Korpus

bestehend aus englischen Wikipedia Artikeln mit Qualitätsmängeln typischerweise unter 5% liegt und von Artikel zu Artikel erheblich variiert.

Die Arbeit wird abgeschlossen von einer Zusammenfassung unserer wichtigsten Beiträge und Ergebnisse. Die wachsende Zahl von Plattformen für gemeinschaftliches Schreiben in kommerziellen Anwendungen und in der Bildung, bspw. durch Massive Open Online Learning Courses, verdeutlicht den Bedarf eines besseren Verständnisses gemeinschaftlicher Schreibprozesse sowie eines besseren Supports der beteiligten Autoren. Wir diskutieren auch die Punkte, die im Bezug auf die Forschungsfragen im Hauptteil dieser Arbeit noch offen geblieben sind und skizzieren mögliche Ansatzpunkte für zukünftige Forschung. Da ein Großteil der Experimente, die im Rahmen dieser Arbeit ausgeführt wurden, auf Verfahren der überwachten Textklassifikation zurückgreift, erläutern wir deren grundlegende Konzepte und Technologien im Appendix. Der Appendix enthält außerdem eine Einleitung in das DKPro TC Framework, das im Laufe dieser Arbeit substantiell erweitert wurde.

Acknowledgments

This work would not have been possible without the support I received from several people. First and foremost, I would like to express my gratitude to my supervisor, Iryna Gurevych, for her constant encouragement and investment into my research over the last years. Her support and guidance have substantially shaped this work. Furthermore, I am very thankful for the fruitful discussions with my co-advisor Ofer Arazy. His persistence and insight are reflected in several parts of this work, and I would like to thank him wholeheartedly for the support and advice to my research. I am also thankful for the provocative discussions and advice by my second reviewer, Karsten Weihe.

This work has also been shaped by the inspiring conversations with my colleagues at UKP Lab (in particular during coffee breaks). I particularly want to mention Christian M. Meyer, Torsten Zesch, and Ivan Habernal, who were always asking the right questions, Richard Eckhart de Castilho, who taught me what programming is really about, and Nicolai Erbs, who was a reliable partner for a short snack break in the afternoon. My work has benefitted substantially from the collaboration with Oliver Ferschke, who has been of inspiration to much of my own research on Wikipedia. I am also thankful for the support of Emily Jamison who organized the crowd-sourcing study on Wikipedia edit-turn-pairs.

I would like to thank Lisa Beinborn, Nicolai Erbs, Oliver Ferschke, and Christian M. Meyer for their feedback on earlier versions of this work. The typesetting and layout of this thesis have originally been proposed by Christian M. Meyer, and further developed by Oliver Ferschke. This work has been supported by the Hessian research excellence program "Landesoffensive zur Entwicklung Wissenschaftlich-Ökonomischer Exzellenz" (LOEWE) as part of the research center "Digital Humanities".

I am most thankful to my family, who has given immense support to the non-scientific part of my PhD career. My parents have encouraged and enabled me to pursue an education in the field of my choice. And last but not least I owe most gratitude to my wife who is the best person to share my life with. All glory to God, source of life, hope, and inspiration.

Contents

Chapter 1

Introduction

The creation of a body of literature is a unique feature of humankind. Written artifacts (referred to as documents throughout this thesis), like other human-created artifacts, are typically the product of a complex process involving a substantial amount of knowledge, thought, and creativity. Most of the literature of humankind is made up of complex documents such as books, whose creation often takes many years. Teaching students to compose well-formed documents requires effort beyond teaching to read and write a language. Writing or composing documents is a process rather than a single event (Fitzgerald, 1987). Before the text contained in a book or a newspaper is published, it is typically revised one or several times. The process of *revision* is in the center of this work. Revision records much of the knowledge involved in the writing process, e.g. the detection and correction of a spelling error. Document revision is a field of study in educational science, computer science, and (psycho-)linguistics. We will deal with questions from all of these fields, although we will mostly use the Natural Language Processing (NLP) perspective when inspecting revision and user interaction in online mass collaboration.

While for many centuries, revision was the task of either the author of a document or another person in charge of copy-editing, the digital age has substantially increased the possibilities to *collaboratively* create and revise documents. The Web and especially the Collaborative Web or Web 2.0 have opened new windows into analyzing the process of development of documents and the underlying user interaction when multiple authors work on the same document. We refer to this process as *collaborative writing* (CW). In this thesis, we will open some of these windows with the intention to get a better understanding of the writing process and related phenomena in online mass collaboration. On one hand, CW has become a popular tool in education, e.g. to teach students in composition classes. On the other hand, CW is also used in industry and academia, e.g. in company wikis which facilitate storing, updating and sharing internal knowledge resources.

The CW process is a particularly interesting phenomena as its study reveals insights not only about the process of writing and revision itself, but also about how authors work together. Whether intended or not, when more than one author works on the same document, interaction takes place. The interaction does not necessarily involve direct communication, but may also happen by editing a piece of text which has been written earlier or by augmenting an existing document from another author. The success of this complex process is often considered to be dependent on the degree of coordination among the authors (Allen et al., 1987). In this thesis, we study two modes of interaction. While coordination and communication is a rather *direct* mode of interaction, revision expresses a kind of *indirect* interaction.

We chose the online encyclopedia Wikipedia as source of data for the main analyses of our work. Wikipedia is well-known both inside and outside of the Web 2.0 community. As of May 2015, wikipedia.org is ranked the 6th most popular web page globally.[1] With about 30,000 active authors on monthly average in 2014 and more than four million documents (articles), the English Wikipedia constitutes an extraordinary resource to study online mass CW. A very powerful yet often completely ignored feature of Wikipedia's underlying wiki software is the revision history, which retains every single edit to each page. The revision history of Wikipedia articles reveals much about the process of CW in the encyclopedia. Throughout this work, we will make substantial usage of this feature, especially in the context of indirect interaction. While document revision histories contain information about the authors of revisions and the textual changes they performed, the reason why a particular revision has been made might remain blurry. There are many reasons why people edit Wikipedia, and these are not always as obvious as in the case of spelling error correction. Therefore, Wikipedia also supports coordination and communication via direct interaction. For each document in Wikipedia, users can coordinate the process of revising in a dedicated discussion space. Like this, it is also possible to follow and analyze the otherwise lost motivation behind some of the revisions to documents. Figure 1.1 visualizes the two CW interaction modes discussed in this thesis by the example of Wikipedia.

Wikipedia is one of the most successful online collaboration projects so far and many studies have tried to explain this success from diverse perspectives. In the course of this thesis, we want to increase the knowledge about mass collaboration in Wikipedia with the goal to better understand some of the features behind the success of the online encyclopedia. Before we give an overview of our research questions and contributions, we will motivate the topics raised in this thesis.

[1] According to the Alexa rank: `http://www.alexa.com/siteinfo/wikipedia.org`, accessed May 25, 2015.

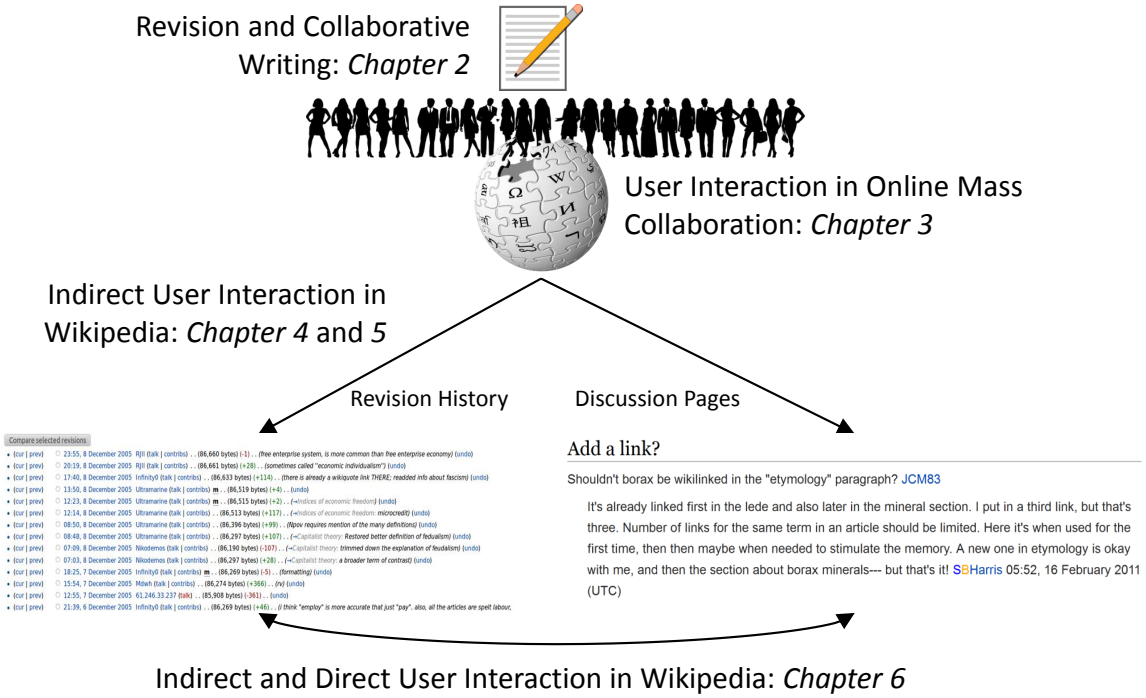


Figure 1.1: Structure of this thesis: analyzing CW in online mass collaboration based on indirect user interaction in the revision history and direct interaction on discussion pages of Wikipedia.

1.1 Motivation

The following paragraphs serve as motivation for the high-level research questions raised in this thesis. Beyond the general motivation, we also motivate the individual contributions in the respective chapters. Additionally, we show the connections between our findings and potential applications in educational science, computer science and (psycho-)linguistics in the individual chapters.

High-level Research Questions This thesis analyses the writing process in online CW. As a central aspect of online CW, we will discuss the interaction between the users involved in the writing process. The central question of this research is thus: *How does CW work in online environments and how is it different from offline writing? In other words: What happens, if hundreds of people, who do not know each other, start writing on the same document?* Lowry et al. (2004) describe CW as an iterative and social process that "includes the possibility of many different writing strategies, activities, document control approaches, team roles, and work modes." Furthermore, CW "involves a team [...] that negotiates, co-

ordinates, and communicates during the creation of a common document." We center our research questions around their observations and consequently want to know:

- Which writing strategies are successful?
- Do roles of authors actually matter?
- How can we measure the impact of coordination and communication techniques on the writing process?

These questions have guided and inspired the more precise research questions presented in section 1.2. If we get the answers to the questions raised above, we expect to

- enable a more targeted use of CW in the classroom,
- do a better job in CW projects in academia and industry by using and developing the rights tools, and
- do a better job in CW projects in academia and industry by applying suitable strategies under given circumstances.

It has been suggested that patterns of interaction in CW and the creation of knowledge are closely related (Onrubia and Engel, 2009). Thus, we can assume that a better understanding of CW will have positive effects on educational applications. Technological advances in communication and resource sharing have substantially lowered the barriers to collaboration across company and university borders. Global competition and international research often require individuals to join forces in collaborative projects, which increasingly involve CW. As a result, the need to find answers to the above questions is growing. We have chosen Wikipedia as a very successful online CW project to find answers to some of these questions.

The Current State Most research on revision and the writing process has been carried out in the context of education. Thus, little focus has been put on the particularities which arise when the writing process is carried out by many authors in an online setting. Those studies which have a focus on online mass collaboration have frequently tried to solve their research questions with manual effort. In her study of on a small sample of Wikipedia articles, Kallass (2012) found that the writing process in Wikipedia does not follow a linear process but should be described as a recursive process in which authors react to changes of their co-authors and which varies substantially across different documents. Furthermore, the set of involved authors changes over time, as do the authors' motivations to contribute to the project. Kallass (2012) also identified the importance of interaction among authors, which is carried out in several ways, and which substantially distinguishes online mass collaboration from traditional offline writing.

In the present study, we analyze the writing process in online mass collaboration on a large scale, by developing tools which are suitable for the automatic analysis of a large number of revisions made by different authors across extended periods of time. To do so, we use different revision category taxonomies designed for the task at hand. Some previous studies have tackled the problem of automatic revision classification in online mass collaboration (Liu and Ram, 2011; Bronner and Monz, 2012), however, none of them has presented a fine-granular taxonomy and model to classify individual edits. While the influence of the interaction between authors on revision has been studied in classroom settings (Yagelski, 1995), little is known about this matter in online collaboration. In our work, we bring together direct user interaction through discussion *and* indirect user interaction through collaborative revision. We use hand-tailored corpora to find answers to the posed research questions. In the next section, we present an overview of these corpora, along with a summary of our contributions and findings.

1.2 Contributions and Findings

We have already presented some of the high-level research questions of this work in section 1.1. In the following, we give a compact overview of the most important findings and contributions. Novel tasks, corpora, methods and concepts proposed in this thesis are summarized at the end of the section.

- We present a comprehensive model for the study of online mass CW, based on the notion of direct and indirect user interaction. Direct user interaction happens when authors communicate during the CW process using oral or written speech, e.g. in a dedicated discussion space. Indirect user interaction happens when two or more authors edit the same document, but do not use oral or written communication, e.g. when revising a text previously written by a co-author.

- To understand the content and intentions behind revisions in Wikipedia articles, we (a) introduce the concept of edits which allows a fine-grained analysis of revisions, and (b) present a novel taxonomy for classifying Wikipedia edits into 21 categories. On a higher level, the taxonomy divides changes into text-base (meaning-changing) and surface (meaning-preserving) edits.

- We present and discuss two novel corpora of Wikipedia edits, extracted from English and German article revision histories. Both corpora have been manually annotated with edit categories based on the newly introduced taxonomy.

- We introduce and analyze a model for the automatic classification of English and German Wikipedia edits. The model is trained and tested on the manually annotated corpora. We show that our model reaches state-of-the-art performance and how it

can be used to automatically classify article revision histories. Additionally, we show that our model performs well (a) when classifying changes on revision rather than edit level, and (b) when trained and tested on a corpus annotated with a different taxonomy of revision categories.

- For the first time, we show that the information content in high-quality articles in Wikipedia tends to become more stable after their promotion. This finding is based on an analysis of the distribution of edits labeled as text-base or surface edits in both the annotated and the automatically classified data.

- We present a novel corpus of 1000 representative articles from the English Wikipedia, based on a double-stratified sampling procedure. The corpus contains articles from 25 topical categories and four maturity stages.

- For the first time, we apply a large corpus of manually annotated English Wikipedia revisions from previous work to automatically label the revision history of a representative set of Wikipedia revisions. We use the result to detect and analyze activity-based roles in the CW process of Wikipedia. Activity-based roles are detected using an unsupervised machine learning approach over user profiles of edit behavior. Due to the novel revision category taxonomy used to classify our data, we discover several roles that have not been detected in previous work.

- We show that the nature of activity-based roles in Wikipedia remains stable across different models of the user space. The same is true for distinct time periods in the CW process. Although activity-based roles are quite stable across time, the users frequently change their roles over time, even within the same article.

- We introduce the concept of edit-turn-pairs to analyze the relationship between direct and indirect user interaction in Wikipedia. Corresponding edit-turn-pairs are instances of collaboration where a user expresses an explicit performative on a discussion page and an edit from the respective article revision history corresponds to this performative.

- We present a novel corpus of corresponding and non-corresponding edit-turn-pairs. We show how to overcome the class imbalance problem inherent to the nature of edit-turn-pairs, as the vast majority of randomly selected edit-turn-pairs are non-corresponding.

- For the first time, we analyze the impact of corresponding and non-corresponding edit-turn-pairs in Wikipedia. In a corpus of English Wikipedia articles suffering certain quality flaws, we find that articles have typically less than 5% corresponding edit-turn-pairs.

Finally, we briefly summarize the contributions listed above along the dimensions of concepts, corpora, tasks, and methods proposed as part of our work. The novel **concepts** elaborated and introduced in this work include:

- CW as *direct and indirect user interaction*
- *Edits* as fine-granular units to analyze collaborative revision in Wikipedia
- A *taxonomy of edit categories*, which distinguishes text-base and surface edits
- *(Non-)corresponding edit-turn-pairs* as interface between direct and indirect user interaction in Wikipedia

One of the central outcomes of this thesis are hand-tailored **corpora** to analyze online CW. In table 1.1, we give a short overview of the corpora we have created and/or used as part of the work, including their availability. Except for one corpus, all of them have been created as part of this work. The overview also serves as a reference for the acronyms used to refer to corpora in the main part of the thesis. To work with the corpora presented in table 1.1, we propose several new **tasks**:

- Quality assessment based on the *stability of content* in high-quality articles
- Quality assessment based on *the homogeneity of collaboration patterns* in high-quality articles
- *Stability analysis of activity-based roles* across user spaces and time
- *Detection of corresponding and non-corresponding edit-turn-pairs* in Wikipedia

The novel **methods** we introduce to solve the posed research questions can be summarized as follows:

- *Automatic classification of Wikipedia edits* across 21 categories
- *Automatic classification of Wikipedia revisions* across 12 categories
- *Automatic classification of edit-turn-pairs* into corresponding and non-corresponding pairs

The individual research questions raised in each of the chapters 4 through 6 are stated at the beginning of each chapter.

Corpus	Full Name	Main Purpose	Lang.	Self-Created	Avail./License	Described in Chapter
WPQAC	Wikipedia Quality Assessment Corpus	Balanced sample of Wikipedia articles for quality assessment	en	yes	yes, no license	4.3.2
WPEC	Wikipedia Edit Category Corpus	Classification of revisions on edit level	en	yes	yes[a], CC-by-SA	4.3.2
WPEC-GER	German Wikipedia Edit Category Corpus	Classification of revisions on edit level	de	yes	yes, upon request	4.4.1
WPREP	Representative Wikipedia Revisions Corpus	Sample of Wikipedia revisions for analysis of revision category distribution	en	yes	no	5.2
ETP-GOLD	Wikipedia Edit-Turn-Pair Corpus	Detecting pairs of edits from articles and turns from discussion pages	en	yes[b]	yes[c], CC-by-SA	6.2.2
WPRC	Wikipedia Revision Category Corpus	Classification of revisions	en	no	no	4.5.1

[a] `http://www.ukp.tu-darmstadt.de/data/wiki-edits`.
[b] The Human Intelligence Task (HIT) Layout and Setup were carried out by Emily Jamison.
[c] `http://www.ukp.tu-darmstadt.de/data/edit-turn-pairs`.

Table 1.1: Overview of corpora used and/or created as part of this thesis.

1.3 Publication Record

In the following, we list our previously published work. All of these publications have been peer-reviewed by researchers from NLP and related fields and most of them were presented at major conferences. Parts of the publications have been reused in this thesis. We indicate if and where a publication has been reused, and whether verbatim quotes from this publication are to be expected. Some of the contributions covered in this thesis are still under review at the time of publication and can thus not be included in this list.

In Daxenberger and Gurevych (2012), we introduce and analyze the concept of edits in Wikipedia, based on article revision histories. We present a novel taxonomy for edit categories, which we use to manually annotate a corpus of English Wikipedia edits. Based on the annotated edits we show that, once they are promoted, high-quality articles tend to become more stable in terms of their information content. Parts of this publication (including verbatim quotes) have been reused in chapter 4, in particular in the sections 4.1, 4.2, and 4.6.

In Daxenberger and Gurevych (2013), we apply the corpus of edits created in Daxenberger and Gurevych (2012) for the automatic classification of Wikipedia edits. We present

and analyze a machine learning system for multi-labeling previously unseen edits. Using this system, we classify a larger portion of edits from high-quality articles in the English Wikipedia, showing that the collaboration patterns within such articles are distinct from those in lower-quality articles. Parts of this publication (including verbatim quotes) have been reused in chapter 4, in particular in the sections 4.3 and 4.6.

In Daxenberger and Gurevych (2014), we introduce the concept of edit-turn-pairs, which connects Wikipedia article revision histories and discussion pages. We explain how the potential correspondence between edits and turns can be used to analyze the CW process in Wikipedia and create a small manually annotated corpus of corresponding and non-corresponding edit-turn-pairs. Furthermore, we present a machine learning system which is able to automatically detect previously unseen corresponding edit-turn-pairs in Wikipedia articles. Parts of this publication (including verbatim quotes) have been reused in chapter 6.

In Daxenberger et al. (2014), we identify and discuss the requirements of a text classification system. We show how these requirements are met in the modular architecture of the text classification framework DKPro TC. We demonstrate the prototypical usage of DKPro TC with the help of a tweet classification use case. Minor parts of this publication (including verbatim quotes) have been reused in appendix A and B.

In Ferschke et al. (2013), we present and discuss state-of-the-art approaches which make use of Wikipedia as a dynamic resource. In particular, we survey NLP applications based on data extracted from Wikipedia's article revision history and discussion pages. Minor parts of this publication (including verbatim quotes) have been reused in chapter 3, in particular in section 3.2.3.

1.4 Thesis Outline and Term Conventions

In the following, we give a high-level overview of the organization of this thesis. While chapters 2 and 3 introduce the necessary background and theory about revision and online mass collaboration, chapters 4 through 6 contain our main contributions and answers to the research questions discussed in this work. In each of these chapters, we additionally discuss implications and applications of the developed methodology which go beyond our use case, Wikipedia.

In chapter 2, we discuss theoretical foundations in writing research, revision, and CW. This includes the clarification of concepts and terminology, and a review of the most important previous work. We explain the relevant fields of research on the writing process and revision classification. Furthermore, we review the history and recent developments with respect to CW and present an organization of strategies, tasks, and roles in CW.

In chapter 3, we first discuss user interaction in online CW. We draw the basic distinction between direct and indirect interaction and discuss the practice of these modes of in-

teraction in online mass collaboration. We then turn to collaborative editing in Wikipedia, presenting the basic CW functionality of wikis and in particular, Wikipedia. We discuss the implications of the encyclopedic setting of Wikipedia and its authors for the CW process and show how direct and indirect user interaction are realized in Wikipedia. Finally, we introduce parameters for article quality in Wikipedia, which we will refer to in the course of the thesis.

In chapter 4, we present a detailed analysis of indirect user interaction in Wikipedia. Based on the concept of edits which are extracted from Wikipedia revisions, we present a novel taxonomy for the classification of changes in Wikipedia. We manually label two corpora with this taxonomy and apply the resulting data to train and test a machine learning classifier. Furthermore, we train and test our machine learning model on a different taxonomy and corpus, which is based on revisions and not edits. With the help of the models analyzed before, we classify a larger number of revisions from high-quality articles in the English Wikipedia and analyze potential relationships between indirect user interaction and article quality.

In chapter 5, we explore activity-based roles in Wikipedia, based on indirect user interaction. To this end, we apply one of the models explained in chapter 4 to classify a large number of revisions from the English Wikipedia, and cluster authors based on the kind of changes they have performed. We present the resulting seven clusters and dicuss their meaning for the nature of activity-based roles in online mass collaboration. Furthermore, we assess the stability of these roles over different configurations of the input user space and over time.

In chapter 6, we connect indirect interaction to direct interaction with the help of edit-turn-pairs. Edit-turn-pairs indicate a potential correspondence between Wikipedia article edits and discussion page turns. We define corresponding and non-corresponding edit-turn-pairs, and create a small corpus of edit-turn-pairs annotated according to our definition. We use this corpus to train and test a machine learning model to automatically detect corresponding edit-turn-pairs. With the help of this model, we classify a large amount of edit-turn-pairs from the English Wikipedia and analyze the distribution of (non-)corresponding edit-turn-pairs across various articles.

In chapter 7, we summarize our main contributions and give implications and recommendations based on our findings. Furthermore, we discuss open issues and potential future work.

The appendix contains several technical descriptions, which explain some of the concepts discussed in this thesis in more depth. It contains a very short introduction to the foundations of supervised learning on textual data, as many of the findings explained in this work have been produced with such approaches. In particular, we focus on the text classification framework DKPro TC, which was substantially extended as part of the work

described in this thesis. Finally, we also append a condensed version of the guidelines for the annotation studies which have been carried out in the course of this work.

The typographical and terminological conventions in this thesis are as follows. Indexed or otherwise important terms are printed in *italics* when they are first introduced. Names of corpora produced or consumed in this work are printed in CAPITALS. Wikipedia edit and revision categories are printed in SMALL CAPITALS. We use the term *users* to refer to (usually active) editors of the online encyclopedia Wikipedia, whereas the term *author* is used in a broader context and generally refers to any participant in the (collaborative) writing process. The product under revision by the authors is called *document* in general, and *page* or *article* in the context of Wikipedia. Throughout the thesis, we will abbreviate the expression *collaborative writing* as CW.

Chapter 2

Writing and Revision

This chapter lays the theoretical foundations for the concepts and methods discussed in the subsequent chapters. The main focus of our work is an in-depth exploration of the writing process in online mass collaboration environments. Before we focus on online mass collaboration in chapter 3, we give a broader introduction to the process of writing and collaborative writing (CW). We will start this chapter with an introduction to writing research (section 2.1). Then, we will analyze the writing process, and in particular, revision, in more detail (section 2.2). Finally, CW will be the focus of the third part of this chapter (section 2.3).

2.1 Writing Research Foundations

We open this chapter with a very brief history of writing, before we turn to the disciplines involved in writing research, their motivation to do research, and domain-specific applications. The first known pictographic writings are said to be older than 4000 years (Daniels and Bright, 1996). Proto-Cuneiform, the oldest known writing system was used in ancient Mesopotamia, on modern-day Iraqi ground. The history of writing is closely connected to the instruments and materials that were available at certain times and in distinct regions, e.g. the media which were used to produce and carry the writing. Throughout centuries, writing by hand was the only way to produce texts. The printing press and subsequent printing techniques enabled mass printing, rapid reproduction of texts and thereby, mass communication. In the 19th century, typewriters started to produce the shift from handwriting to typing. Since the end of the 20th century, digital writing has revolutionized the writing process, and the act of writing and the act of printing have been largely separated (Daiute, 1986; Vacc, 1986). In addition, new forms of publishing have been developed, including instant publication in online media such as (micro-)blogs and social networks.

Writing is quite different from speaking, and it is more than just a physical representation of what we think and say. The visual representation of one's own or other's thoughts can have surprising effects.[2] One of the simplest but also most effective representations of cognitive concepts is to write them down. Writing can be "a way to explore one's feelings and thoughts" (Zamel, 1982, p. 205). When we talk about writing, we refer to its composing sense, rather than transcribing or copying. Composing a written document is a creative act which involves a broad range of cognitive procedures (Flower and Hayes, 1981). It is a *process* which usually implies revision. *Revision* is the visible product of consuming, rethinking and rewriting of what has been written before, and therefore an expression of cognitive development. As Sommers (1980, p. 387) puts it, experienced writers have "a sense of writing as discovery".

2.1.1 Concepts and Terminology

In the following, we present a couple of important concepts which will be relevant in the remainder of this work. We will first explain these concepts by their journalistic definition. Although some terms are used quite differently across disciplines and communities, all of them usually root in the domains of journalism, composition, and publishing.

Composition, the act of writing, is the top-level concept in writing research. We refer to the producer of a written composition as its *author*, and to the product as *document* (e.g. a book). In this study, we are particularly interested in the concepts of revising and rewriting a document. Whereas the term *revising* emphases reconsidering and rethinking existing content (from Old Latin *revisere*: revisit, go back, look at again),[3] *rewriting* refers to the actual act of composing a document or parts of it again, usually involving substantial changes as compared to its original state.[4]

The term *editing* is used in a broad range of contexts, and often closely related to revising. However, in the editorial process, the two concepts are quite different. The Oxford Dictionaries define editing as "prepare (written material) for publication by correcting, condensing, or otherwise modifying it".[5] Editing or *copy-editing*[6] (i.e. editing the *copy* of a document for publication), as part of the last stages in the publishing process, may involve a considerable amount of changes to a document to ensure accuracy and consistency. However, it is not supposed to change the content of the document itself. In professional settings, (copy-)editing is often performed by a another person than the author of the original writing. In addition, *proofreading*, i.e. correcting spelling, grammatical and typographical er-

[2] For example, one might discover shortcomings in reasoning and argumentation, as expressed in the saying "Now that I see it on paper...".

[3] `http://www.oxforddictionaries.com/definition/english/revise`, accessed May 25, 2015

[4] `http://www.oxforddictionaries.com/definition/english/rewrite`, accessed May 25, 2015

[5] `http://www.oxforddictionaries.com/definition/english/edit`, accessed May 25, 2015

[6] Editing and copy-editing are often used interchangeably, while they are sometimes considered separate steps in the editorial process.

rors, may be considered separately from copy-editing (Einsohn, 2011). In the publishing process, proofreading is the last step before publication and therefore strictly limited to very specific edits correcting serious errors which have not been corrected before.

The verb *editing* originally stems from *Editor*, i.e. "a person who is in charge of and determines the final content of a newspaper, magazine, or multi-author book",[7] which itself stems from Medieval Latin *edere*: to give out, put forth, publish. An *edition* is a volume of identical or nearly identical copies of a book. Reprints of popular books resulting in a new print run do not necessarily contain any substantial changes (beyond minor typographical changes or spelling corrections) and therefore do not constitute a new edition. However, major changes such as adding new content, or rewriting parts of the content, result in a new edition of the book. The numbering of book editions is a special kind of *revision control* and a wide field with a long tradition. Although there are bibliographical definitions and guidelines, in many cases the versioning and especially the naming of an edition (e.g. "2nd, *revised* edition") are rather subjective decisions, often bound to marketing strategies.

2.1.2 The Writing Process

While research on writing systems has a long tradition, research on the *writing process* is a rather young discipline which did not get much attention until the late 1970s (Emig, 1971; Fitzgerald, 1987). This is connected to a shift in educational research and teaching. Throughout decades, teachers in essay writing and composition focused on the final product (product-focused view). According to the traditional model of the writing process, it consists of three stages, namely prewriting, writing and rewriting (Rohman, 1965; Murray, 1978a). Faigley and Witte (1981) call this the tidying-up view, since it considered revision mostly as copy-editing. This view was supported by the fact that in teaching usually only the relationship between certain pedagogical approaches and the final writing (i.e. the product) was analyzed. When in the late 70s attention shifted towards a process-oriented view (Fitzgerald, 1987), teachers eventually started to teach and train revision, i.e. to intervene their students while writing (Zamel, 1982). Today, both in research and teaching, writing is studied as a process which includes revision, and therefore revision has become its own subject of study and training.

In an attempt to better understand the cognitive processes of writers during the writing process, Flower and Hayes (1981) suggest four basic aspects of the writing process, based on a protocol analysis (basically a think-aloud protocol of writers while composing a text on a given subject). We list them here in a slightly shortened version:

1. the writing process is made up of a set of distinctive cognitive processes, which are applied in a non-linear fashion

[7] `http://www.oxforddictionaries.com/definition/english/editor`, accessed May 25, 2015

2. these cognitive processes are embedded and organized hierarchically
3. the overall writing process is goal-directed
4. authors create their own high-level goals and sub-goals and sometimes change high-level goals based on what they have learned during writing

These aspects have been quite fundamental to subsequent studies. Flower and Hayes (1981) summarize their findings in a so called writing model. In this model, the writing process is connected to two other elements, namely the task environment (e.g. the writing assignment) and the writer's long-term memory (e.g. the writer's knowledge of the writing assignment and the audience).

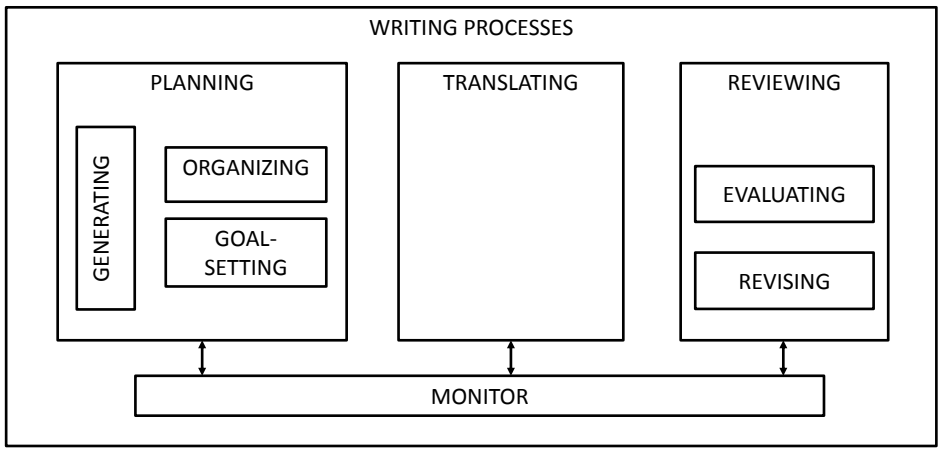


Figure 2.1: The writing process, part of the writing model, as proposed by Flower and Hayes (1981).

The most important conclusions with respect to the writing process from Flower and Hayes (1981)'s writing model are visualized in figure 2.1. The cognitive processes involved in the writing process are hierarchically embedded (as visualized through the boxes) and they are not executed in a linear order. The force behind these processes are high-level goals and sub-goals which might be adapted during the writing. The writing process itself constitutes three processes: planning, translating and reviewing. *Planning* refers to the author's act of building an internal representation of ideas and knowledge. The involved sub-processes are *generating ideas*, i.e. drawing relevant knowledge from the long-term memory, *organizing*, i.e. structuring the ideas, and *goal-setting*, i.e. a set of ambitions developed by the author which drive idea generation and organization. *Translating* is the process of putting the ideas generated during planning into visible language. This process depends on the grammatical and formal proficiency of the author. *Reviewing* includes the subprocesses *evaluating* and *revising*, which may occur at any time and usually trigger new planning and translating processes. The *monitor* enables the author to switch between processes, determined by their goal and personal habits.

As we have shown, the writing process is a complex area of research, with *revision* being one of the most important concepts in writing. Without losing the bigger picture of the overall process, in the following we will address revision on a more detailed level.

2.2 Revision

Given the changes in the perspective on writing explained in the previous section, research on revision itself – a concept inseparably connected to the writing process and sometimes used interchangeably – started to gain momentum. In the 1980s, several studies tried to answer questions about the nature of revision: about how much and when to revise, and what kinds of revision are performed (Sommers, 1980; Faigley and Witte, 1981; Fitzgerald, 1987). Fitzgerald (1987, p. 484) proposes the following definition of revision: "Revision means making any changes at any point in the writing process. It involves identifying discrepancies between intended and instantiated text, deciding what could or should be changed in the text and how to make the desired changes. Changes may or may not affect the meaning of the text, and they may be major or minor. Also, changes may be made in the writer's mind before being instantiated in written text, at the time the text is first written, and/or after the text is first written [...]". This definition includes both the revising and the rewriting aspect of composition introduced in section 2.1.1. Fitzgerald (1987) also analyzed revision as part of the writing process. She points out that the term *revision* was used equally to editing or error correcting for many centuries. When researchers started to investigate the concept of revision with a process-oriented view on writing, three new aspects of revision were introduced (Fitzgerald, 1987):

1. revision may occur any time in the writing process, i.e. "before, while and after putting the pen to paper" or typing (Fitzgerald, 1987, p. 483)

2. revision can be meaning-based (affecting the text-base) or surface-based (not meaning-changing) (Faigley and Witte, 1981)

3. (visible) revision is directly connected to what happens in the mind of the revising author during the revision (revision as a mental process, learning through revision)

The focus of this thesis is collaborative revision, but we will address all of these aspects at varying levels of detail. In chapter 4, we shed light on different categories of revision in Wikipedia (both meaning-changing and meaning-preserving). In chapter 6, we turn to events which are directly related to revision but do not change the document itself. We also take a look at the authors in chapter 5, where we analyze the roles of Wikipedia users within the writing process.

As Faigley and Witte (1981) state, expert writers tend to revise in very diverse ways, and the number and extent of changes does not necessarily correlate with the quality of the

text after revision. We will discuss this matter in more detail in section 2.2.2. Faigley and Witte (1981) also highlight the importance of the "situational variables for composing" (e.g. why the document is written, on which medium it is written, the author's familiarity with the subject). They draw a clear connection between success in revision and the author's planning and reviewing skills (Flower and Hayes, 1981) and point out the importance of getting students in writing courses to "see [what they have written] again" and then revise their writing where necessary.

2.2.1 Taxonomies of Revision Categories

The most basic categorization of revisions is the distinction between changes that influence the meaning of a text (on all levels) and changes that do not affect its meaning. According to (Faigley and Witte, 1981, p. 402), the meaning of a text is affected if "new information is brought to the text" or "old information is removed in such a way that it cannot be recovered through drawing inferences". Murray (1978a) makes a similar distinction and calls the two forms of revision *internal* and *external*.

Apart from the basic distinction between revision which changes the meaning and revision which does not change the meaning, several more fine-grained taxonomies of revision have been developed. Hildick (1965) analyzed the revisions of several writers of fiction. He suggested the following revision categories: tidying-up changes, roughening-up changes, power changes, structural alterations, ideologically determined changes, and miscellaneous. Stallard (1974), as cited in Fitzgerald (1987), distinguishes spelling, syntax, multiple-word, paragraph, punctuation, and single-word changes in the revision of essays of 12th graders. In later studies (Bridwell, 1980; Sommers, 1980), linguistic and syntactic levels were separated and the applicability of revision categories to textual changes was improved (Fitzgerald, 1987).

Faigley and Witte (1981) present the first elaborated taxonomy capturing the intentions behind a textual change, as displayed in figure 2.2. Changes which affect meaning are called *text-base changes* and 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*. The latter includes six categories, namely additions, deletions, substitutions, permutations, distributions and consolidations. All of them add, delete, substitute, permute, split, and respectively merge words or longer text sequences without changing the meaning of the text. 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. Like meaning-preserving changes, microstructure changes and macrostructure changes are further divided into additions, deletions, substitutions, permutations, distributions and consolidations. However, as opposed to meaning-preserving changes, microstructure changes and macrostructure changes do change the meaning. Faigley and Witte (1981) give examples for their cat-

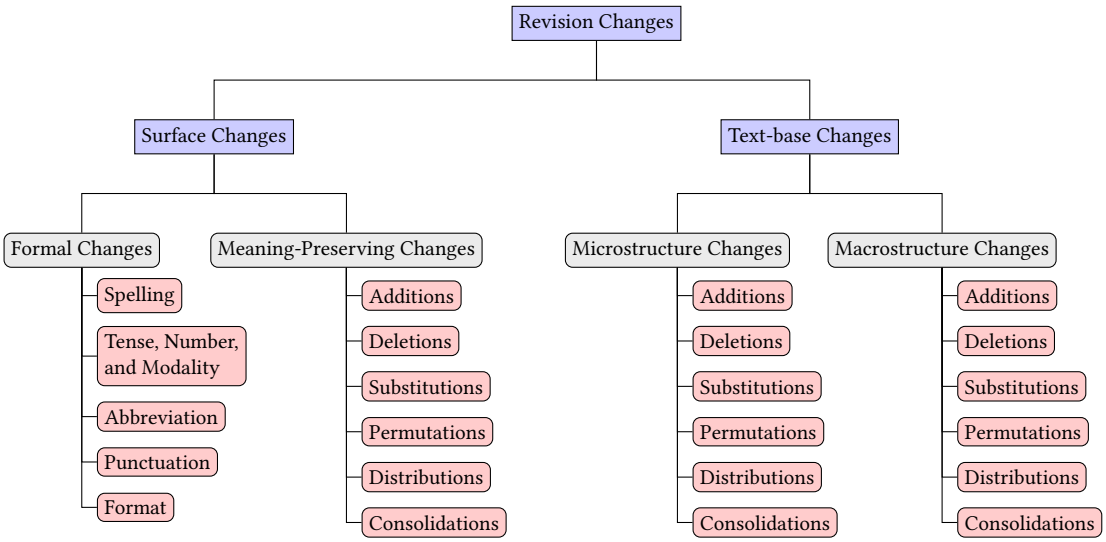


Figure 2.2: Faigley and Witte's (1981) taxonomy of revision changes.

egories, but do not specify in much detail where to draw the border between meaning-changing and meaning-preserving revisions. Their taxonomy includes changes from all stages of the editorial process (copy-edits as well substantial revisions).

2.2.2 The Relationship between Revision and Text Quality

Analyzing the relationship between revision and text quality is one of the ways to understand the cognitive processes and knowledge creation in the mind of the author during the process of revision. If revision (beyond editing) is found to substantially improve the quality of a document, this apparently shows a cognitive development and potentially new knowledge in the mind of the author. Certainly, a definition and analysis of the quality of a document is a subject of research of its own. We therefore limit our discussion about quality to the domain of education, where quality is often measured as the expertise of writers, e.g. undergraduate students compared to graduate students (Perl, 1979). We will return to discuss document quality within the limited context of Wikipedia in section 3.2.7.

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. Faigley and Witte (1981) confirm Sommers's (1980) findings to the largest extent. They tested and verified their taxonomy (see section 2.2.1) using the revisions made by 18 writers with different writing skills. With respect to the distribution of surface and text-base changes, they found substantial differences between experienced and unexperienced writers. Although

the inexperienced writers performed more changes overall, they made very few text-base changes. The experienced writers made fewer changes in total, but more changes affecting the meaning as compared to the inexperienced writers.

Apart from the connection between the kind of revision (surface vs. text-base change) and the expertise of writers, several studies have analyzed the relationship between collaborative revision and quality. Most studies agree that documents created using collaborative revision and peer reviewing are of higher quality as compared to those which have been written and revised by a single author (Storch, 2005). In particular, a thoroughly coordinated process of collaboration during revision seems to improve the quality of documents in classroom settings (Erkens et al., 2005; Wichmann and Rummel, 2013).

2.2.3 Analyzing Revision: Motivation

Several fields are involved in research on revision, each one with a different scope. We list the most important ones here, explaining their motives and applications. In the course of this study, there will be repeated references to research questions, insights and applications from these fields.

Educational Science Certainly the main driving force behind the research on revision stems from the questions raised by teachers and educational scientists. Since the shift in the teaching paradigm in the late 70s, when essay composition classes started to focus more on the process of writing than on the final product, a lot has changed (Murray, 1978b; Zamel, 1982, 1983; Harris, 2003; Horning and Becker, 2006). Today, revision is a tool carefully thought and practiced in the classroom. The literature about revision in essay composition and foreign language learning is vast (Xue and Hwa, 2010; Mizumoto et al., 2011). The main motivation behind research on revision in teaching is the analysis of the relationship behind revision and either text quality (Sommers, 1980) or the learning effect (Murray, 1978b). Other studies have analyzed peer-reviewing as a tool to teach students to review their own writing with the eyes of their potential readers (Hyland and Hyland, 2006). We will deal with another shift in education in section 2.3, namely the switch from single-author writing to multiple-author writing.

Computer Science Revision is a well-known concept in computer science. Similar to book editions, computer programs can be modified and released in a newer version. Versioning the various revisions of a piece of code or a program is essential to ensure its compatibility with other programs, operation systems etc. The main motivation for developing complex version control systems is the huge market of computer programs used and developed by a wide audience, e.g. in the context of open-source software. Revision control is a basic need in multi-developer environments, where multiple people simultaneously work on the

same software code. In software development teams with a high number of people, seamless collaboration and fruitful collaboration patterns become more and more essential (Fan et al., 2012), so that the requirements for revision control systems continuously grow and often go beyond the mere numbering of changes. Whereas revision control refers to the fine-granular, incremental numbering of each single change entered into a version control system, the version number of a released software product is typically determined by its authors.

Version control systems originally developed to track software changes have also influenced the versioning of textual revisions. From a technical viewpoint, developing and maintaining a software program is quite similar to writing a text. The functionality of modern tools for CW (see section 2.3) is based on the features of software version control systems. This includes the possibility to revert (undo) malicious or unwanted changes, compare revisions side-by-side, monitor metadata of revisions (author, time stamp) for the entire history, and change/commit notifications (Sharples et al., 1993, p. 27).

For many years, computer programs include tools to support writers on digital media (Mahlow and Piotrowski, 2008), e.g. spell checkers. With the rise of artificial intelligence (AI), intelligent writing assistants have grown beyond simple spell and grammar checkers and offer more or less sophisticated real-time feedback about the writing of their users (Heidorn, 2000). The use of AI for revision assistance is still limited by the performance of the underlying models. Although such models usually learn from real-world data, they are typically black boxes and may thus not be accurate and comprehensible enough to be used (e.g. in the classroom). However, the big data paradigm and technological advances in the construction of hardware have increased the viability of AI-based tools and it is to be expected that the use of such systems will increase in the near future, see e.g. Simon (2013).[8]

Psycholinguistics In very close connection to the educational scholars, researchers from psychology and linguistics started to discover writing as a psychological process (Emig, 1971; Daiute, 1982; Scardamalia and Bereiter, 1985). Their research focuses on the creation and development of new ideas and knowledge during the process. Like most of the educational scholars, psycholinguists concentrate on the processes that happen in the mind of the writer during revising. Many of the questions stemming from this line of research are inseparably connected to the CW paradigm. Along these lines, the *knowledge building theory* (Scardamalia and Bereiter, 1994) which eventually became very popular within the

[8] "Big data" has become a buzzword for many concepts involving large data collections, and it is disputed what "big" actually means. Big data programs have become popular in industry, academia, and government. For example, the Obama administration in the US made use of big data during the 2012 election campaign by addressing individual voters on the Web, social networks, and smartphones (see e.g. `http://www.technologyreview.com/featuredstory/508836/how-obama-used-big-data-to-rally-voters-part-1/`, accessed May 25, 2015).

CSCL (computer supported collaborative learning) community, developed. We will discuss the relevant research questions from this area in more detail in section 2.3, but they essentially center around the following question: How is knowledge building and particularly, collaborative knowledge building, connected to writing strategies and processes? Based on the answers to this question, computer-supported collaborative writing researchers want to determine how computers are able to support writers.

2.3 Collaborative Writing

In this section, we lay the theoretical and terminological foundations of collaborative writingCW with a particular focus on computer-supported collaborative writing. In an attempt to contribute to the consistency and clarity of CW terminology, we follow the terminology of Lowry et al. (2004) with respect to CW and related concepts. Building on the finding that experienced writers were able to read their own writing with the eyes of potential readers, US composition teachers in the late 1980s increasingly relied on peer reviewing among students (Herrington and Cadman, 1991; Tuzi, 2004). Furthermore, they discovered that the process-like nature of revision can be conveniently taught to students with the help of a CW task. CW naturally incorporates two important concepts of revision:

1. Shifts of perspective as the writers need to coordinate their writing and
2. non-linear, iterative text production as writers need to revise the writing of their fellow writers.

Given these connections, it is not surprising that CW has become an integral part of writing classes. American college composition classes have been using CW for many years to improve both the students' social skills as well as their writing skills (Sharples et al., 1993). However, CW has not only proven to be useful in teaching, but was also increasingly used in industry and academia (Faigley and Miller, 1982; Allen et al., 1987; Ede and Lunsford, 1990; Rimmershaw, 1992). To a large extent, the latter was a consequence of the rapid development of computer technology (Jones, 2005). Lately, the rise of the Web 2.0 has brought further momentum to the paradigm of CW. CW was and is also used in the creation of fiction, but to a much lesser degree and with rather little success so far (Ede and Lunsford, 1990). Despite its success in the classroom and beyond, CW is a complex process which has drawn the attention of a vast amount of research in past years and it is not fully understood to date (Lowry et al., 2004).

The complexity and the broad spectrum of CW research is reflected by the various names that have been used over the years to denominate CW, including co-authoring, collaborative authoring, joint authoring, collaborative editing, cooperative writing, group writing, and shared-document collaboration (Lowry et al., 2004). Where not explicitly stated otherwise, we summarize all of these terms under the common concept CW. Some studies relate

collaboration with the focus on a common final product, whereas *cooperation* might involve more than one final product (Johnson et al., 1994; Onrubia and Engel, 2009). As suggested by Lowry et al. (2004), we do not distinguish between the terms *cooperative* writing and *collaborative* writing. In CW various writers cooperate to create a single document (Galegher and Kraut, 1994), whereas in *single-author writing* (Flower and Hayes, 1981) only one writer plans, drafts and and revises his or her own writing. As highlighted by Rice and Huguley J.T. (1994, p. 163f.), "... collaboration is any writing performed collectively by more than one person that is used to produce a single text ...".

Galegher and Kraut (1994) explain CW as a non-linear, dynamic process. Although this process might involve sequential elements, its components and the roles of authors are hard to predict and therefore contribute to the complex nature of CW (Noël and Robert, 2004; Lowry et al., 2004). Lowry et al. (2004) propose six axioms to capture the nature of CW:

1. single-author writing involves planning, drafting, and revising
2. CW extends single-author writing by involving multiple parties
3. CW therefore involves social activities, such as building consensus
4. CW requires effective group dynamics including coordination and communication
5. CW should involve pre-task and post-task activities (e.g. group formation and task delivery)
6. CW requires group tasks carried out in team work

These axioms clearly reflect the crucial importance of (successful) group activity and coordination during the process of CW, a view which is supported by an exhaustive number of studies (Allen et al., 1987; Posner and Baecker, 1992; Galegher and Kraut, 1994; Erkens et al., 2005; Wichmann and Rummel, 2013). Lowry et al. (2004, p. 73f.) summarize as follows: "CW is an iterative and social process that involves a team focused on a common objective that negotiates, coordinates, and communicates during the creation of a common document. The potential scope of CW goes beyond the more basic act of joint composition to include the likelihood of pre- and post-task activities, team formation, and planning. Furthermore, based on the desired writing task, CW includes the possibility of many different writing strategies, activities, document control approaches, team roles, and work modes." Although we will have to refine our perspective on CW in chapter 3, this definition covers all important aspects of CW discussed in this work.

A growing stream of research is carried out to analyze the construction of knowledge during the process of CW (Scardamalia and Bereiter, 1994). This kind of collaborative knowledge construction is based on the knowledge building theory developed by Marlene Scardamalia and Carl Bereiter. Scardamalia and Bereiter (2003) define *knowledge building*

"as the production and continual improvement of ideas of value to a community, through means that increase the likelihood that what the community accomplishes will be greater than the sum of individual contributions and part of broader cultural efforts." As opposed to the learning process of an individual, knowledge building is seen as the "creation or modification of public knowledge". Scardamalia and Bereiter (2003) further state that knowledge building "goes on throughout a knowledge society and is not limited to education". In a CW project, knowledge is created when authors add or modify information based on their individual skills and knowledge, and as they contribute their own knowledge, they learn from co-authors based on their previous or subsequent revisions. This kind of knowledge building has been studied within several CW tools, e.g. in the context of wikis which are discussed in section 3.2.1.1 (Cress and Kimmerle, 2008; Moskaliuk et al., 2009).

Any research about CW needs to take into account the social dimension of this phenomenon. Related questions usually gather around the following aspects, cf. Posner and Baecker (1992):

- the *strategies*, activities, and work modes of CW,
- the *roles* of the authors in the CW process,
- the *technical issues* addressing the needs of writers to communicate and coordinate their writing,
- and the relationship between the *quality* of the outcome and the CW process (Storch, 2005; Onrubia and Engel, 2009; Hanjani and Li, 2014).

In the following, we will outline the questions and findings which are relevant to our work in more detail, before we turn to computer-supported collaborative writing in section 2.3.3.

2.3.1 Organization of Collaborative Writing

Several studies have analyzed strategies behind a CW activity, i.e. the high-level approaches for coordinating the writing (Ede and Lunsford, 1990; Posner and Baecker, 1992; Lowry et al., 2004). Obviously, the strategy employed depends on a lot of factors such as the desired outcome, the involved writers, the group size, the setting (e.g. classroom, academic or industry) etc. Despite these factors, a more or less coherent set of strategies has been suggested in previous studies, e.g. Bisaillon (2007). Based on the suggestion of CW strategies in Ede and Lunsford (1990), Lowry et al. (2004) define the following four *collaborative writing strategies*:

- *group single-author writing*: a team decides over the content of the document that should be writen, but only one author writes the final document; only used for simple CW tasks, where consensus on the written document is not very important

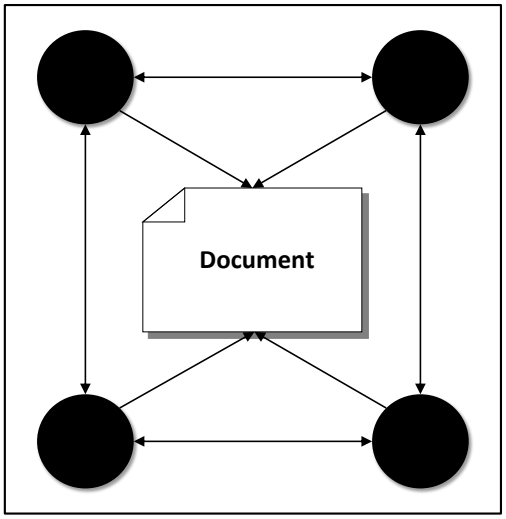


Figure 2.3: The reactive CW strategy, where authors work in real-time on the same document, reacting to changes from co-authors; as suggested by Lowry et al. (2004).

- *sequential writing*: one author writes after another, either a part of the document or a complete draft which is revised by the next author
- *parallel writing*: authors work in parallel, each on a different part of the document (e.g. a section or chapter of the entire document; *horizontal-division writing*) or on a different subtask expressed by the role of the author (e.g. reviewer or editor; *stratified-division writing*)
- *reactive writing*: authors write and react to others' changes on the same document in real-time, as opposed to parallel writing this can happen in the same parts of the document and is usually not preplanned and not explicitly coordinated; visualized in figure 2.3

Noël and Robert (2004, p. 76f.) found that, in practice, parallel writing is the most popular CW strategy. Onrubia and Engel (2009) confirm this finding in their study of the in-depth structure of CW strategies in a classroom setting. They find that the several groups they had inspected applied a variety of strategies, with parallel approaches predominating. In a study about undergraduate student CW, Marttunen and Laurinen (2012) partly contradict this finding, since single-author writing approaches were preferred by their students. As noted by Ferschke (2014) and explained in more detail in section 3.2.1.1, reactive writing is the only suitable CW strategy for online mass collaboration, where authors are often not able to preplan and extensively coordinate their writing due to the distributed setting and the huge number of authors.

Previous research suggests that there are recurring activities typically involved in a CW task (Posner and Baecker, 1992; Galegher and Kraut, 1994; Lowry et al., 2004; Onrubia and Engel, 2009). Galegher and Kraut (1994) analyzed student CW projects and found three high-level phases in the following (typical) order: plan, write, revise. However, the time spent on the individual phases varied significantly for different communication modalities (e.g. face-to-face, computer). Going a bit more into detail, Lowry et al. (2004) suggest a list of seven *collaborative writing activities*:

- *brainstorming*: collecting new ideas for the document
- *converging on brainstorming*: among all authors, processing the results of the brainstorming
- *outlining*: creating a high-level structure and direction of the document
- *drafting*: writing the initial text of the document, usually incomplete
- *reviewing*: reading and annotating the result from drafting, suggesting improvements to content, grammar and style
- *revising*: responding to the comments from reviewing, by applying changes to the document
- *copy-editing*: applying final changes to the document, typically related to consistency and carried out by a single author, cf. section 2.1.1

Although there is a natural order to these activities, they do not necessarily occur in a fixed order but are flexible and often iteratively applied, depending on the task and the needs of the involved authors. As opposed to Flower and Hayes (1981), Lowry et al. (2004) consider reviewing and revising separately.

Another crucial factor in CW settings is the *work mode*. Lowry et al. (2004) define the CW work mode along two dimensions: the physical closeness of the authors and the synchronization of the writing activity. The resulting four modes are summarized in figure 2.4. CW work modes are directly related to the group awareness, i.e. the understanding of an author's own work in the context of the activities of other group members (Lowry et al., 2004).

2.3.2 Roles in Collaborative Writing

Previous research suggests that authors in the CW process assume roles (Posner and Baecker, 1992; Lowry et al., 2004; Marttunen and Laurinen, 2012). These roles can be determined based on a formal dimension (e.g. administrator) or on an activity-based dimension (e.g. reviewer), see also Leland et al. (1988) and Sharples et al. (1993). Roles can be assigned by a

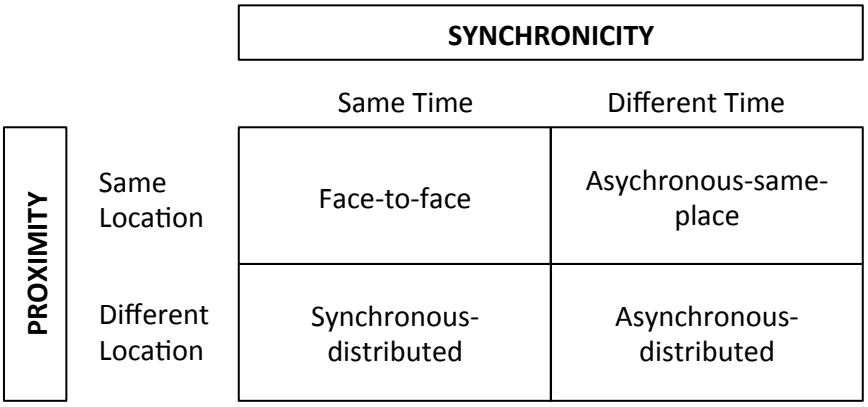


Figure 2.4: CW working modes, taken from Lowry et al. (2004) and first suggested by Johansen (1988).

leader, self-assigned (consciously or unconsciously) or introduced by constraints of the author or the task setting. Lowry et al. (2004) suggest the following *roles*, mixing both formal and activity-based dimensions:[9]

- *writer*: writes a portion of the document
- *consultant*: provides general feedback but has no ownership of the document (usually external person)
- *editor*: responsible for the overall content produced by the writers, may change the document
- *reviewer*: provides content-specific feedback without responsibility to change the document
- *team leader*: leads the team, is usually fully involved in the CW process
- *facilitator*: leads the team through appropriate processes but does not give content-specific feedback (external person)

Lowry et al. (2004) suggest that roles are flexible (writers may change their role during the CW project) rather than static (writers keep a single role throughout the entire project). There is less agreement about the importance of roles in CW, although Stratton (1989) argues that such roles make the CW process more efficient as authors can be deployed according to their skills. The meaning and division of roles is certainly influenced by the size of a CW team. In a small team with five or less people, the establishment and

[9]The first four roles are also mentioned by Posner and Baecker (1992).

assignment of roles is rather straightforward. A small team is also able to handle flexible role assignments as changes can be announced and reacted to quickly. In online mass collaboration, however, this is not necessarily the case due to the potentially large number of authors, who are involved to very diverse degrees in the CW process. Occasional authors might not be aware of (the meaning of) roles, which will make coordinating the CW process more complicated. If, for example, a senior co-author modifies a revision contributed by a newcomer to comply with consistency or quality standards, the newcomer might feel that his or her edit is not welcome. We will discuss CW roles in online mass collaboration in section 3.2.2 and in much more detail in chapter 5.

2.3.3 Computer Supported Collaborative Writing

Research on *computer-supported collaborative work* studies the use of computers to support CW (Sharples et al., 1993). Computer-supported collaborative writing is not to be confused with *computer-supported cooperative work*, although the two are closely related and share the same acronym (CSCW). Research on computer supported cooperative work focuses on group interaction and the use of computers for any kind of collaboration (Lowry et al., 2004, p. 92). The resulting tools are often referred to as *groupware*. Computer-supported collaborative writing is also different from *computer-mediated communication*. According to Lowry et al. (2004), computer-supported collaborative writing tools expand upon computer-mediated communication with extended support for coordination and communication, document sharing facilities, and process structures.

The use of computers to support CW is almost as old as the first extensive studies about CW itself. Starting in the late 80s until today, an exhaustive body of research has been carried out to better understand the needs of writers in the CW process and to build systems which adequately support users in this process (Leland et al., 1988; Galegher and Kraut, 1994; Rimmershaw, 1992; Posner and Baecker, 1992; Lowry and Nunamaker, 2003; Jones, 2005). The effect of computers on the writing process and on writing quality has been studied mostly in education. Daiute (1986) found that students writing on a computer correct their own errors more often as compared to students writing on paper. Goldberg et al. (2003) performed a meta-analysis on studies over a period of ten years, suggesting that digital writing improved the work of students both quantitatively as well as qualitatively. In addition, they found that students working with computers also collaborated more often as compared to those working in a pencil and paper environment.

Several case studies about computer-supported collaborative writing tools have been carried out, see e.g. Sharples (1993). Based on interviews with collaborative writers from different backgrounds, Rimmershaw (1992) argues that computer-supported collaborative writing tools need to support activity-based roles. The latter requires computer-supported collaborative writing tools to be very flexible since the CW practices applied by the authors

are very diverse. This finding is also confirmed by Posner and Baecker (1992), who applied a similar method to analyze the writers' needs in CW.

Many researchers argue that making proper use of computer-supported collaborative writing tools improves coordination and eventually document quality (Lowry and Nunamaker, 2003; Erkens et al., 2005; Passig and Schwartz, 2007), because these tools account for the requirements in CW projects as outlined in section 2.3.1. Passig and Schwartz (2007) argue that peer-to-peer (online) CW can produce better results than face-to-face CW. It is yet another question, how much use writers in industry and academia actually make of computer-supported collaborative writing tools and how such tools influence their working and writing behavior (Jones, 2005). In the following section, we will address this question and discuss the most important online CW tools.

2.3.4 Tools for Collaborative Writing

In their study from 2004, Noël and Robert (2004) found that the predominant means of communication during computer-supported collaborative writing were email, face-to-face and phone. They found that very few CW projects were actually using specialized computer-supported collaborative writing tools. Due to the rapid change of technology usage, ten years later, the situation has substantially changed. Although specialized offline computer-supported collaborative writing programs never gained widespread acceptance, web-based CW platforms such as wikis have become common tools in many companies (Tapscott and Williams, 2008; Arazy et al., 2009). Due to that reason, we focus on introducing web-based computer-supported collaborative writing tools in this section, rather than presenting an exhaustive list of groupware supporting CW. Behles (2013) found that 85% of technical communication practitioners and students use online CW tools. Popular web-based CW tools include Zoho Writer[10], Google Drive[11], Etherpad[12], or Authorea[13]. Wikis, such as Twiki[14], Foswiki[15] and MediaWiki[16], are another widespread web-based tool for CW. We will discuss wikis and in particular, the MediaWiki-based Wikipedia, in section 3.2.

Many of the traditional offline computer-supported collaborative writing tools offer a generic or one-fits-it-all solution. However, the complexity of many CW projects required that computer-supported collaborative writing tools were extensible and customizable or at least able to interact with other specialized CW programs (Sharples et al., 1993). Several successful web-based CW tools, e.g. most wikis, are open-source, i.e. everybody can modify and extend their functionality. Typically, open-source software is supported by a

[10] `http://writer.zoho.com`, accessed May 25, 2015
[11] `http://drive.google.com`, accessed May 25, 2015
[12] `http://etherpad.org`, accessed May 25, 2015
[13] `http://www.authorea.com`, accessed May 25, 2015
[14] `http://www.twiki.org`, accessed May 25, 2015
[15] `http://www.foswiki.org`, accessed May 25, 2015
[16] `http://www.mediawiki.org`, accessed May 25, 2015

community which can flexibly react to the practical needs of authors. MediaWiki, the wiki software used by the online encyclopedia Wikipedia, is an impressive example of a successful computer-supported collaborative writing tool, which has been continuously updated according to the needs of the CW community in Wikipedia.

2.4 Conclusion

In this chapter, we discussed two major shifts in research on writing. The first shift occurred in the late 1970s, with more and more scholars and teachers placing emphasis on writing as a non-linear process, involving various iterations of revision. The second shift was caused by technological developments enabling writers to work jointly on a shared document from various locations, both simultaneously and asynchronously. Both of these shifts have generated a substantial body of research and had major impact on teaching in composition classes. The concepts associated with these shifts, writing as process and CW, will be very relevant to the remainder of this work.

Most scholars agree that CW potentially increases the quality of documents compared to single-author writing. Reynolds et al. (1911) already stated at the beginning of the 20th century:[17] "Thus, the three of us have done together, as well as we could, what neither of us separately could have done at all – which, surely, is the essence of collaboration." At the same time, CW increases the complexity of the writing process a lot. As a result, many studies highlight the importance of coordination during CW. Revision and revision control help to coordinate the process of collaboration, but are not sufficient by themselves, as writers need to adjust their strategies, roles and activities to each other and the task at hand. A lot of essential cognitive steps during the writing process remain hidden for co-authors, e.g. trains of thought during drafting. Those needs are addressed with the help of computer-supported collaborative writing, where specialized tools for CW have been developed. While most of the traditional offline CW tools were never used on a large scale, new technologies and in particular, web-based tools such as wikis, have made computer-supported collaborative writing available to a wide audience. Online CW projects such as the open encyclopedia Wikipedia, have attracted millions of authors, who collaboratively create and revise documents. In the next chapter, we will present a more detailed discussion of CW in online mass collaboration and the underlying technologies, with a focus on wikis.

[17] Quoted from Sharples et al. (1993), p. xii.

Chapter 3

Online Mass Collaboration

The web has become the predominant market for collaboration in recent years. People are spending a substantial amount of their time online, and growing availability of high-speed internet access on mobile devices is likely to make this trend continue.[18] The biggest social network Facebook has more than one billion users.[19] Obviously, only few of the many users who are *interacting*, are actually *collaborating*. Online mass collaboration is a typical example of *peer production*, a phenomenon studied in several contexts, e.g. the open-source software community (Benkler, 2002, 2006). Peer production communities can create high-value and high-quality resources such as the operation system Linux, and work very differently as compared to traditional economic production systems. CW platforms on the web usually form peer production communities, with a typically large number of authors producing a shared good, i.e. one or more documents (Viégas et al., 2007b).

In this chapter, we present the theoretical foundations of CW in the web and in particular in Wikipedia. As discussed in chapter 2, coordination among the participants of the CW process is of crucial importance to create high-quality documents. The latter is true for most online peer production projects. Consequently, section 3.1 will deal with strategies in which users of online collaboration platforms coordinate their writing, while in section 3.2, we turn to one of the most successful and popular online CW projects, the encyclopedia Wikipedia. We will take a look at Wikipedians, the people behind Wikipedia. Further, we will analyze the writing process in Wikipedia, which is driven by the concept of revisions. To complete the analysis of user interaction in Wikipedia, we also need to consider the concept of discussion pages, a space that serves CW participants to meet, discuss and coordinate writing-related tasks. Finally, we will discuss some aspects of document quality in Wikipedia.

[18] This has been shown by several studies. For the US, see the following link to an article in the Huffington Post: `http://www.huffingtonpost.com/2013/08/01/tv-digital-devices_n_3691196.html`, accessed May 25, 2015

[19] `http://online.wsj.com/article/SB10000872396390443635404578036164027386112.html`, accessed May 25, 2015

3.1 User Interaction as a Means to Coordinate Collaborative Writing in the Web

Coordination is an important prerequisite for successful CW experience (Allen et al., 1987; Erkens et al., 2005). Even within a closely defined task setting such as writing an encyclopedic article about a certain subject, coordination is one of the crucial factors determining the quality of the outcome (Kittur and Kraut, 2008). Diverging views on the subject, lack of consistency, and unclear division of roles and/or workload can severely complicate the CW process. The ability to interact before, during and after the writing process is a substantial part of most CW projects. Coordination in such projects can either be achieved by a governing body which coordinates, decides and mediates in a top-down manner, or through direct and indirect interaction between all participants. Online communities, especially open (source) communities, are known for a strong emphasis on democratic approaches to organization and governance structures (Forte et al., 2009). While most scholars agree that healthy virtual communities must provide policies, rules of conduct, and penalties on misconduct, the creation of such rules and policies is often developed in a bottom up approach (self governance) rather than imposed by the leaders of the community (Viégas et al., 2007b). The focus of this study is on user interaction rather than governance in online communities. However, where appropriate, we will also discuss concepts of power and control, especially with respect to Wikipedia.

Interaction between users, as an important means of coordination, can take place in two ways in online CW scenarios: through direct interaction and through indirect interaction. While direct user interaction is established through communication between the users, indirect interaction is happening without direct communication, but when two or more users are working on the same document or piece of document.

The CW axioms presented in 2.3 back this view of interaction during the CW process. Direct interaction involves social activity, e.g. by arguing over a neutral point of view for an article in Wikipedia. Group dynamics include direct and indirect interaction, as authors often react to edits by fellow authors, either by re-editing their writing (indirect) or via direct communication on the discussion page or via comments associated to edits.

We define *user interaction* in the context of online CW projects as a social action between two or more authors which has an effect on all participants in the action. This action is social in the sense that it (typically) responds to the actions of other authors, but it does not necessary involve any kind of relation between the authors. There are many ways for CW participants to interact, however, we suggest that all of them can be classified as either *direct* or *indirect*. Put simply, direct interaction *coordinates* the CW process and indirect interaction *produces* content. In sections 3.1.1 and 3.1.2, we will show how modern web techniques support these two ways of interaction and how to analyze them. Later, in section 3.2, we turn to the concept of wikis and how they support direct and indirect interaction.

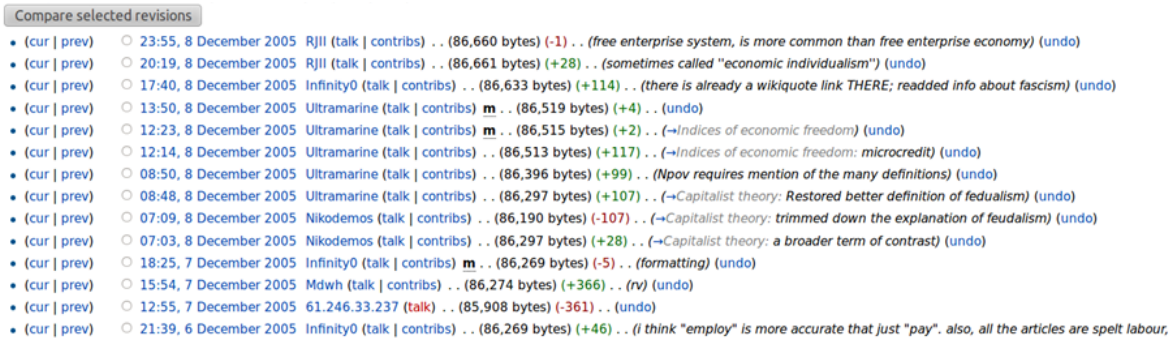


Figure 3.1: A snippet of the revision history of an article as displayed in the English Wikipedia.

3.1.1 Indirect User Interaction

We define *indirect user interaction* as a type of interaction between two or more authors without verbal or written communication. In the context of CW, indirect interaction happens when two or more authors are editing the same part (sentence, paragraph) of a document, or more general, the same document. There are several ways to analyze such interaction. We discuss three levels of exploration to analyze indirect interaction in online mass collaboration: on the level of a single revision, sequences of revisions and on a global level (entire revision history).

3.1.1.1 Collaborative Revision

At the textual level, CW means work on a joint document, which (sooner or later) involves editing a (piece of) text written by another author (called *co-author*). As already discussed in section 2.2.3, this results in a couple of challenges, which are usually technically solved by means of version control systems. On the web, different revisions of documents are often versioned and collected in a *revision history*. Basically, every co-author successively edits the document, and when finished, *commits* the changes, resulting in a new *revision* of the document. Figures 3.1 and 3.2 show two examples for revision history visualizations in online CW tools.

In mass collaboration on the web, conflicts are likely to arise when more than one author works on the same document at the same time. This problem can be addressed in several ways:

- **Document locking (single master copy)**: The author who starts editing a document first (in time), gets temporary ownership and the document is blocked until the current owner has finished editing, e.g. FosWiki; this basically corresponds to sending the document back and forth, e.g. via email.
- **Manually solving edit conflicts (master copy on a server, many local copies)**: Several authors are allowed to work on a single document, resulting in an *edit conflict*

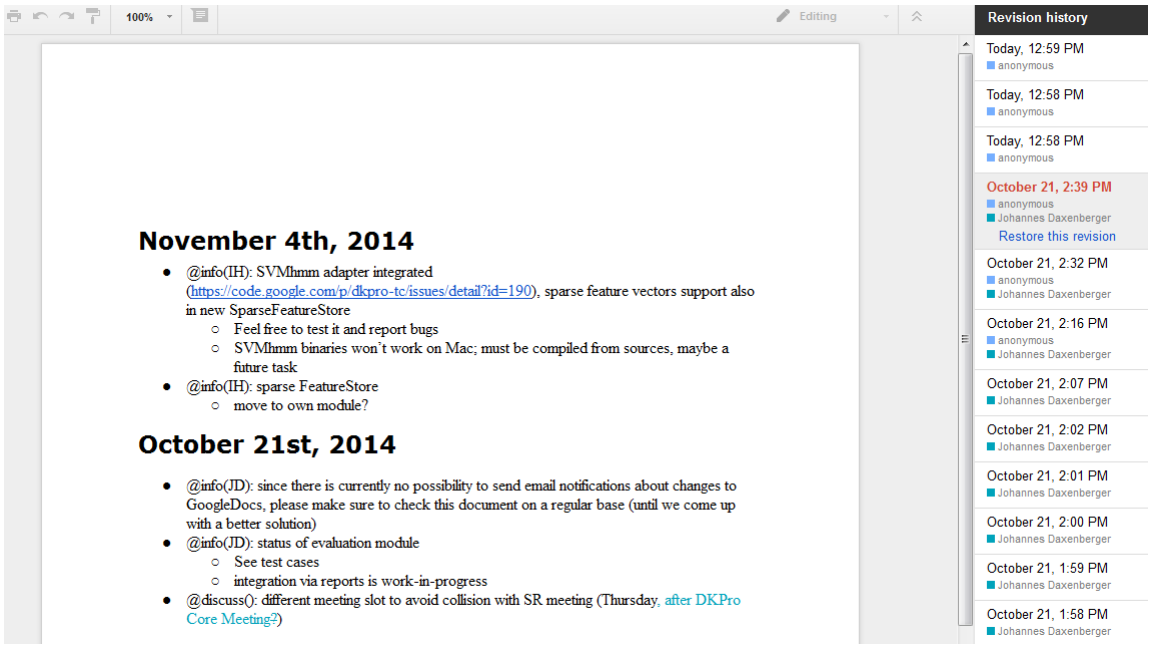


Figure 3.2: Revision history of a document created with GoogleDocs.

for those authors who finish editing after any of the co-authors who were editing at the same time. The conflict needs to be resolved manually by merging all pending changes, e.g. MediaWiki, and many version control systems for software development.

- **Real-time collaborative editing (master copy on a server, local copies and master are kept in sync)**: Several authors are allowed to work on a single document, and possibly resulting edit conflicts are solved automatically, e.g. Google Docs[20]

Regardless of the way in which the authors are presented with the fact that they are collaborating on a single document, they will automatically interact with their own or their co-authors' writing. This interaction can be supported by mechanisms such as *commit comments*, which authors publish together with the changes to the document. A commit comment can serve different purposes, but most importantly, it should summarize (and, if necessary, explain or justify) the changes to the document. Furthermore, many CW systems offer access to the revision history of documents. The revision history stores all versions of the document including the time stamps of their creation, their author, and possibly a commit comment. It is typically used to restore previous revisions of a document (*revert*), which is an important tool to combat *vandalism* (i.e. malicious edits by co-authors with

[20] The "operational transformations" and the "collaboration protocol" behind Google Doc's automatic edit conflict resolution are quite sophisticated. More details about the technology can be found in the following and subsequent blog entries: `http://googledrive.blogspot.de/2010/09/whats-different-about-new-google-docs_21.html`, accessed May 25, 2015.

bad intentions). Some CW systems additionally offer support for comparing two or more revisions. This can be as simple as displaying the number of added or removed bytes, but might also involve a complex visualization of the interaction by highlighting all changes. The functionality of such tools is important for complex CW tasks and may have substantial influence on the success of a CW system, as it helps authors to quickly understand revisions by their co-authors.

An important property of indirect interaction is the category of a revision, as explained in section 2.2. Human-readable information about the type of or the reason for a change can often be given in the commit comment. For example, an author intending to correct a spelling error, might indicate this change with a comment such as "typo". However, the length of commit comments is typically limited to a short number of characters, which might not be enough to explain all of the performed changes. Additionally, commit comments are usually not enforced. Consequently, a lot of revisions are not explained with the help of comments by their authors. If a revision contains several local changes, it is not clear to which of the multiple changes the comment refers. Consequently, to find out about the category of a revision, it is often necessary to visually inspect the differences between two revisions. By showing how revision categories can automatically be detected in Wikipedia, we will address this problem in detail in chapter 4.

3.1.1.2 Sequential Pattern Mining

Another approach to analyze indirect user interaction at a higher level makes use of the inherent chronological order of sequences of revisions. Revisions are discrete events which can be ordered based on their time stamps. Thus, the revision history of a document can be turned into a sequence. The revision history of a Wikipedia article, displayed in figure 3.1, represents such a sequence. Time stamps, authors and optionally edit comments are displayed along with each revision. Using sequential pattern mining (Mabroukeh and Ezeife, 2010), sequences of revisions can be searched for recurring patterns. In a mass collaboration system for CW, several attributes of revisions can be used to create patterns. A revision could be represented by the author who created it. In this scenario, a revision sequence models the succession of users editing a document. Another property which could be used to represent a revision is its category. In that case, the sequence models the succession of edit categories, e.g. a malicious edit followed by a revert. We refer to the latter as *collaboration pattern.*

3.1.1.3 Co-Author Networks

Indirect user interaction, at the global level, can be represented as a network of collaborating authors participating in the CW process, i.e. a co-author network. Social network research defines co-author networks as directed or undirected graphs where nodes typically represent agents, and edges between agents an interaction or relationship between

those agents (Aggarwal, 2011). In the case of indirect interaction in CW, nodes represent authors, and edges represent indirect interaction as defined above, i.e. editing the same (part of a) document. Co-author networks have been frequently used to analyze scientific networks, with edges representing a co-authored publication (Newman, 2001). Using co-author networks to analyze CW allows to apply network-specific measures and algorithms to understand collaboration. A promising approach is a motif analysis, which is able to uncover frequent collaboration patterns within the network (Krumov et al., 2011).

3.1.2 Direct User Interaction

We define *direct user interaction* as a type of interaction between two or more authors which involves verbal or written communication. In online mass collaboration, face-to-face communication is often impossible.[21] Instead, the communication which is necessary to coordinate the CW task takes place in written form, so that all participants are able to follow it.[22] There is a vast amount of online communication networks such as question and answer sites and discussion forums or groups (e.g. Slashdot, Stack Exchange, Google Groups). In this section, we are looking at technologies which offer functionality similar to communication networks, but are targeted towards online CW. Although the following list is not exhaustive, it covers the most important technologies currently used for written indirect interaction in CW projects. Online systems for CW often combine several of the listed technologies.

3.1.2.1 Comments within the document itself (asynchronous)

As an additional layer to the text that is written, comments may serve as a means to communicate between authors and to coordinate the current and future writing tasks. The advantage of comments within the document under revision is that they can be placed in specific locations, e.g. nearby the text segment they are referring to and that everybody editing this particular text segment will see them until deleted. To their disadvantage, comments are rather static and do not allow for long discussions with many participants.

Figure 3.3 shows the comment function in Google Docs.

[21] In this study, we do not consider video conferencing and other means to transmit speech or video at long distance. Please note that, in small CW projects with only a handful of people involved, the usage of such media to enable face-to-face communication might be of high importance to the CW task.

[22] The growing availability and usage of video chats, such as Google(Plus) Hangouts, in large-scale open online platforms (e.g. Massive open online courses, MOOCs) is changing this. Analyzing verbal and visual communication in mass online CW projects is out of the scope of this work, but will likely become a crucial factor in the near future.

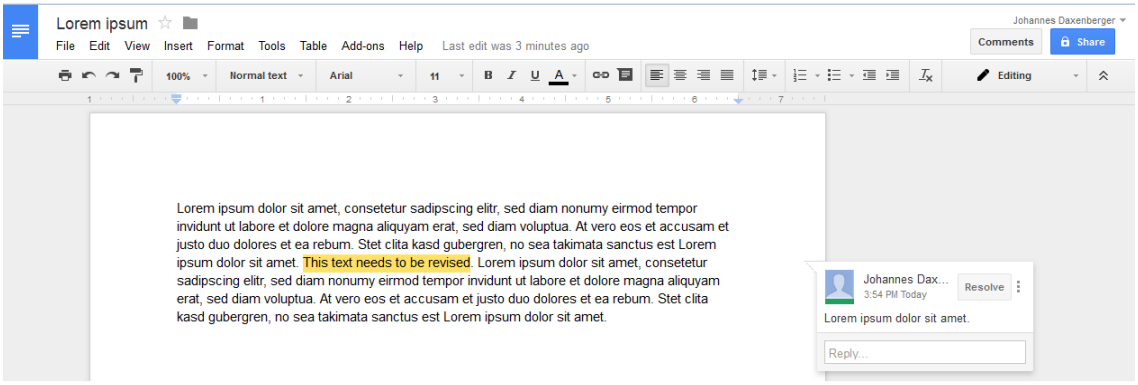


Figure 3.3: A document created with the online service GoogleDrive, showing the comment function.

3.1.2.2 Comments on a separate discussion space (asynchronous)

A simple form of direct communication which does not require additional technical effort is to define another document within the CW platform exclusively for communication purposes about the CW task. The advantage of this technique over commenting within the document is that it allows for a (more or less) structured discussion between two or more authors which is preserved for later reference. The level of support for a forum-like discussion space is often limited in real-world applications. Many wikis including Wikipedia offer pages exclusively dedicated to discussion. Rather than providing one page for all matters to be discussed during the CW process, larger CW projects such as Wikipedia offer dedicated discussion pages which are bound to discussion about a particular topic, e.g. an article in the encyclopedia. In addition to dedicated discussion pages, some CW platforms offer authors a way to hold asynchronous one-to-one discussions, e.g. personal discussion pages of individual users.

3.1.2.3 Real-time chatting (synchronous)

Some CW tools, especially those designed for rather small CW groups (e.g. Etherpad or GoogleDrive) offer an instant messaging service which is directly integrated within the CW system and enables real-time conversations. The disadvantage of instant messaging in the context of CW projects is that the chat protocol is often not available to all CW participants, so that relevant information about the content of a discussion might be lost. External services for instant messaging can obviously be used for direct interaction regardless of the CW system at hand. However, as such services are not bound to the document under revision, it cannot be guaranteed that all participants have access and are aware of the interaction.

3.1.2.4 Communication on external channels

Except for external chat forums, all of the previously listed mechanisms operate within the realm of the respective CW platform. However, direct interaction can also happen outside the particular platforms. A simple but popular way to communicate and coordinate CW projects is email conversation, and in particular, due to the higher visibility, mailing lists. Mailing lists can be run by the providers of the CW platform itself or external organizations, but they need to be clearly communicated to all authors, as these usually need to subscribe to the lists themselves. The advantage of mailing lists is that their focus can be restricted to certain subjects, so that authors can sign up for particular areas of their interest and expertise. Their disadvantage in the context of online CW is that they are not directly related to the product under revision. Thus, mailing lists often serve to announce and discuss higher-level organizational issues, e.g. technical innovations. Discussions on mailing lists are usually archived and can thus be searched for particular topics. Although discussion forums have evolved out of mailing lists and can be used in a similar fashion, mailing lists remain a very popular tool to spread information and discuss issues related to the CW process.

3.1.3 Further Aspects of Interaction in Mass Online Collaboration

Technical support for direct and indirect interaction (communication, editing and version control) is not the only factor for the success of an online CW system. Therefore, before looking at one of the most successful examples of mass online collaboration, we will address further issues which influence mass collaboration in practice. While the findings and contributions of this thesis are clearly focused around the analysis of indirect and indirect user interaction as defined in section 3.1.1 and 3.1.2, the following aspects of mass collaboration systems also influence direct and indirect user interaction.

Edit Notifications Awareness about the edit activity of other authors is an important requirement for participants of CW systems (Kirby and Rodden, 1995). In online mass collaboration, it is to be expected that not all authors will be active and online at the same time. Especially in long-term CW projects, authors might not be aware of new revisions to a document they have previously edited. Therefore, automatic *notifications* informing the co-authors about new revisions or comments play an important role in the CW process.[23] Notifications could be issued via email (e.g. commit notifications in open-source projects), either instantly or as a daily summary. However, email notifications might not be a convenient solution for frequently changing documents in mass collaboration, so that other means need to be used to inform the author about the changes they are interested in

[23] Manual notification by the author of a change might be common in small scale CW projects (Kim and Eklundh, 2001), but is not feasible in online mass collaboration systems.

(Moran et al., 2001). RSS feeds are an alternative to email notifications. However, change awareness might also be raised via intelligent real-time summaries of changes within the CW system itself. Like this, co-authors can follow the changes they are interested in wherever they want and instantly revise the respective documents (Tam and Greenberg, 2006).

Formal Organization Another crucial issue in online mass collaboration systems is formal organization, also referred to governance (O'Mahony and Ferraro, 2007; Butler et al., 2008). This involves pre- and post-task activities (cf. section 2.3), the creation and maintenance of policies, guidelines and other organizational tasks. The organization of a CW project is tightly bound to the task at hand. For example, the collaborative creation of a user manual for a product within a company has different legal implications as compared to writing an encyclopedic article in Wikipedia. Typically, a small number of authors are assigned to roles which grant them special user privileges, such as starting new or deleting existing documents. Depending on the degree of bureaucracy and on the process that is necessary to be granted special permissions, formal organization can impede the growth and dynamics of a CW system (Halfaker et al., 2012). However, in practice, such organizational measures are necessary to ensure quality and to avoid vandalism, especially in open mass collaboration.

Reputation and Feedback Another way to tackle quality issues and to improve the reliability of content in online mass collaboration systems are reputation and/or feedback mechanisms. McNally et al. (2013) model interaction in the social web along the dimensions of feedback and reputation. Feedback can be given either directly (user-to-user) or indirectly via giving feedback about an item produced by another user. Likewise, reputation can be explicit (user-to-user) or implicit by means of using different ways of expressing trust (e.g. by following a user on Twitter).

3.2 Collaborative Editing and User Interaction in Wikipedia

After we have discussed the general implications of user interaction in CW online platforms, we now turn to one of the most popular and well-known examples of a successful online CW project, the open encyclopedia *Wikipedia*. Since our analysis in chapters 4 through 6 is based on data from Wikipedia, a thorough understanding of the wiki technology and in particular Wikipedia is essential. In the following, we will explain why we have picked Wikipedia as the base for our analysis and show how the previously discussed CW concepts are applied in wikis.

Figure 3.4: The edit interface of the English Wikipedia, showing the article about "Technische Universität Darmstadt" including the source wiki markup (as of December 18, 2014).

3.2.1 Wikipedia Foundations

Wikis have been designed as content management systems with a particular focus on fast, simple and open collaboration (Leuf and Cunningham, 2001). The online encyclopedia Wikipedia is one of the most remarkable instances of a wiki. Wikipedia's slogan, "The free encyclopedia that anyone can edit", highlights this paradigm. Editing an article is as fast and simple as clicking the "Edit" button on top of the article and changing whatever necessary in the source of the article. Potential authors are encouraged to register before editing, but are not forced to do so. Figure 3.4 shows the editing interface of the English Wikipedia article "Technische Universität Darmstadt".

History Wikipedia was founded in 2001 by Jimmy Wales and Larry Sanger. Like its predecessor Nupedia (Sanger, 2005), which was founded one year earlier and closed in 2003 after only three years of operation, it was founded as a free online encyclopedia, written by volunteers. However, while Wikipedia is fully based on the wiki-technology which allows instant edits from anybody, Nupedia had an extensive peer-review system before articles got published. Wikipedia was originally designed as a draft platform to create articles for Nupedia with the help of mass CW. However, it quickly turned out that Wikipedia's concept

was much more successful than the expert-driven peer review system of Nupedia. As of October 2014, 287 languages have an official Wikipedia, and the largest project, the English Wikipedia, comprises more than 4.5 million encyclopedic articles. Wikipedia is run by the Wikimedia Foundation, which also develops the open-source wiki MediaWiki, Wikipedia's underlying software.

Table 3.1 lists the five largest Wikipedias, as measured by number of revisions. The English Wikipedia has already been edited more than half a billion times, and is currently changing by a rate of almost three million edits per month (see figure 3.6a). The German Wikipedia is the second largest, with only one tenth the number of registered users, but still over 1.7 million articles. The Spanish Wikipedia has more users than the German Wikipedia, but they are apparently less active.[24] Although it comprises almost 2 million articles, the Swedish Wikipedia is not mentioned in table 3.1, as it has a much lower revision count.[25]

Why Wikipedia? Wikipedia is a unique resource to discover the writing process in online mass collaboration. Given that Wikipedia is fully wiki-based (see section 3.2.1.1), it covers all the concepts discussed in section 3.1. In particular, it offers:

- a revision history for every page, including information about who has changed how much and when (indirect interaction), see section 3.2.3
- a discussion space with a forum-like structure for asynchronous communication, discussions are typically bound to single documents, e.g. encyclopedic articles (direct interaction), see section 3.2.5
- information about the quality of documents, see section 3.2.7

This covers most of the important features of research methodology to study the cognitive aspects of revision mentioned in Fitzgerald (1987, pp. 497f.). Furthermore, Wikipedia is seen as one of the most successful online CW projects (Giles, 2005; Mesgari et al., 2015).

Limitations Certainly, the encyclopedic nature of Wikipedia limits its text type and, to a certain extent, the composition of its authors. However, given that it covers a very broad range of topics and languages, it remains a unique resource to study the writing process in online mass collaboration. Furthermore, Wikipedia contains a substantial amount of non-encyclopedic content (mostly policy and direct interaction spaces; in the English Wikipedia, there are about the same number of encyclopedic articles and discussion pages[26]), which have made it the subject of many studies in human communication, and in particular, in discourse analysis (Viégas et al., 2007a; Ferschke et al., 2012a).

[24] The number of revisions might be skewed by bots, see section 3.2.2.2.

[25] Many, if not most articles in the Swedish Wikipedia have been created by bots, see `http://blog.wikimedia.org/2013/06/17/swedish-wikipedia-1-million-articles/`, accessed May 25, 2015.

[26] As of August 2014, excluding redirects and discussion archives.

Language	Article Pages	Pages	Revisions	Users
English	4,632,391	34,089,573	740,600,378	22,923,766
German	1,769,920	4,886,229	140,886,033	1,991,673
French	1,555,851	6,887,117	110,859,758	1,979,446
Spanish	1,134,038	4,703,567	83,234,527	3,341,609
Russian	1,158,230	3,930,865	78,759,898	1,408,521

Table 3.1: The number of pages in the article (main) namespace with at least one internal link, pages in all namespaces, number of revisions in all namespaces and users in the five largest Wikipedias, excluding redirects (as of October 27, 2014). Extracted from `http://wikistats.wmflabs.org/`, accessed May 25, 2015.

3.2.1.1 The Wiki technology

This section explains important concepts behind the *wiki* technology. Furthermore, we introduce most of the Wikipedia-related terminology used in chapters 4 through 6. Wikipedia is running on a MediaWiki system, hence most of its functionality is provided by this wiki.

Terminology and Technical Aspects We will use the term *page* to refer to any document in Wikipedia from any namespace, including articles, discussion pages, policy pages and others. The *namespace* system defined by the MediaWiki software distinguishes content pages and administrative pages. Table 3.2 lists and explains the namespaces in the English Wikipedia. In each namespace, there are *subject pages* and *discussion page*. The latter are used to discuss any issues about the content of the corresponding subject page. Hence, each subject page is bound to one ore more discussion pages, although discussion pages may also exist for non-existent subject pages and vice versa. An *article* is a page from the Main namespace, containing encyclopedic content. We refer to the Wikipedian who creates a new or edits an existing page as its author. Whenever an author saves changes to a page, a new revision of the edited page will be created. We call any version of a Wikipedia page a *revision*, denoted as r_v. v is a number between 0 and n, r_0 is the first and r_n the present version of the page, revisions are chronologically ordered. Registered authors can be identified by their user name, unregistered authors by the IP of the machine they are editing from. Wikipedia stores all textual changes of all authors for each of its pages. This way, it is possible to detect invalid or vandalistic changes, but also to trace the process of evolution of an article. Changes can be reverted. A *revert* is a special action carried out by users to restore a previous state of a page. Effectively, that means that one or more changes by previous authors are undone, mostly due to vandalism. Authors can revert the latest

Namespace	Usage
Main	Encyclopedia articles, lists, disambiguation pages, redirects
User	Pages of authors for personal use, e.g. to introduce themselves
Wikipedia	Content related to the Wikipedia project incl. policies
File	File descriptions for images, videos or audio files
MediaWiki	Used by the MediaWiki software to generate automatic messages
Template	Templates, i.e. structured content which can be included in other pages, e.g. to generate infoboxes
Help	Help for Wikipedia readers and authors
Category	List of pages and subcategories which were added to this category
Portal	Entry pages associated with certain topical areas or WikiProjects, mainly intended to help navigating through encyclopedic content
Special	Pages generated on demand, e.g. category trees
Media	Links directly to a media file (rather than its description in the File namespace)

Table 3.2: Major namespaces in the English Wikipedia. The prefix for each namespace is its name with a colon appended, except for the main namespace which does not have a prefix. Special and Media are virtual namespaces.

page version to any past state or edit it in any way they wish.[27] A revert will result in a new revision of the reverted page.

Layout and formatting in wikis are handled with a lightweight markup language which we refer to as *wiki markup* (also called wikitext). Wiki markup is a simplified version of HTML, with varying syntax between different wiki implementations. Frequent elements such as links or boldfaced text can be created using a simple syntax, e.g. `[[This is a link.]]`. Most wikis also support a range of HTML elements. MediaWiki (the wiki implementation running Wikipedia) also supports more complex elements which help to structure article text, so called *templates*. Templates are indicated by double curly brackets and are used to include text from other pages, creating standardized messages, infoboxes, or other automated text generation tasks.

Wikis as Collaborative Writing Tools The CW strategy applied in wikis is reactive writing (see section 2.3.1 and figure 2.3). Authors in a wiki are (at least potentially) simultaneously working on the same document, and their edits often react to changes by their co-authors. While wiki edits might be responding to other edits, they are typically not coordinated and preplanned, unless direct interaction techniques (cf. section 3.1.1) are used. We will discuss the implications and supporting technologies of this collaboration strategy

[27] However, pages can be protected from editing by privileged users, as stated in the Wikipedia Protection Policy, see `http://en.wikipedia.org/wiki/WP:Protection_policy`, accessed May 25, 2015.

in Wikipedia in detail and with examples in section 3.2.3 and section 3.2.5. The CW working mode in Wikipedia is Synchronous-distributed (see section 2.3.1), as several authors from different locations can work on the same article simultaneously. In Wikipedia, edit conflicts arising when several authors edit the same page at the same time (see section 3.1.1.1) are partially solved automatically. However, manual resolving can become necessary when several authors work on the same text segment at the same time.[28]

The nature of CW roles in Wikipedia can be defined along two dimensions (Callero, 1994; Merton, 1968): formal and activity-based (see section 2.3.2). *Formal roles* are determined by the rights and responsibilities of an author (e.g. administrator), whereas activity-based roles can be assigned based on the activity or behavioral patterns of authors (e.g. copy-editor) (Welser et al., 2011; Arazy et al., 2014; Laniado et al., 2011). To date, the literature on CW communities has paid particular attention to formal roles. Empirical studies investigating the organizational structure of Wikipedia have delineated a formal role hierarchy (Arazy et al., 2014; Stvilia et al., 2008). Other studies have described the functions and responsibilities associated with formal roles (Arazy et al., 2015; Butler and Sproull, 2007), and the literature on promotion processes explains how contributors progress between roles (Bryant et al., 2005; Burke and Kraut, 2008). Wikipedia has an extensive formal role system, which we will discuss in more detail in section 3.2.2.2. *Activity-based roles* in Wikipedia are a less studied area (Welser et al., 2011). Chapter 5 of this work will be dealing with this concept in depth.

3.2.2 Wikipedians

The authors in Wikipedia are called Wikipedians (as opposed to readers, who do not edit). A lot of research has been carried out to understand more about the nature of Wikipedians. Who is actually editing the encyclopedia, and with which motivation? Are articles written by experts, laymen, or both? Some of the facts that are repeatedly found show that content in Wikipedia is strongly biased considering its authors' demography.[29] This results in the so called *systematic bias*, given that Wikipedians do not constitute a representative sample of US-American citizens, and even less so, of the world population. Several researchers suggest that the systematic bias is a major threat to long-term quality assurance of Wikipedia (Reagle and Rhue, 2011; Callahan and Herring, 2011).

Fun, and the ideology to freely share knowledge are among the main factors motivating Wikipedians to edit, as indicated by several studies (Nov, 2007; Yang and Lai, 2010). A lot of studies have analyzed how article quality in Wikipedia relates to the so called wisdom

[28] Details about this process can be found here: `http://en.wikipedia.org/wiki/Help:Edit_conflict`, accessed May 25, 2015

[29] From `http://en.wikipedia.org/wiki/Wikipedia:Systemic_bias`, accessed May 25, 2015: the average Wikipedian is male, formally educated, an English speaker, aged between 15 and 49, from a developed nation, and from the Northern Hemisphere.

	English	Russian	Spanish	Japanese	German	French
2014	921,495	67,695	97,310	82,098	151,483	99,463
2013	846,675	61,156	87,131	74,351	142,719	90,414
2012	769,099	52,618	76,151	66,022	132,034	80,065
2011	684,931	43,394	64,850	57,522	120,261	69,343
2010	592,011	33,548	54,020	49,264	106,947	58,756
2009	492,948	22,288	42,470	40,550	92,476	47,607
2008	384,838	12,651	30,314	30,872	77,132	36,121
2007	264,942	6,593	17,936	19,884	57,725	24,155
2006	122,769	2,665	7,502	9,156	35,425	13,204
2005	33,992	671	2,055	3,355	16,543	4,557
2004	9,828	94	492	1,353	5,625	1,208
2003	2,521	10	102	132	753	229
2002	661		29		79	25
2001	84		1		7	1

Table 3.3: Number of Wikipedians who have created at least ten revisions in one of the six largest Wikipedias, August 2014. Extracted from `http://stats.wikimedia.org/EN/TablesWikipediansContributors.htm`, accessed May 25, 2015.

of the crowds. In other words: are high-quality articles created by a few experts or by many authors with different expertise? We will come back to this question in section 3.2.7. We can only discuss selected issues related to the users of the largest online encyclopedia and will therefore concentrate on the core aspects, i.e. the Wikipedia community and user interaction.

3.2.2.1 Wikipedia Community and Organization

Wikipedians form a community of mostly volunteers. The backbone of this community is the idea to work on a free and open online encyclopedia. Wikipedia's core principles (referred to as "Five Pillars"[30]) are:

- Wikipedia is an encyclopedia
- Wikipedia is written from a neutral point of view
- Wikipedia is free content that anyone can use, edit, and distribute
- Editors should treat each other with respect and civility
- Wikipedia has no firm rules

[30] `http://en.wikipedia.org/wiki/Wikipedia:Five_pillars`, accessed May 25, 2015

Beyond the fundamental principles, Wikipedia has developed an extensive body of policies and guidelines over the years (Reagle, 2010). The core content policies are: *Neutral point of view* (all significant views on a topic need to be represented fairly in the article), *Verifiability* (claims and material in articles must be referenced with reliable sources), *No original research* (material in articles must stem from reliable sources).[31] The English Wikipedia's "Manual of Style" sets standards for writing Wikipedia articles.[32] It seeks to promote clarity and consistency to its millions of articles, including aspects such as capitalization and punctuation, but also about how to write the lead section of an article.

With growing size of users, organization obviously became more important. The legal administration of Wikipedia is organized by the Wikimedia Foundation, a non-profit corporation with its headquarter based in San Francisco, USA. The administration of content in Wikipedia is mainly overseen by Wikipedians with special privileges (formal roles). Another means to organize content are WikiProjects (Morgan et al., 2014).[33] WikiProjects constitute a form of a subgroup in a CW project (Sharples et al., 1993, p. 19), and are organized around a certain topic or activity, e.g article quality assessment (cf. section 3.2.7).

User pages in the User namespace are intended to enable one-to-one direct user interaction, and to help authors to organize their work (Kittur et al., 2007b). User pages may also be used to display a limited amount of personal content which can help to overcome the deindividuation and anonymity of Wikipedians. A means to deal with conflict management (Sharples et al., 1993, p. 18) in Wikipedia are behavioral guidelines, such as "Assume good faith" (Generally assume that any author edits articles with the goal to improve Wikipedia) and "Don't bite the newbies" (Treat new authors with kindness and patience).[34]

Since 2012, the Wikimedia Foundation hosts a central knowledge platform called *Wikidata*.[35] As of September 2015, Wikidata stores more than 14 million data items such as entities, and associated statements like birth dates for persons.[36] Since 2004, Wikimedia's *Commons* platform stores media data including images, audio, and video files.[37] In September 2015, Wikimedia Commons contained more than 28 million files.[38]

Development of the Wikipedia Community As clearly shown in figure 3.5, there has been a significant break in the growth of the English Wikipedia community around the years 2006/2007 (Suh et al., 2009). Furthermore, the number of active Wikipedians has

[31] `http://en.wikipedia.org/wiki/Wikipedia:Core_content_policies`, accessed May 25, 2015
[32] `http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style`, accessed May 25, 2015
[33] `http://en.wikipedia.org/wiki/Wikipedia:WikiProject`, accessed May 25, 2015
[34] `http://en.wikipedia.org/wiki/Wikipedia:Assume_good_faith`, accessed May 25, 2015, `http://en.wikipedia.org/wiki/Wikipedia:Please_do_not_bite_the_newcomers`, accessed May 25, 2015
[35] `http://www.wikidata.org`, accessed May 25, 2015
[36] `http://www.wikidata.org/wiki/Special:Statistics`, accessed September 30, 2015
[37] `http://commons.wikimedia.org`, accessed May 25, 2015
[38] `http://commons.wikimedia.org/wiki/Special:Statistics`, accessed September 30, 2015

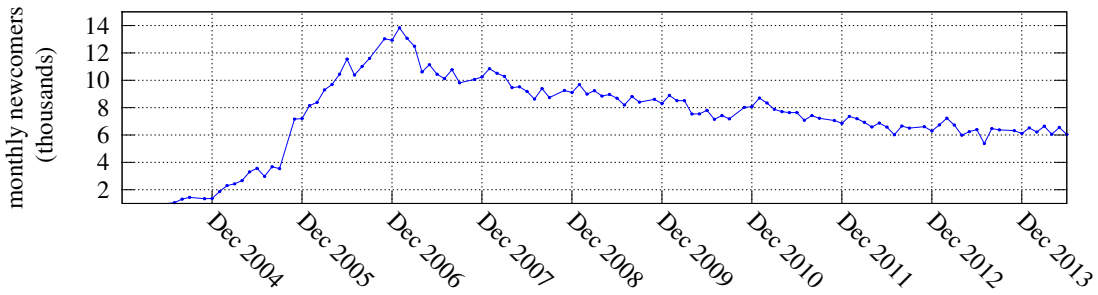


Figure 3.5: Newcomers in the English Wikipedia: Number of new Wikipedians who have created at least ten revisions since they started editing. Extracted from `http://stats.wikimedia.org/EN/TablesWikipediansNew.htm`, accessed May 25, 2015

passed its peak in 2007, and has been declining rather than increasing since then.[39] A bit later, the edit peak (revisions per month, cf. figure 3.6a) has also been passed. Halfaker et al. (2012) explain this shift with the introduction of quality and consistency management tools, which tend to reject newcomers' edits. This shift marks a significant change in the open collaboration paradigm of the English Wikipedia. Other Wikipedias have also been suffering this decline, although not all to the same extent. In the German Wikipedia, the number of monthly revisions also started decreasing in 2006, see figure 3.6b.

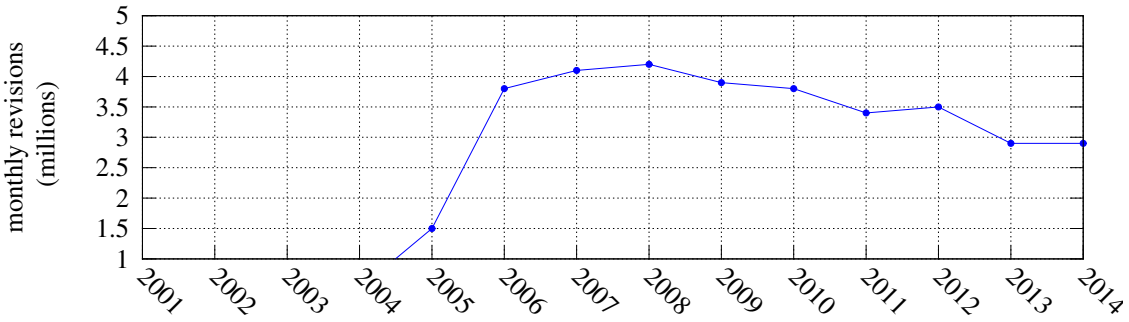
3.2.2.2 Further Aspects of Interaction in Wikipedia

In the following, we will shortly address aspects of interaction in Wikipedia which implicitly affect direct and indirect user interaction in Wikipedia (cf. section 3.1.3). Although these concepts are not directly covered in our research, we believe that they have a significant practical impact on CW in Wikipedia.

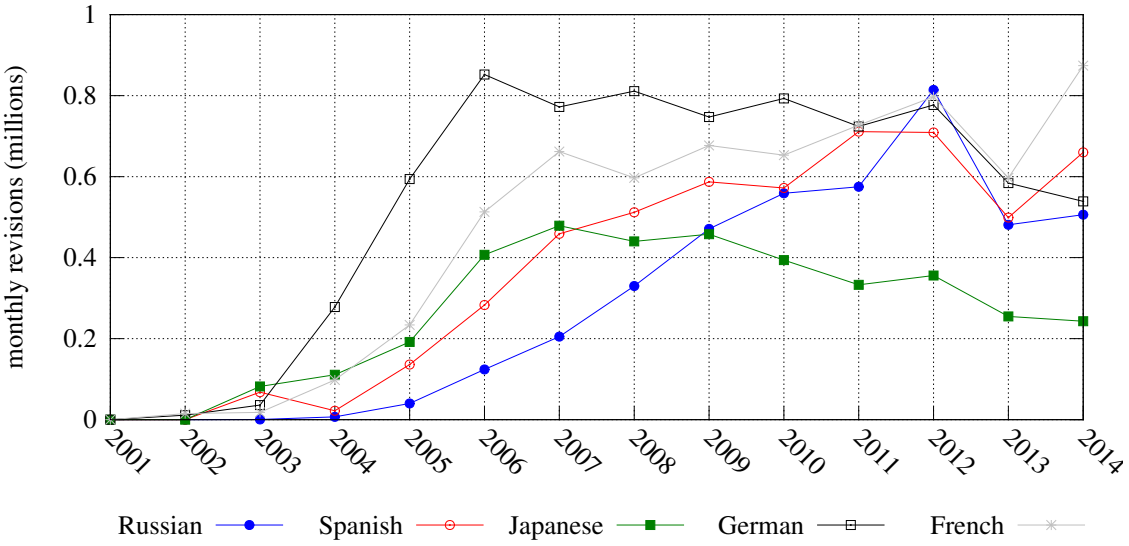
Edit Notifications As explained in section 3.1.3, edit notification are an important trigger of activity in CW systems. In Wikipedia, this is mostly done via *Watchlists*, which enable users to watch selected Wikipedia pages.[40] Other means to follow changes in Wikipedia are an IRC (Internet Relay Chat) server run by Wikimedia. The IRC sends automated messages about the activity in any language version of Wikipedia.

Formal Roles in Wikipedia The formal role system in Wikipedia (Arazy et al., 2014) is widely determined by the Wikimedia Foundation's *user group* system.[41] For the English Wikipedia, roles are defined via a *user access level*.[42] The user access level is bound to a

[39] See `https://stats.wikimedia.org/EN/TablesWikipediansEditsGt5.htm`, accessed May 25, 2015
[40] `http://en.wikipedia.org/wiki/Help:Watching_pages`, accessed May 25, 2015
[41] `http://meta.wikimedia.org/wiki/User_groups`, accessed May 25, 2015
[42] `http://en.wikipedia.org/wiki/Wikipedia:User_access_levels`, accessed May 25, 2015



(a) English Wikipedia: Number of revisions (millions).



(b) Russian, Spanish, Japanese, German, and French Wikipedia: Number of revisions (millions).

Figure 3.6: Number of changes per month in the largest Wikipedias. Extracted from `http://stats.wikimedia.org/EN/TablesDatabaseEdits.htm`, accessed May 25, 2015.

user account (either a manually registered account or via the IP of anonymous users) and determines whether a user is allowed to perform special actions. Some user access levels are automatically assigned (e.g. `autoconfirmed`), but most are manually assigned by a user with higher authority. Not all user access levels are used to manage privileges, but some also serve as flags to mark users (e.g. bots or blocked users). Table 3.4 lists access levels with more than 100 members active in one or more of the English Wikipedia namespaces listed in table 3.2.

Arazy et al. (2014) have analyzed the organizational structure behind the raw user access levels. Based on a study of 10,496 users with special access levels, they found that Wikipedia has a more hierarchical and bureaucratic structure than previously assumed. Flat hierarchies and openness are essential properties of any peer production system (Tapscott and Williams, 2008); hence the slowing community growth explained in section 3.2.2.1 can

User access level	No. Users	Permissions
`reviewer`	6,141	may review edits to protected pages
`rollbacker`	5,094	may perform instant reverts ("rollback")
`autoreviewer`	3,100	may create new articles which are automatically patrolled (rather than manually)
`sysop`	1,407	may perform several special actions incl. page deletion and blocking
`bot`	742	may perform edits which do not show up among recent changes
`filemover`	372	may work with and rename files
`ipblock-exempt`	239	may edit from previously blocked IP addresses
`abusefilter`	173	may create and modify edit filters
`accountcreator`	117	may create more accounts per day than default
`autoconfirmed`	appr. 14m	may move pages and edit semi-protected pages

Table 3.4: User access levels with more than 100 users in the English Wikipedia, numbers as of July 2014. The listed permissions are not exhaustive and might change over time.

also be explained with the gradual extension of quality assurance measures which increase governance and bureaucracy. The overall number of users with special access levels might seem small, however, considering that there are only about 30,000 monthly active users, this becomes a quite substantial number.[43]

Despite their small number (see table 3.4), *bots* carry out a significant amount of work in the English Wikipedia and other Wikipedias, e.g. the above mentioned Swedish Wikipedia, see section 3.2.1. Bots carry out diverse tasks such as vandalism combating (Geiger and Halfaker, 2013), spelling correction, category assignment and many tedious tasks related to formatting and markup of pages. During his 3-day study in 2013, Steiner (2014) found that about half of all revisions in all Wikipedia language versions and Wikidata are created by bots. He also found that the number of bot edits varies considerably across languages.

Reputation and Trust in Wikipedia One way to tackle quality problems and to improve the reliability of content in online mass collaboration systems are reputation and/or feedback mechanisms (cf. section 3.1.3). Wikipedia's reputation system is implicit, since users cannot explicitly rate each other, but there are ways to gain reputation, e.g. by frequent participation in discussion (Wöhner et al., 2011; Kittur and Kraut, 2008). Feedback in Wikipedia is mostly indirect, since it is typically given about an item written or edited by a user (i.e. a Wikipedia article). However, there are ways to give direct feedback to users, via so called Barnstars, which are rewards for special achievements and hard work (Kriplean

[43] See `http://stats.wikimedia.org/EN/TablesWikipediansEditsGt5.htm`, accessed May 25, 2015; the numbers are as of August 2014.

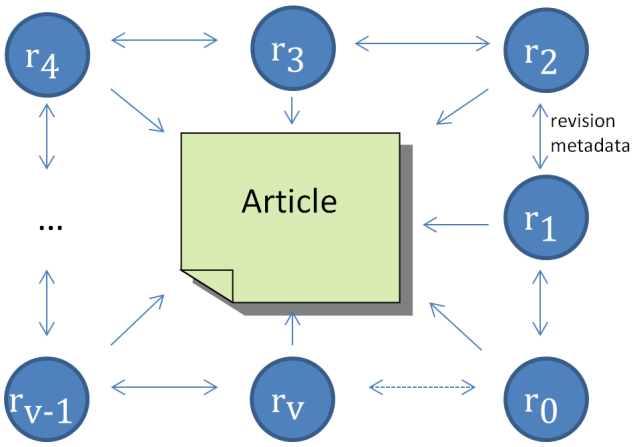


Figure 3.7: Reactive, collaborative writing in Wikipedia, adapted from Lowry et al. (2004).

et al., 2008).[44] They are issued by fellow Wikipedians, are free to give and usually placed on the recipient's user discussion page in the form of an image.

3.2.3 The Concept of Revision in Wikis

Wikipedia's revision history reflects a strictly indirect type of interaction between authors. Communication only takes place via meta data related to each revision in Wikipedia such as the author comment, the revision time stamp and the author's user name or IP address. Based on Lowry et al. (2004)'s definition, the writing process in Wikipedia can best be described as *reactive writing*, where real-time collaboration is possible and authors often directly react to changes by other authors as they adjust their own writing, see figure 3.7. We replaced the collaborating authors by their respective revisions (r_x), as several revisions can have the same author (reacting to their own change). Real-time collaborative editing of Wikipedia pages is possible only to limited extent, as it may likely cause edit conflicts; see section 3.2.1.1.

3.2.3.1 Revision history

The *revision history* of a Wikipedia page shows every revision of that page with a time stamp (date and time of creation), the author, an optional flag for minor changes applied by the author, the size of the changes in bytes and an optional comment given by the author, see figure 3.1. We call these items *revision meta data*, as opposed to the textual content of each article revision. The changes between pairs of revisions can easily be accessed through Wikipedia's web page by so called diff pages. *Diff pages* display a line-based comparison of the wiki markup text of two revisions (see figure 3.8). In particular, the diff page for a pair

[44] `http://en.wikipedia.org/wiki/Wikipedia:Barnstars`, accessed May 25, 2015

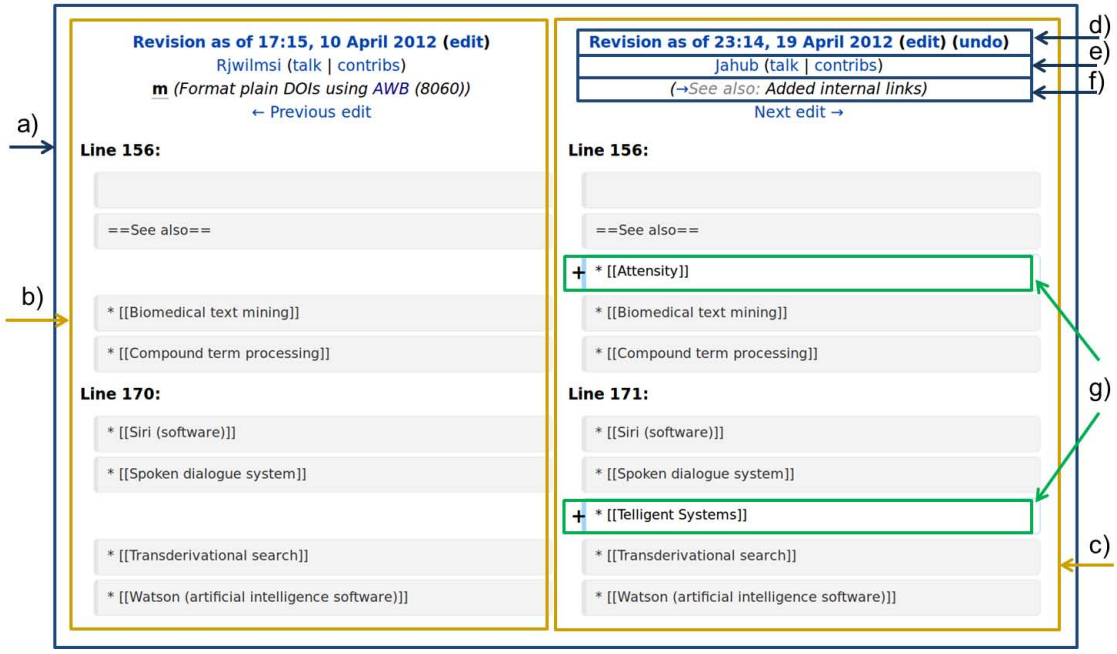


Figure 3.8: A diff page: *a)* entire diff, *b)* older revision r_{v-1}, only the changed part and its context are displayed, *c)* newer revision r_v *d)* time stamp with edit and revert ("undo") button, *e)* author, *f)* comment, *g)* individual edits. *d)*, *e)* and *f)* are meta data of r_v

of chronologically adjacent revisions r_v and r_{v-1} reflects the editing activity of one author at a certain point of time in the history of a page. We call the set of all changes from one revision to another a *diff*.

A single diff in an article's revision history can be reverted if subsequent changes do not conflict with it, i.e. modify text affected by the reverted diff. As changes can affect one or several parts of a page, a diff can consist of various edits. An *edit* is a coherent local change, usually perceived by a human reader as one single editing action. In figure 3.8, two consecutive revisions r_v and r_{v-1} are displayed in a diff page consisting of two edits inserting internal links. With respect to the meta data, the revisions in figure 3.8 have different authors. Both r_v and r_{v-1} are accompanied by comments. The time stamps indicate that the two versions have a time difference of approximately nine days.

Wikipedia is a huge data source for generating training data for edit category classification, as all previous versions (revisions) of each page in the encyclopedia are stored in its revision history. It is not surprising that the number of studies extracting certain kinds of Wikipedia edits with the help of rules or filters keeps growing. Among the latter, there are NLP applications such as the detection of lexical errors (Nelken and Yamangil, 2008), sentence compression (Nelken and Yamangil, 2008; Yamangil and Nelken, 2008), summarization (Nelken and Yamangil, 2008), simplification (Yatskar et al., 2010; Woodsend and Lapata, 2011), textual entailment (Zanzotto and Pennacchiotti, 2010; Cabrio et al., 2012), in-

formation retrieval (Aji et al., 2010; Nunes et al., 2011), paraphrasing (Max and Wisniewski, 2010; Dutrey et al., 2011), spelling error correction (Max and Wisniewski, 2010; Zesch, 2012), preposition error correction (Cahill et al., 2013), bias detection (Recasens et al., 2013), event detection (Georgescu et al., 2013), and fluency edit detection (Bronner and Monz, 2012).

Several researchers have studied *reverts* as a special kind of user interaction (Rzeszotarski and Kittur, 2012; Segall and Greenstadt, 2013; Flöck et al., 2012). Reverts typically express a negative relation between the author who reverts r_v and the author of r_v. They are mainly intended to quickly undo vandalism, however, users also (mis-)use reverts as a way to delete page content that does not reflect their opinion. When two or more authors apply reverts to fight over contradicting opinions, an edit war arises (Sumi et al., 2011; Yasseri et al., 2012).[45] The English Wikipedia has a long history of edit wars, with several odd examples making press appearance.[46]

One of the most serious problems in Wikipedia, caused by its open editing policy, is vandalism. About 6 to 7% of all revisions in the English Wikipedia are estimated to be vandalized (Buriol et al., 2006; Potthast, 2010). In short, vandalism or spam is "any addition, removal, or change of content, in a deliberate attempt to compromise the integrity of Wikipedia".[47] Vandalistic additions, removals or changes to an article can only be detected using revision history data, because at least two revisions need to be compared: a trustworthy, not vandalized revision r_{v-1} and a possibly vandalized revision r_v. Malicious edits are supposed to be reverted as quickly as possible by other users, which in practice seems to work quite well. Different median survival times for vandalized revisions are reported, ranging from less than three minutes (Viégas et al., 2004) to 11.3 minutes (Kittur et al., 2007b), depending on the type of vandalism. Several studies have addressed the important task of vandalism detection in Wikipedia. Vandalism detection in Wikipedia has mostly been defined as a binary machine learning task, with the goal to classify a pair of adjacent revisions as vandalized or not-vandalized based on edit category features. In Adler et al. (2011), the authors group these features into meta data (author, comment and time stamp of a revision), reputation (author and article reputation), textual (language-independent, i.e. token- and character-based) and language features (language dependent, mostly dictionary-based). They carry out cross-validation experiments on the PAN-WVC-10 corpus (Potthast and Holfeld, 2011). Classifiers based on reputation and text performed best. Adler et al. (2011) use a Random Forest classifier (Breiman, 2001) in their experiments. This classifier was also used in the vandalism detection study of Javanmardi et al. (2011) where it outperformed the classifiers based on Logistic Regression and Naive Bayes.

[45] http://en.wikipedia.org/wiki/Wikipedia:Edit_warring, accessed May 25, 2015

[46] Some not-so-serious edit wars can be found at http://en.wikipedia.org/wiki/Wikipedia:Lamest_edit_wars, accessed May 25, 2015. Also see http://news.slashdot.org/story/13/12/13/0056226/wikipedias-lamest-edit-wars, accessed May 25, 2015

[47] From http://en.wikipedia.org/w/index.php?title=Wikipedia:Vandalism&oldid=638930398. The same page also offers a list of frequent types of vandalism.

	Pfeil et al. (2006)	Jones (2008)	Liu and Ram (2011)	Antin et al. (2012)
Wikipedia Policy	Vandal. Reversion	Vandal. Revert Disambiguation	Revert	Vandal. Deleting Vandal.
Text-base	Add Inform. Delete Inform.	Signif. Addition Signif. Delet. Struct. Change	Sentence Creat. Sentence Delet. Sentence Modific.[a]	Creating Articles Adding Content Deleting Content
	Add Link Delete Link Fix Link	Add Link Fix or Delete Link	Link Creat. Link Delet. Link Modific.	
			Reference Creat. Reference Delet. Reference Modific.	Adding Citations
		Add Image Fix or Delete Image		
Surface	Style/Typogr. Spelling Grammar Format Mark-up Lang. Clarify Inform.	Style/Readability		Fixing Typos Reorganizing Text Chang. Wiki Markup Rephr. Exist. Text

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

Table 3.5: Three studies classifying revisions in Wikipedia and the categories they use.

3.2.4 Wikipedia Edit Category Taxonomies

Various studies have classified edits in Wikipedia; we compare them in table 3.5. 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 categories include Vandalism, Reversion (Reverts), Add Information, Delete Information, Clarify Information, Add Link, Delete Link, Fix Link, Style/Typography, Spelling, Grammar, Format, Mark-up Language. The taxonomy is based on an analysis of the data at hand, rather than existing research on revision. Two annotators manually examined and labeled the 500 revision pairs in their corpus. Revisions may be labeled with multiple categories. Jones (2008) analyzes differences in the CW 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, and includes the following categories: Vandalism, Revert, Disambiguation, Significant Addition, Significant Deletion, Structural Change, Add Image, Fix or Delete Image, Add Link, Fix or Delete Link, and Style or Readability. For the annotation process, he relies on revision comments that have been generated either by the authors or

automatically, but not on the actual edits. Liu and Ram (2011) created their taxonomy of Wikipedia edit categories aiming to analyze patterns of collaboration in Wikipedia. Since they applied rules to analyze edits rather than machine learning techniques, there is no explicit distinction between text-base and surface edits (cf. section 2.2.1). Their categories are limited to the following set: Sentence insertion, Sentence modification, Sentence deletion, Link insertion, Link modification, Link deletion, Reference insertion, Reference modification, Reference deletion, and revert. Antin et al. (2012) propose a list of ten categories, based on Kriplean et al.'s (2008) "editing work" types: Adding citations, Adding content, Changing Wiki markup, Creating articles, Deleting content, Fixing typos, Reorganizing text, Rephrasing existing text, Vandalism, and Deleting vandalism. Their taxonomy was applied in a crowdsourcing annotation study, where the annotators labeled entire revisions with one or more categories.

The Wikipedia revision taxonomies listed in 3.5 are applicable to tasks other their original use cases. Further taxonomies have been proposed, however their focus is limited to specific applications and thus hardly transferable to different problems. 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 distinction between content and format changes. Content includes text, links and images, format refers to HTML/CSS and templates. Bronner and Monz (2012) use very course-grain categories and only distinguish between factual and fluency edits. They segment adjacent revisions into edits and classify them in a supervised machine learning system. Further studies tried to detect reverts (Rzeszotarski and Kittur, 2012; Flöck et al., 2012).

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 with a category from the set of categories assigned to a pair of adjacent revisions.

3.2.5 The Concept of Discussion Pages in Wikipedia

The support for direct interaction is an important part of a CW system, cf. section 3.1.2. In Wikipedia, the most prominent tool for direct interaction are the so called talk or *discussion pages* (Viégas et al., 2007a; Wang and Cardie, 2014; Ferschke et al., 2012a; Kittur and Kraut, 2010; Kittur et al., 2007b). The CW axioms explained in section 2.3 also cover the importance of direct interaction tools to build consensus and enable coordination and communication. Although Wikipedia has created policies such as "neutral point of view" as a means to help building consensus and to decide over the relevant and irrelevant content, much coordination is necessary to put the theory into practice when CW takes place. Wikipedians make heavy use of discussion pages so that the discussions for popular or controversial articles

can grow really large. To maintain an acceptable level of clarity on vivid discussion pages, they get archived (either manually or automatically) after exceeding a certain size or age.[48]

As of July 2014, 9.2% of all revisions in the English Wikipedia Main and Talk namespaces are revisions of discussion pages, showing that discussion plays an important role in Wikipedia; but indirect interaction still outscores direct interaction by 10:1 (Kittur and Kraut, 2010). Three years earlier, in April 2011, this number was 9.7%, so it is to be assumed that this number has been rather stable over the last years. As of July 2014, the total number of discussion pages in the English Wikipedia is 4.85 million (4.91 including archived pages), as compared to 4.64 million article pages. The existence of a discussion page does not necessarily mean that there is an active discussion or any discussion at all going on about the article content. Many discussion pages only contain boilerplate content like information about the WikiProject responsible for this article, generated via templates (cf. section 3.2.1.1). As opposed to discussion pages in the Main namespace, discussion pages in the User namespace (user discussion pages) are intended for one-to-one direct user interaction, e.g. to award Barnstars (cf. section 3.2.2.2).

3.2.5.1 Usage, policies and best practices of discussion pages

The English Wikipedia guidelines for the usage of discussion pages state that discussion pages should not be used to express personal opinion, but to coordinate the development of the article with the goal to improve the encyclopedia.[49] According to Schneider et al. (2010), discussion pages are used mainly for:

- requests/suggestions for editing coordination
- requests for information
- references to Wikipedia guidelines and policies
- references to internal and external resources
- references to vandalism or controversial edits
- requests for peer review
- references to edits in the corresponding article
- requests for help

In Schneider et al.'s (2010) sample, requests for coordination were the most frequent usage category, outscoring all other categories by far.

[48] `http://en.wikipedia.org/wiki/Help:Archiving_a_talk_page`, accessed May 25, 2015
[49] `http://en.wikipedia.org/wiki/Wikipedia:Talk_page_guidelines`, accessed May 25, 2015

Add a link?

Shouldn't borax be wikilinked in the "etymology" paragraph? JCM83

> It's already linked first in the lede and also later in the mineral section. I put in a third link, but that's three. Number of links for the same term in an article should be limited. Here it's when used for the first time, then then maybe when needed to stimulate the memory. A new one in etymology is okay with me, and then the section about borax minerals--- but that's it! SBHarris 05:52, 16 February 2011 (UTC)

Figure 3.9: A topic from the discussion page of the English Wikipedia article "Boron" with two turns.

Discussion pages are technically identical to article pages, with the exception that they are located in a different namespace. However, discussions should be structured in a thread-like manner, i.e. with named topics and posts by users. Users can respond to posts by other users or create a new topic. Ferschke et al. (2012a) analyzed the structure of discussion pages in depth and refer to a user post in the sense of the smallest unit in a discussion page as *turn*. On a higher level, discussion pages are segmented into *topics* based on the structure of the page (in Wikipedia, topics are separated by headlines). Ferschke et al. (2012a) retrieved individual turns from topics by considering the revision history of the discussion page. This procedure successfully segmented 94% of all turns in a corpus from the Simple English Wikipedia. Each turn is associated with meta data, namely the name of the user who added the turn, the time stamp, and the name of the topic to which the turn belongs. Figure 3.9 shows a topic ("Add a link?") from a discussion page with two turns. The first turn has not been correctly *signed*, as it only contains the user's signature, but not the date.

3.2.6 Wikipedia Co-Author Networks

Several studies have analyzed co-author networks and collaboration patterns in the Wikipedia revision history (Brandes et al., 2009; Laniado and Tasso, 2011; Liu and Ram, 2011; Sepehri Rad et al., 2012; Wu et al., 2011). In most of these studies, networks of one or more articles were created. In these, nodes correspond to the authors of the article(s) and/or the articles themselves and edges represent a kind of direct or indirect interaction between authors or between authors and articles: editing the (same) article, editing the same sentence within an article, discussing the same topic etc. Laniado and Tasso (2011) find that "core" authors tend to interact with "peripheral" Wikipedians rather than among themselves, based on an analysis of the English Wikipedia. Laniado et al. (2011) analyzes a Wikipedia co-author network for discussion pages, based on the following kinds of (direct) interaction: direct replies on discussion pages, direct replies on user discussion pages, and messages on user discussion pages. They find that discussions are mainly focused on solving controversies according to the respective Wikipedia policies.

3.2.7 Aspects of Article Quality in Wikipedia

Although the quality of the information content produced through CW is not the main focus of this study, it is an important aspect of the CW process in Wikipedia. Furthermore, in chapter 4, we will show how to apply the methods we have developed to assess and understand article quality in Wikipedia. Ever since Giles (2005) suggested that "Wikipedia comes close to [Encyclopaedia] Britannica in terms of the accuracy of its science entries", research has tried to understand the development of high-quality articles in Wikipedia and how these distinguish from the ones with lower quality. A recurring question within the context of mass online CW is whether high-quality documents are written by few experts or by many laymen (Kittur et al., 2007a; Feldstein, 2011; Kittur and Kraut, 2008; Warncke-Wang et al., 2015). Onrubia and Engel (2009) have shown for the context of CSCL environments in university student groups, that "[...] in many cases, the final product is not the result of real joint construction, but a juxtaposition of the individual contributions, more or less controlled and directed by one of the group members." Although several studies show that the vast amount of work in online mass collaboration is done by few authors (Feldstein, 2011; Priedhorsky et al., 2007), the importance of the "long tail", i.e. thousands of authors with few edits, is also acknowledged (Wilkinson and Huberman, 2007; Kittur et al., 2007a). In line with previous literature on CW, some argue that successful mass collaboration is only possible when the "the wisdom of the crowd" is coordinated properly (Arazy et al., 2010; Wilkinson and Huberman, 2007; Kittur and Kraut, 2008).[50]

Stvilia et al. (2007, 2008) identify different types of information quality problems in Wikipedia and suggest actions to be taken to prevent them. There are many approaches to measure and/or predict information quality in Wikipedia: simple metrics based on meta information, e.g. number of words, revisions, or discussions (Wöhner and Peters, 2009; Han et al., 2011; Blumenstock, 2008; Stvilia et al., 2008; Wilkinson and Huberman, 2007), but also metrics based on the text itself such as readability or style (Hasan Dalip et al., 2009; Stvilia et al., 2007). To the opposite, indicators for quality problems in Wikipedia have also been studied. In Wikipedia, certain (known) problems (e.g. missing references) are marked with *quality flaw* markers (Anderka et al., 2012; Ferschke et al., 2013). These markers can be used to learn more about the underlying problems with respect to information quality. For an extensive discussion about mass collaboration and information quality in Wikipedia, see Ferschke (2014).

Wikipedia-Internal Quality Criteria Wikipedia has developed extensive guidelines and policies for improving the quality of its content. One of these efforts is the awarding of *fea-*

[50] Also see this interesting read, about the influence of stewards: https://medium.com/matter/the-36-people-who-run-wikipedia-21ecca70bcca, accessed May 25, 2015

Class	Criteria	Articles (%)
Featured article	all featured article criteria	0.11
A	essentially complete content, good organization	0.03
Good article	all good article criteria	0.47
B	mostly complete content	2.07
C	substantial content, might lack reliable sources	3.94
Start	incomplete content, lacks reliable sources	25.61
Stub	very basic content	54.10

Table 3.6: The quality classes in the English Wikipedia defined by the Wikipedia 1.0 Assessment Team (as of November 4, 2014).

tured article[51] or *good article*[52] status. Awarded articles have been manually assessed and confirmed to fulfill quality criteria defined by the community.[53] These criteria state that articles should be well-written and comprehensive, follow a certain structure, and contain images where appropriate, among others. Articles which do not fully conform to these criteria, can also be labeled with other "classes". The Wikipedia 1.0 Assessment Team evaluates articles on a scale from very short and incomplete articles to featured articles.[54] As can be seen in table 3.6, the percentage of high-quality article is quite low compared to the overall number of articles in the English Wikipedia.

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 of the Wikipedia 1.0 Assessment Team is substantial. Several studies have compared featured articles with non-featured articles based on a wide range of properties of the articles (Wilkinson and Huberman, 2007; Jones, 2008; Lipka and Stein, 2010; Stvilia et al., 2008, among others). Distinguishing featured articles and non-featured articles has been shown to be rather easy, with state-of-art approaches achieving around 0.9 F1 score (Lipka and Stein, 2010). Further research has also used the full Wikipedia 1.0 Assessment Team range of classes to learn about article quality in Wikipedia (Warncke-Wang et al., 2013; Hasan Dalip et al., 2009).

It should be noted, that due to the collaborative and open nature of Wikipedia, article quality is a moving target. An article might be featured at a certain point of time after going through a peer reviewing process and at a later point be demoted from featured article status in the same process. This can happen either because the quality of the articles has lowered despite continuous editing, or because the article did not change while the quality standard did increase over time. As the quality standards applied by the community may

[51] `http://en.wikipedia.org/wiki/Wikipedia:Featured_articles`, accessed May 25, 2015

[52] `http://en.wikipedia.org/wiki/Wikipedia:Good_articles`, accessed May 25, 2015

[53] `http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria`, accessed May 25, 2015, resp. `http://en.wikipedia.org/wiki/Wikipedia:Good_article_criteria`, accessed May 25, 2015

[54] `http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment`, accessed May 25, 2015

change over time, inconsistencies in the quality of articles assessed at different times are to be expected. For example, the article on "Windows XP" has been featured in 2005, but was demoted from featured article status in 2008 due to a lack of appropriate and consistently formatted citations.[55]

3.3 Conclusion

In the past 15 years, successful online CW projects such as Wikipedia have shown that mass collaboration with little or no explicit governance is possible. Wikis have become a very popular CW tool and are an integral part in many companies, classrooms and online platforms. Certainly, wikis and other web-based collaboration tools are not the only way to jointly create documents, but less sophisticated tools can also produce good outcomes (Dishaw et al., 2011). However, online platforms have become a widely used instrument to enable and improve collaborative document creation (Behles, 2013; Arazy et al., 2009; Lowry and Nunamaker, 2003; Majchrzak et al., 2006).

As stated by Sharples et al. (1993, p. 23), "we need to analyze the patterns of interaction [... in CW...] tasks". To do so, we have defined interaction in CW along two dimensions, direct and indirect interaction. We have shown how these dimensions are supported by wikis and in particular, Wikipedia. We note that our definition is not the only way to characterize interaction in Wikipedia.[56] However, as we will show in chapters 4 through 6, it is a comprehensive way to structure and analyze the various patterns of interaction in online mass collaboration. We have shown that online mass CW is a complex phenomenon which involves several technical, cognitive and social aspects. Several aspects of CW, such as the social conflict which is implied in changing other people's text (Birnholtz and Ibara, 2012), have not been discussed in this chapter. Rather than covering all aspects of CW in this work, we focus on certain aspects of the writing and coordination process inherent to online mass CW.

Precisely, we address three important concepts of CW which can be analyzed by the example of Wikipedia. *Indirect interaction* will be the main focus of chapter 4. Based on the aspects introduced in chapter 3.2.3 and 3.2.4, we will analyze the content of and intentions behind edits in Wikipedia and explain what our findings reveal about the CW process in online mass collaboration. Second, we will discuss *activity-based CW roles* in chapter 5. We will show how the Wikipedia community is naturally dividing editing tasks based on the

[55] `http://en.wikipedia.org/wiki/Wikipedia:Featured_article_review/Windows_XP/archive1`, accessed May 25, 2015

[56] For example, the type of collaboration in Wikipedia has also been described as stigmergy, i.e. a kind of self-organization, where each action builds upon another action, resulting in a seemingly intelligent structure. A discussion on the Wikipedia Research Mailing List from 2012 has addressed this issue in detail, see `http://lists.wikimedia.org/pipermail/wiki-research-l/2012-July/thread.html#2231`, accessed May 25, 2015.

authors' preferences and discuss the implications of this finding for the overall CW process in online mass collaboration. Finally, in chapter 6 we analyze the CW strategy applied by Wikipedians as an *interface between direct and indirect interaction*. In particular, we will discuss and evaluate our approach to link direct interaction in discussion pages to edits on the article pages.

Chapter 4

Wikipedia Revision Classification

In this chapter, we present our in-depth investigation of indirect interaction in Wikipedia. To find answers to the questions raised in chapter 1, it is essential to understand the details of the processes underlying indirect user interaction in mass online collaboration. As discussed in the previous chapter, indirect interaction through working on the same document, often reacting to changes by co-authors, makes up the majority of time and effort invested by the authors of the open online encyclopedia Wikipedia (90% when measured in revisions of articles as compared to revisions of discussion pages). As such, Wikipedia is a very rich resource to study indirect user interaction in practice. To this end, we will address the following research questions:

1. What is the content of and the intention behind revision in Wikipedia?
2. How can we automatically categorize revisions in Wikipedia?
3. Is there a relationship between article quality and indirect interaction in Wikipedia?

To analyze the revision process in Wikipedia, we first need to establish a way to extract and process revisions (section 4.1). We already introduced the concept of *edits* in section 3.2.3.1. Edits are local modifications, calculated from consecutive pairs of wiki revisions, and will be our unit of analysis throughout most of this chapter. To answer the first research question, we designed a novel taxonomy of edit categories in Wikipedia (section 4.2). In section 4.3, we present a hand-tailored corpus of English Wikipedia revisions. Next, our edit category taxonomy is used to label edits in this corpus in a manual annotation study. Addressing the second research question, we propose a novel set of features for machine learning algorithms, and evaluate the classifier trained on this data. We test the language dependency of this model in section 4.4, by transferring knowledge from the English model to a classifier for German revisions. In section 4.5 we test our machine learning classifier on a different taxonomy for Wikipedia revisions rather than individual edits, and show

that our proposed feature set works well on a different taxonomy. Finally, addressing the third research question, we relate our findings about edit categories in Wikipedia to article quality (section 4.6).

4.1 Extraction and Segmentation of Wikipedia Revisions

The Wikimedia Foundation regularly provides database dumps of the wikis it is hosting. The dumps are stored in an XML format, including the source texts and metadata.[57] Wikipedias are dumped at least once a month, under a Creative Commons license.[58] The compressed sources of the English Wikipedia including all revisions sum up to more than 100 GB. This size is partly caused by the fact that the current XML format of the dumps encodes the full text of each revision.

4.1.1 Extracting Consecutive Revisions

The raw data for our analysis is extracted from Wikipedia dumps, as explained above. We process the revision content (text with markup) and metadata using the Java Wikipedia Library JWPL (Zesch et al., 2008) and the Wikipedia Revision Toolkit (Ferschke et al., 2011). The revision text is not parsed, we keep the sources including wiki markup. Each article in Wikipedia is associated with a chronologically ordered list of revisions $r_0, r_1, r_2 ... r_n$, where r_0 is considered an empty revision without metadata. We extract pairs of consecutive revisions r_{v-1}, r_v, with $0 < v \geq n$, so that each new revision is represented by a pair (the newly created revision and the previous revision, as determined by their time stamps). The creation of a new article is represented by the pair r_0, r_1, and the first revision to this newly existing article as r_1, r_2. In Wikipedia, revisions are labeled with unique IDs, and we refer to a pair of consecutive revisions by the id of the newer revision. For instance, the diff page in figure 4.1 displays the changes made in the revision with the ID 488248874, as compared to the previous revision of this article, ID 486656058.[59]

4.1.2 Segmenting Revision Pairs into Edits

We implemented a customized segmentation algorithm for Wikipedia revision pairs, represented in figure 4.2 as pseudocode. For each (r_{v-1}, r_v)-pair, we calculate all of the n changes

[57] `http://dumps.wikimedia.org`, accessed May 25, 2015

[58] CC-BY-SA 3.0, a copyleft license which ensures that the content can be modified and freely distributed as long as the appropriate credits are given: `http://creativecommons.org/licenses/by-sa/3.0`, accessed May 25, 2015.

[59] Wikipedia's web interface allows to easily access diff pages, by appending the ID of the newer revision to the following link: `http://en.wikipedia.org/w/index.php?&diff=prev&oldid=`. E.g. `http://en.wikipedia.org/w/index.php?&diff=prev&oldid=488248874` can be used to access the diff page displayed in figure 3.8.

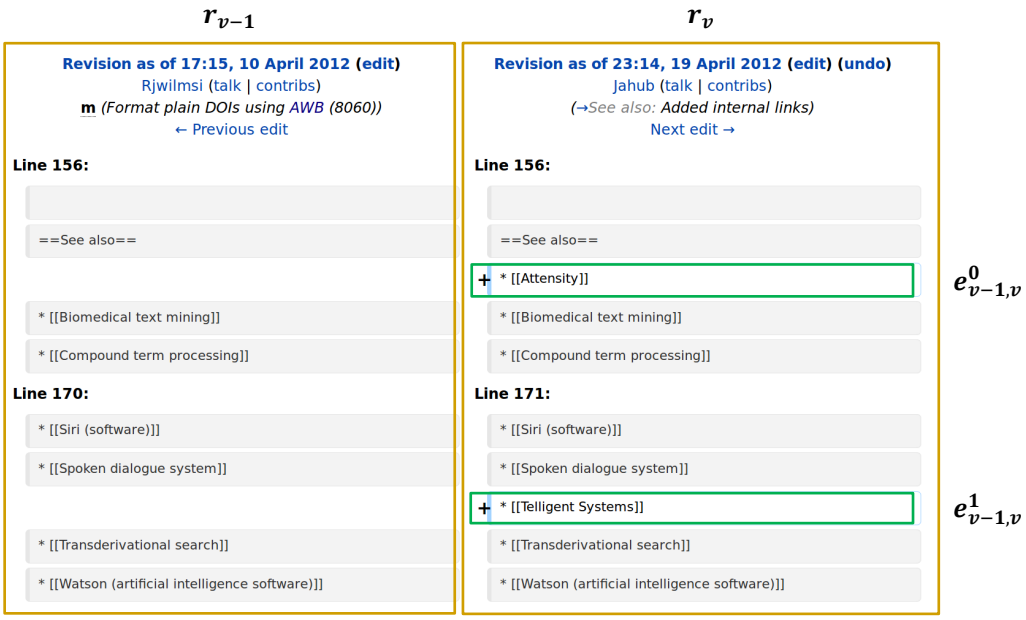


Figure 4.1: A Wikipedia diff page, displaying two consecutive revisions, with two edits.

$e^k_{v-1,v}$ that have been made to the newer revision via an adapted version of the diff comparison algorithm by Heckel (1978), with $0 \leq k < n$. The algorithm splits each revision into its lines (i.e. paragraphs) and numbers them. Then, it compares each line in r_{v-1} with each line in r_v to find differences in terms of the type of change (inserted, deleted, modified and relocated lines). Consecutive lines with the same type of change are merged into a single edit (figure 4.2, line 3).

Inside modified lines, we additionally detect and mark changes on character level (deletions, insertions and modifications) in situ using Neil Fraser's google-diff-match-patch library.[60] Figure 4.3 describes the post-processing of edits in more detail. The google-diff-match-patch library is a fast state-of-the-art framework for file comparison, based on the diff algorithm proposed by Myers (1986). To avoid splitting heavily edited lines into a very high number of counterintuitive edits, the last step is only executed where the ratio of the number of overall changes in that line to the number of tokens in that line does not exceed a threshold (figure 4.3 lines 11–23).[61] If, for example, stop words like "the" or "a" are the only unchanged segments inside a modified line, we mark the entire line as modified. We do further post-processing to recognize and merge associated edits. For example, we merge `[[` and `]]` for edits that add a link, as shown in figure 4.2 in lines 16–29. It should

[60] `http://code.google.com/p/google-diff-match-patch/`, accessed May 25, 2015
[61] The threshold values were determined empirically; for all of the experiments described in this work we set $\gamma = 5$, $\delta = 0.1$, $\epsilon = 0.3$, and $\zeta = 0.5$.

```
Data: the source text of a pair of consecutive revisions
Result: a list of edits
1  oldText ← source text of r_{v−1};
2  newText ← source text of r_v;
3  lineBasedEdits ← line-based diff algorithm by Heckel (1978) on oldText and newText;
4  processedEdits ← [];
5  foreach edit e ∈ lineBasedEdits do
6      if e.type = MODIFICATION then
           /* perform character-based diff, possibly splitting e                */
7          postProcessedEdits ← postprocess e;
8          add all postProcessedEdits to processedEdits;
9      else
           /* inserted, deleted, relocated lines: only add                      */
10         add e to processedEdits;
11     end
12 end
13 sort processedEdits ascending by position in source text;
14 processedEditPairs ← create list of all consecutive edit pairs in processedEdits;
15 wikifiedEdits ← [];
16 foreach editPair ep ∈ processedEditPairs do
17     e_0 ← ep.firstEdit;
18     e_1 ← ep.secondEdit;
       /* only for inserted, deleted, modified lines                           */
19     if e_0.type = e_1.type AND e_0.type != RELOCATION then
           /* if e_0 and e_1 contain matching opening and closing markup symbols */
20         if e_0, e_1 fulfill matching criteria then
               /* create new edit by merging edits e_0 and e_1 using the text from the
                  older revision                                                */
21             e_merged ← merge e_0, e_1 using r_{v−1};
22             add e_merged to wikifiedEdits;
23         else
24             add e_0, e_1 to wikifiedEdits ;                  /* only add */
25         end
26     else
           /* relocated lines or different type: only add                       */
27         add e_0, e_1 to wikifiedEdits;
28     end
29 end
30 return wikifiedEdits
```

Figure 4.2: Overview of the edit segmentation process. The post-processing procedure is explained in figure 4.3 (lines 5–12). Lines 15–30 show how logically associated edits are merged.

```
Data: an edit
Result: a list of edits
instantiate parameters γ, δ, ϵ, ζ;
oldLine ← text from edited line in r_{v−1} newLine ← text from edited line in r_v /* diff
    part types: inserted, deleted and equal                                    */
diffParts ← Character-level Fraser diff on oldLine, newLine;
ratioDiffPartsToTokens ← diffParts.size / newLine.numberTokens;
changedTextSize ← [];
foreach diffPart d ∈ diffParts do
    if d.type != EQUAL then
        changedTextSize += d.text.length;
    end
end
if diffParts.size < γ then
    /* this is a minor edit                                                   */
    if ratioDiffPartsToTokens < δ then
        /* split diffParts into edits based on their type                     */
        postprocessedEdits ← transform diffParts
    else
        add e_l to postprocessedEdits
    end
else
    /* this is a major edit                                                   */
    if ratioDiffPartsToTokens < ϵ and changedTextSize < ζ then
        /* split diffParts into edits based on their type                     */
        postprocessedEdits ← transform diffParts
    else
        add e_l to postprocessedEdits
    end
end
return postprocessedEdits
```

Figure 4.3: The post-processing part of the edit segmentation algorithm. It transforms a line-based edit into smaller units given certain criteria.

be noted that this process is not fully accurate as wiki markup is a context-sensitive language and hence difficult to parse (Dohrn and Riehle, 2011). The segmentation process is language-independent.[62]

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,

[62]The parameters γ, δ, ϵ, and ζ are not language-independent and might need to be adjusted for languages with significantly deviating average word and sentence length.

deletions, modifications (inside and across lines) and relocations (only on line level). Correspondingly, each (r_{v-1}, r_v)-pair can create more than one object to classify, depending on the number of edits it contains.

4.2 A Classification Scheme for Wikipedia Edits

To understand the nature of the collaborative editing process in Wikipedia articles, we need not only quantitative information about revisions but also qualitative measures. More precisely, we want to analyze

- the effect(s) of a new revision to the content of the article (e.g. reducing the number of spelling errors), and
- the intention and motivation of the author of the revision with respect to a particular change (improving orthography in the article).

Based on these factors, we have identified the following requirements for a Wikipedia edit category taxonomy:

- it should be fine-grained enough to capture the different purposes of edits
- it should account for the particularities of changing the wiki source text (wiki markup syntax and other technical means which influence the editing process, e.g. links and templates)
- it should be possible to infer higher-level information about the effect of a change (i.e. whether it has an effect on the meaning of the article or not)

Additionally, the taxonomy should be grounded in previous work on revision inside and outside Wikipedia, and it should be applicable to label Wikipedia edits by human annotators intuitively.

We have discussed edit category taxonomies in section 3.2.4 (cf. table 3.5). None of these taxonomies fulfills all of the above outlined requirements, as they either lack the level of detail required for an accurate description of edits, or do not properly distinguish between surface and text-base edits (see section 2.2.1). We therefore propose a new edit category taxonomy which meets our requirements and, at the same time, is based on existing revision taxonomies.

4.2.1 Classifying Edits: A Multi-Label Problem

Despite our fine-granular approach to revision analysis, a single edit $e^k_{v-1,v}$ might still encode various changes with different effects. For example, adding a new sentence to the end

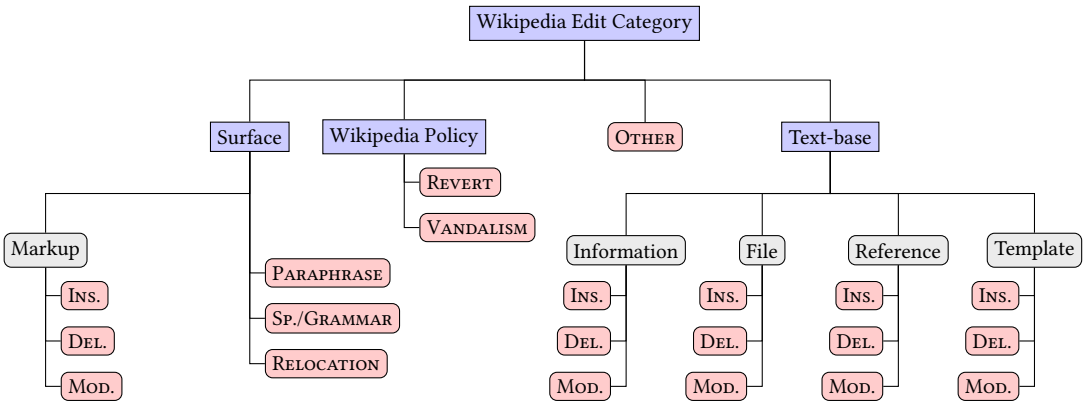


Figure 4.4: The hierarchical Wikipedia Edit Category taxonomy.

of an existing paragraph introduces new information and potentially new (internal or external) references in the form of links. Wikipedians are expected to link new material they introduced with reliable sources.[63] Adding new material in the form of text and new references in the form of a link are different changes (a Wikipedian could also add new material without referencing it or add a reference to existing material). Rather than splitting such changes into separate edits, we decided to allow for multi-labeling edits. It is important to find the right degree of granularity for labeling edits. If the granularity is very coarse (e.g. at the level of the entire text of the revision), many changes might be encoded in the same edit without the possibility to quantify and allocate inherently different changes. This makes the manual annotation of edits harder, and annotators are more likely to overlook edits. The other extreme, very fine granularity (e.g. if every single edited markup character needs to be labeled separately), makes the manual annotation of edits very tedious and the meaning behind a change might be blurred by trivial material. As we will explain later in section 4.5.3, the choice of the level of edit granularity also depends on the application in which the processed edits will serve.

4.2.2 The Proposed Taxonomy

We based our approach to classify edits $e^k_{v-1,v}$ extracted from (r_{v-1}, r_v)-pairs on Faigley and Witte's (1981) generic revision category taxonomy. Wikipedia editing does not follow traditional editorial processes (cf. section 2.1.1 and section 3.2.3). Hence, Faigley and Witte's (1981) taxonomy is particularly suited for the purpose of classifying Wikipedia revisions as it accounts for changes from all stages of the editorial process, from minor copy-edits to substantial revisions on the macro-level.

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. To keep the

[63] `http://en.wikipedia.org/wiki/Wikipedia:Citing_sources`, accessed May 25, 2015

Taxonomy Proposed by	Granularity	Fully Wiki Markup Aware	Text-base vs. Surface	Multi-labeling
Pfeil et al. (2006)	revision	✓	✗	✓
Jones (2008)	revision	✗	✓	✗
Liu and Ram (2011)	edit	✗	✗	✓
Antin et al. (2012)	revision	✓	✗	✓
This study	edit	✓	✓	✓

Table 4.1: Compatibility of various Wikipedia revision category taxonomies with the requirements specified in section 4.2.

taxonomy manageable, we do not follow Faigley and Witte's (1981) fine-grained distinction of textual edits in ADDITIONS, DELETIONS, SUBSTITUTIONS, PERMUTATIONS, DISTRIBUTIONS, and CONSOLIDATIONS. As shown in figure 4.4, our taxonomy is hierarchical with the three top layers Wikipedia Policy, Surface and Text-base. The taxonomy also reflects some of the technical particularities in Wikipedia (most of them apply to any wiki), including reverts and vandalism, templates, usage of media files (images, video etc.), and wiki markup text. Table 4.2 presents a short explanation and example for each category. In table 4.1, we compare our taxonomy with previous work along the dimensions of the requirements outlined above.

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 (see the examples in Table 4.2) 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 bibliographical citations. Different from Liu and Ram (2011), we do not distinguish between links and citations, as these edits have the same effect in the sense of referencing something. Finally, the TEMPLATE category labels all edits related to templates.

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

Category	Description	Edited Text Segment from r_{v-1}	Edited Text Segment from 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 page state	**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	... –> {{...}} **[[Category** ...	...**Dynasty}}** {{...}} ***[[Chinese**...
MARKUP-I	Edits affecting markup segments	an infant Parasaurolophus	an infant "Parasaurolophus"
MARKUP-D		'"AcDec"'	"AcDec"
MARKUP-M		=== The geometry of gravitation ===	== The geometry of gravitation ==
text-base *Edits affecting the meaning of the text*			
INFORMATION-I	Textual edits affecting information content	effects were tested by	effects were **both** tested by
INFORMATION-D		it is **not** a sacrament	it is a sacrament
INFORMATION-M		steppes of Kashmir and **Siberia**.	steppes of Kashmir and **Manchuria**.
FILE-I	Edits affecting files (media content)		**[[File:Dholeskull.jpg**|...**]]**
FILE-D		**[[Image:Victoria_Cross_bar.JPG**|...**]]**	
FILE-M		{{Infobox...|image=UMCLogo.**svg**...}}	{{Infobox...|image=UMCLogo.**xml**...}}
REFERENCE-I	Edits affecting links or bibliographical references and citations		**[[sl:Erlang]]**
REFERENCE-D		**[[**spondee**]]**	spondee
REFERENCE-M		[[**molar**]]	**[[molar (tooth)**|molar]]
TEMPLATE-I	Edits affecting templates (for including text from other pages, automatic text generation etc.)		**{{cite book**|... **}}**
TEMPLATE-D		**{{clear}}**	
TEMPLATE-M		{{**Unit** m|2|1}}	{{**Convert**|2|m|**ft**|1}}
Other	*Segmentation Errors*	[[pirates**]]**	[[pirate**]]s**

Table 4.2: Classification of Wikipedia edits with truncated examples from our corpus. Where necessary, we added the context, while the actual edit is boldfaced in r_{v-1} and/or r_v. To increase readability, line breaks are omitted. Insertion, Deletion and Modification are abbreviated as I, D and M, respectively.

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 text segment which has been changed from r_{v-1} to r_v. Here, we define a text 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, and for File edits, the embedded file must be changed. Likewise, 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 4.2.

We classify changes to the source text of a wiki page, as opposed to the visual changes on the page's 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 towards edits in a certain text type, namelywikis. It is language-independent.

4.3 Classifying Edits in the English Wikipedia

After defining a taxonomy to categorize edits in Wikipedia, we now want to apply it to real-world data and learn how to automatically use annotated data to classify a large number of revisions in Wikipedia. Before we can apply our taxonomy to the data, we need to create a suitable corpus. We used our taxonomy to annotate two corpora of different language versions of Wikipedia. In this section, we present an annotation study on a corpus of edits from the English Wikipedia, while in section 4.4 we annotate and classify edits from the German Wikipedia. We will explain the tool we used to annotate the data in section 4.3.1. The selection of revisions from the English Wikipedia was driven by the goal to create a corpus applicable in quality assessment scenarios. We perform a detailed analysis of results of the annotation study (section 4.3.2), before we describe and evaluate the machine learning model trained and tested on this corpus (section 4.3.3).

4.3.1 Annotation Tool and Visualization of Edits

For the annotation of edits, we used the Apache UIMA Cas Editor.[64] That way, we were able to directly annotate the source files which are produced by a UIMA pipeline. We use this pipeline to extract the raw text for each revision and to segment each (r_{v-1}, r_v)-pair into a list of edits. The editor displays the source text including wiki markup of the older revision r_{v-1}, highlighting all portions of text that have been deleted, changed, or relocated. Additionally, any text inserted in the newer revision r_v is displayed at the location where it has been added. Figure 4.5 shows a screenshot of the Cas Editor displaying a revision which

[64] Unstructured Information Management System, `http://uima.apache.org/`, accessed May 25, 2015 (Ferrucci and Lally, 2004).

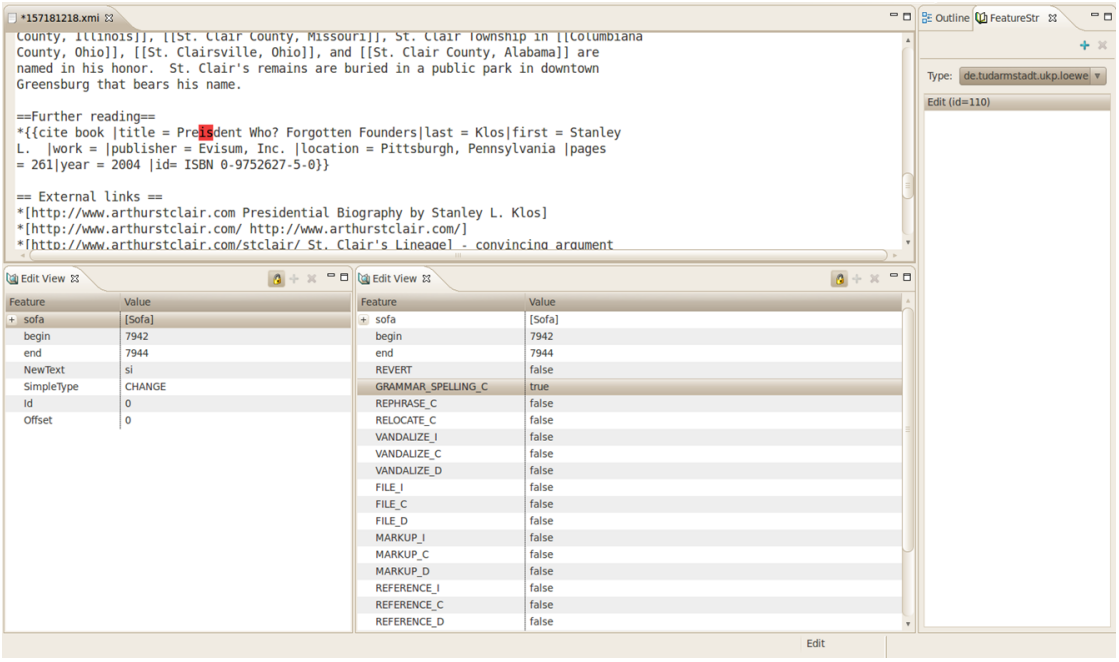


Figure 4.5: The Apache UIMA Cas Editor which we used to annotate edits in Wikipedia revisions. Edits are highlighted within their context. More information about an edit can be found in the lower left window, whereas the category/categories of the edit can be selected in the lower right window. The revision's metadata can be accessed in a separate window.

corrects a spelling error by swapping two characters. The text with which the highlighted portion of text is replaced is displayed in the lower left window. The annotators had access to all metadata information (author name, comment etc.) and the entire text of r_{v-1} and r_v. Further details about the annotation interface and guidelines can be found in appendix C.

4.3.2 The English Wikipedia Quality Assessment Corpus

Focusing on the third research question of this chapter (relationship between article quality and indirect interaction), we compiled a set of English Wikipedia articles from different quality scales. As explained in section 3.2.7, Wikipedia has an internal quality grading scheme, which we apply for this purpose. For each featured article (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 4.3). While 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 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

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

Table 4.3: The size of the latest revision (in characters including wiki markup) and edit frequency (average number of revisions per day) in WPQAC are equal for each FA-NFA pair.

political issues. The youngest article is almost six years old, the oldest one is more than nine years old. We call the result Wikipedia Quality Assessment Corpus (WPQAC).

***Pre* and *Post* Revision Groups** From the article pairs in WPQAC, we selected 891 revisions containing 1995 edits for the annotation study. From the FAs, we determined the revision at the time of promotion to featured status (referred to as r_{prom}) specified on the respective discussion page as the reference and divided the article history into a *pre* and a *post* stage. *Pre* denotes all revisions made 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 4.4. 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 is designed to reflect the entire range of possible edits in Wikipedia, including bot edits, vandalism and reverts. Hence, no further filtering is done.

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

4.3.2.1 Annotation Study

The annotated corpus consists of $N_r = 891$ revisions containing $N_e = 1,995$ edits. We refer to this specific subset of revisions from WPQAC as WPEC (Wikipedia Edit Category corpus). 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.

We hired three undergraduate students with working knowledge of the Wikipedia policies and markup to label the corpus based on written annotation guidelines. To make sure that the annotators had the right understanding of our edit category taxonomy, we carried out several training rounds. 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 4.2 (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 boldfaced 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. The full annotation guidelines can be found in appendix C.1.

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 two annotators, we assign the category l in the gold standard. If all three annotators disagreed, i.e. if an edit was labeled with none of the categories at least two times, it is assigned the Other category in the gold standard. For example, a particular edit changed "...algorithm *will* not terminate..." to "...algorithm *does* not terminate...". The first annotator labeled this edit as Paraphrase, the second as Information-M and the third 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. In the following, we will refer to the final gold standard annotation when we talk about WPEC.

Inter-annotator Agreement To estimate the reliability of the annotations, we computed the inter-annotator agreement per category using the multi-rater Kappa κ measure (Fleiss,

1971), see Table 4.5. 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, 2004) 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 $\alpha = .67$. This is at the lower boundary of what is usually considered to allow for drawing tentative conclusions (Krippendorff, 2004). To the best of our knowledge, no previous 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.

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 4.5, 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 in WPEC 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

$$LD = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i|}{|L|} = .06,$$

where D denotes our data set.

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 $t \in T$, where

Category	κ	P_O	Edits		Revisions		C_e - FA		C_e - NFA	
			C_e	%	C_r	%	*pre*	*post*	*pre*	*post*
Information-I	.64	.91	280	11.67	200	13.11	71	59	81	69
Reference-I	.79	.95	262	10.92	209	13.70	59	37	87	79
Revert	.83	.96	254	10.59	128	8.39	66	55	50	83
Information-M	.58	.90	237	9.88	145	9.50	62	40	72	63
Markup-I	.61	.92	223	9.30	133	8.72	50	54	80	39
Vandalism	.69	.95	163	6.79	98	6.42	50	28	43	42
Spelling/Grammar	.73	.96	161	6.71	80	5.24	32	75	30	24
Information-D	.55	.93	139	5.79	80	5.24	54	32	22	31
Other[a]	.18	.97	139	5.79	86	5.64	42	36	26	35
Markup-D	.58	.95	131	5.46	59	3.87	22	60	23	26
Reference-D	.68	.97	88	3.67	66	4.33	35	6	24	23
Reference-M	.54	.96	88	3.67	78	5.11	24	8	30	26
Template-I	.78	.99	72	3.00	62	4.06	27	20	5	20
Paraphrase	.31	.96	54	2.25	24	1.57	6	12	7	29
Relocation	.71	.99	29	1.21	17	1.11	6	2	17	4
Template-D	.66	.99	26	1.08	20	1.31	13	5	1	7
Markup-M	.25	.97	17	.71	13	.85	8	2	6	1
Template-M	.73	.99	17	.71	9	.59	9	3	0	5
File-I	.78	.997	13	.54	13	.85	5	3	4	1
File-D	.72	.998	5	.21	5	.33	2	1	2	0
File-M	.25	.999	1	.04	1	.07	0	0	0	1
Text-base	.66	.83	1228	51.19	888	58.19	361	214	328	325
Surface	.61	.83	615	25.64	326	21.36	124	205	163	123
Wikipedia Policy	.79	.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 4.5: Inter-annotator agreement in the annotation study on WPEC, 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.

$|T| = 90$. Tsoumakas et al. (2010) refer to this transformation method as *Label Powerset*, as $T \subseteq P(L)$. This transformation helps us to find out more about the relationship between categories, i.e. which categories frequently co-occur. Table 4.6 lists the number of edits labeled with the most frequent unique category set ($Y \geq 2$). Among these, we see that Insertions of Information and/or References together with Markup occur most frequently. This is due to the fact that we seek to capture the effect of an edit in all categories. Hence, an edit which inserts a larger portion of text including links and markup elements needs to be multi-labeled correspondingly.

Category Set	Edits		Revisions	
	N_e	%	N_r	%
Markup-I, Information-I	35	1.75	12	1.35
Markup-I, Reference-I, Information-I	34	1.70	27	3.03
Reference-I, Information-I	32	1.60	12	1.35
Markup-I, Reference-I	13	.65	14	1.57
Markup-D, Information-M	11	.55	3	.34
All	1995	100	891	100

Table 4.6: Absolute numbers of edits N_e and revisions N_r in WPEC for the five largest unique category sets with $Y \geq 2$. N and % are number and percentage of edits resp. revisions in the gold standard.

Error Analysis We created and analyzed confusion matrices over the unique category sets on WPEC (after transformation using the Label Powerset approach, see above) 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 well-defined. Further categories with low agreement are Paraphrase, File-M and Markup-M (cf. table 4.5). 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 4.1.2), 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 (meaning-preserving 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 in some cases. If, for example, an edit *deletes* the word "not" in a phrase like "it is not a sacrament" (cf. table 4.2), this edit also *changes* the meaning, which complicates the annotation of such edits. Across all categories, we observe a higher agreement for Insertions (of Information, References, Markup etc.) as compared to Deletions and Modifications.

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).

4.3.3 Automatic Classification of Wikipedia Edits

In what follows, we describe the set of features we used for the automatic classification of edit categories. Models based on this or part of this feature set are used to classify edits and revisions in the English Wikipedia (this section and section 4.5.2) and edits in the German Wikipedia (4.4.2). For a short introduction to automatic text classification and machine learning, see appendix A.

4.3.3.1 Proposed Feature Set

We grouped our features into *Metadata*, *Textual*, *Markup* and *Language* features. An overview and explanation of all features can be found in table 4.7. The scheme we apply to group edit category classification features is similar to the system used by Adler et al. (2011). We re-use some of the features suggested by Adler et al. (2011), Javanmardi et al. (2011) and Bronner and Monz (2012), as marked in table 4.7. Features are calculated on edited text spans. We label the edited text span corresponding to e_i in r_{v-1} as t_{v-1} and the edited text span in r_v as t_v. In edits which are insertions, we consider t_{v-1} to be empty, while t_v is considered empty for deletions. For Relocations, $t_{v-1} = t_v$. For spell-checking, we use British and US-American English Jazzy dictionaries.[65] Markup elements are detected by the Sweble Wikitext parser (Dohrn and Riehle, 2011).

Metadata Features We consider the comment, author, time stamp or any other flag ("minor change") of r_v as metadata. The Wikimedia user group of an author specifies the edit permissions of this user (see section 3.2.2.2).[66] We indicate whether the revision comments or parts of it have been auto-generated. This happens when a page is blanked, i.e. all of its content has been deleted or replaced or when a new page or redirect is created (denoted by the Comment-is-auto-generated feature). Furthermore, edits within a specific section of an article are automatically marked by adding a prefix with the name of this section to the comment of the revision (denoted by the Auto-generated-comment-ratio feature). Metadata features have the same value for all edits in a (r_{v-1}, r_v)-pair.

Textual Features Textual features are calculated based on a certain property of the changed text. In a preprocessing step, any wiki markup inside t_{v-1} and t_v is deleted. The n-gram feature spaces are composed of n-grams that are present either in t_{v-1} but not t_v, or vice verse. Character n-grams only contain English alphabet characters, token n-grams consist of words excluding special characters.

[65] `http://sourceforge.net/projects/jazzydicts`, accessed May 25, 2015
[66] `http://meta.wikimedia.org/wiki/User_classes`, accessed May 25, 2015

Markup Features As opposed to textual features, wiki markup features account for the Wikimedia specific markup elements. Markup features are calculated based on the number and type of a markup element m and the surrounding context of an edit. Here, m can be a template, an external or internal link, an image or any other element used to describe markup including HTML tags. The type of m is defined by the link target for internal and external links and images, by the name of the template for templates and by the wiki markup element name for other markup elements. Markup features are calculated on text spans t_{v-1} and t_v. Naturally, wiki markup is not deleted beforehand. The edited text spans t_{v-1} and t_v may be located inside a markup element m (e.g. a link or a template). In such cases, our diff algorithm will not label the entire element m, but rather the actually modified text. However, such an edit may change the name of a template or the target of a link. We therefore include the immediate context s_{v-1} and s_v of each edit and compare the type of potential markup elements m in s_{v-1} and s_v. Here, s_v (s_{v-1}) is defined as t_v (t_{v-1}) including all characters from r_{v-1} which precede and follow the edit and which are not separated from t_v (t_{v-1}) by a boundary character (whitespace or line break). If, for example, `[[link1]]` is changed into `[[link2]]`, t_{v-1} corresponds to `1` and t_v to `2`, while s_{v-1} would be `[[link1]]` and s_v `[[link2]]`. The above described features model *what* is actually edited in the text. A number of features are calculated on t_{v-1} only. These features are more likely to inform about *where* an edit is conducted. They specify whether t_{v-1} covers (i.e. contains) a certain wiki markup element and vice versa, i.e. whether t_{v-1} is located inside a text span that belongs to a markup element.

Language Language features are calculated on the context s_{v-1} and s_v of edits, any wiki markup is deleted. For the Explicit Semantic Analysis, we use Wiktionary (Zesch et al., 2008) and not Wikipedia assuming that the former has a better coverage with respect to different lexical classes. POS tagging was carried out using the OpenNLP POS tagger.[67] The vandalism word list contains a hand-crafted set of around 100 vandalism and spam words from various places in the web.

4.3.3.2 Evaluation on WPEC

The pipeline which processes the edits and extracts the features proposed in section 4.3.3.1 is based on a prototype of the DKPro TC framework, cf. appendix B. DKPro TC eases testing several parameter configurations (parameter sweeping) and feature extraction based on the Apache UIMA framework. For the machine learning part, we use Weka (Hall et al., 2009) with the Meka[68] and Mulan (Tsoumakas et al., 2010) extensions for multi-label classification. We randomly split WPEC into 80% training, 10% test and 10% development set, as shown in table 4.8.

[67] Maxent model for English, `http://opennlp.apache.org`, accessed May 25, 2015
[68] `http://meka.sourceforge.net`, accessed May 25, 2015

	Feature	Explanation
Meta Data	Author-group	Wikimedia user group of author
	Author-is-registered*	Author is registered (otherwise: IP user)
	Same-author*	Authors of r_v and r_{v-1} are same
	Comment-length*	Number of characters in comment
	Vulgarism-in-comment	Comment contains a word from vulgarism word list
	Comment-is-auto-generated	Entire comment has been auto-generated
	Auto-generated-comment-ratio	Auto-generated part of comment divided by length of comment
	Incorrect-comment-ratio	Out-of-dictionary word count divided by word count in comment
	Comment-n-grams[a]	Presence or absence of token n-grams in comment
	Is-revert*	Comment contains a word from revert word list
	Is-minor	Revision has been marked as minor change
	Time-difference*	Time difference between r_{v-1} and r_v (in minutes)
	Number-of-edits	Absolute number of edits in the (r_{v-1}, r_v)-pair
Textual	Diff-capitals*	Difference in number of capitals
	Diff-digits*	Difference in number of digits
	Diff-special-characters*	Difference in number of non-alphanumeric characters
	Diff-whitespace-characters	Difference in number of whitespace characters
	Diff-characters*	Difference in number of characters
	Diff-tokens*	Difference in number of tokens
	Diff-repeated-characters	Difference in number of repeated characters
	Diff-repeated-tokens	Difference in number of repeated tokens
	Cosine-similarity	Cosine similarity
	Levenshtein-distance*	Levenshtein distance
	Optimal-string-alignment-distance	Damerau-Levenshtein distance (Damerau, 1964)
	Ratio-diff-to-paragraph-characters	Diff characters divided by length of edited paragraph
	Ratio-diff-to-revision-characters	Diff characters divided by length of r_{v-1}
	Ratio-diff-to-paragraph-tokens	Diff tokens divided by length of edited paragraph
	Ratio-diff-to-revision-tokens	Diff tokens divided by length of r_{v-1}
	Ratio-old-to-new-paragraph	Difference in number of characters in edited paragraph
	Character-n-grams[a]	Presence or absence of n-grams of edited characters
	Token-n-grams[a]	Presence or absence of n-grams of edited tokens
	Simple-edit-type	Modification, Insertion, Deletion or Relocation
Markup	Diff-number*m*	Difference in number of *m*
	Diff-type*m*	Different types of *m*
	Diff-type-context*m*	Different types of *m* within immediate context of edit
	Is-covered-by*m*	Edit is covered by *m* in r_{v-1}
	Covers*m*	Edit covers *m* in r_{v-1}
Language	Diff-spelling-errors*	Difference in number of out-of-dictionary words
	Diff-vulgar-words*	Difference in number of tokens from vandalism word list
	Semantic-similarity	Explicit Semantic Analysis with vector indexes from Wiktionary (Gabrilovich and Markovitch, 2007)
	Diff-POS-tags*	POS tags in symmetric difference
	Diff-type-POS-tags*	Number of distinct POS tags in r_v and r_{v-1}

[a] n-gram features are represented as boolean features.

Table 4.7: List of edit category classification features with explanations. *m* may refer to internal link, external link, image, template, or other markup element. Features marked with * have previously been mentioned in Adler et al. (2011), Javanmardi et al. (2011) or Bronner and Monz (2012).

	N_r	N_e	**Cardinality**
Train	713	1597	1.20
Test	89	229	1.24
Dev	89	169	1.21
All	891	1995	1.20

Table 4.8: Statistics of the training, test and development sets of WPEC. Cardinality is the average number of edit categories assigned to an edit.

		Random	Majority	BR	HOMER	RAKEL
	Threshold[a]	–	–	.10	.25	.33
Example	**Accuracy**	.09	.13	.50	.44	**.53**
	Exact Match	.06	.13	.35	.36	**.44**
	F1	.09	.13	.55	.47	**.56**
	Precision	.10	.13	.54	.46	**.56**
	Recall	.10	.13	**.61**	.50	.60
Label	**Macro-F1**	.10	.06	.49	.35	**.51**
	Micro-F1	.10	.12	.59	.49	**.62**
Ranking	**One Error**	.90	.87	.42	.48	**.34**

[a] Used for creating bipartitions when the classfier outputs a ranking.

Table 4.9: Overall classification results on WPEC with 3 multi-label classifiers and a C4.5 decision tree base classifier, as compared to random and majority category baselines.

Multi-label Classification We report the performance of various machine learning algorithms. Multi-label classification problems are solved by either transforming the multi-label classification task into one or more single-label classification tasks (problem transformation method) or by adapting single-label classification algorithms (algorithm adaption method). Several algorithms have been developed on top of the former methods and use ensembles of such classifiers (ensemble methods). We applied the Binary Relevance approach (BR), a simple transformation method which converts the multi-label problem into $|C|$ binary single-label problems, where $|C|$ is the number of categories. Hence, this method trains a classifier for each category in the corpus (one-against-all). It is the most straightforward approach when dealing with multi-labeled data. However, it does not consider possible relationships or dependencies between categories. Therefore, we tested two more sophisticated methods. Hierarchy of multi-label classifiers HOMER (Tsoumakas et al., 2008) is a problem transformation method. It accounts for possibly hierarchical relationships among

categories by dividing the overall category set into a tree-like structure with nodes of small category sets of size k and leaves of single categories. Subsequently, a multi-label classifier is applied to each node in the tree. Random k-labelsets RAkEL (Tsoumakas et al., 2011) is an ensemble method, which randomly chooses l typically small subsets with k categories from the overall set of categories. Subsequently, all k-labelsets which are found in the multi-labeled data set are converted into new categories in a single-labeled data set using the label powerset transformation (Trohidis et al., 2008). HOMER and BR are among the multi-label classifiers, which Madjarov et al. (2012) recommend as benchmark methods. As underlying single-label classification algorithm, we used a C4.5 decision tree classifier (Quinlan, 1993), as decision tree classifiers yield state-of-the-art performance in related work.

Multi-label Evaluation We denote the set of relevant categories for each edit $e_i \in E$ as $y_i \in C$ and the set of predicted categories as $h(e_i)$. Evaluation measures for multi-label classification systems are based on either bipartitions or rankings. Among the former, we report example-based (weighting each edit equally) and label-based (weighting each edit category equally) measures. The accuracy of a multi-label classifier is defined as

$$\frac{1}{|E|}\sum_{i=1}^{|E|}\frac{|h(e_i)\cap y_i|}{|h(e_i)\cup y_i|},$$

which corresponds to the Jaccard similarity of $h(e_i)$ and y_i averaged over all edits. We report subset accuracy (exact match), calculated as

$$\frac{1}{|E|}\sum_{i=1}^{|E|} I, \text{with } I = 1 \text{ if } h(e_i) = y_i \text{ and } I = 0 \text{ otherwise.}$$

Example-based precision is defined as

$$\frac{1}{|E|}\sum_{i=1}^{|E|}\frac{|h(e_i)\cap y_i|}{|h(e_i)|},$$

example-based recall as

$$\frac{1}{|E|}\sum_{i=1}^{|E|}\frac{|h(e_i)\cap y_i|}{|y_i|},$$

and example-based F1 as

$$\frac{1}{|E|}\sum_{i=1}^{|E|}\frac{2\times|h(e_i)\cap y_i|}{|h(e_i)|+|y_i|}.$$

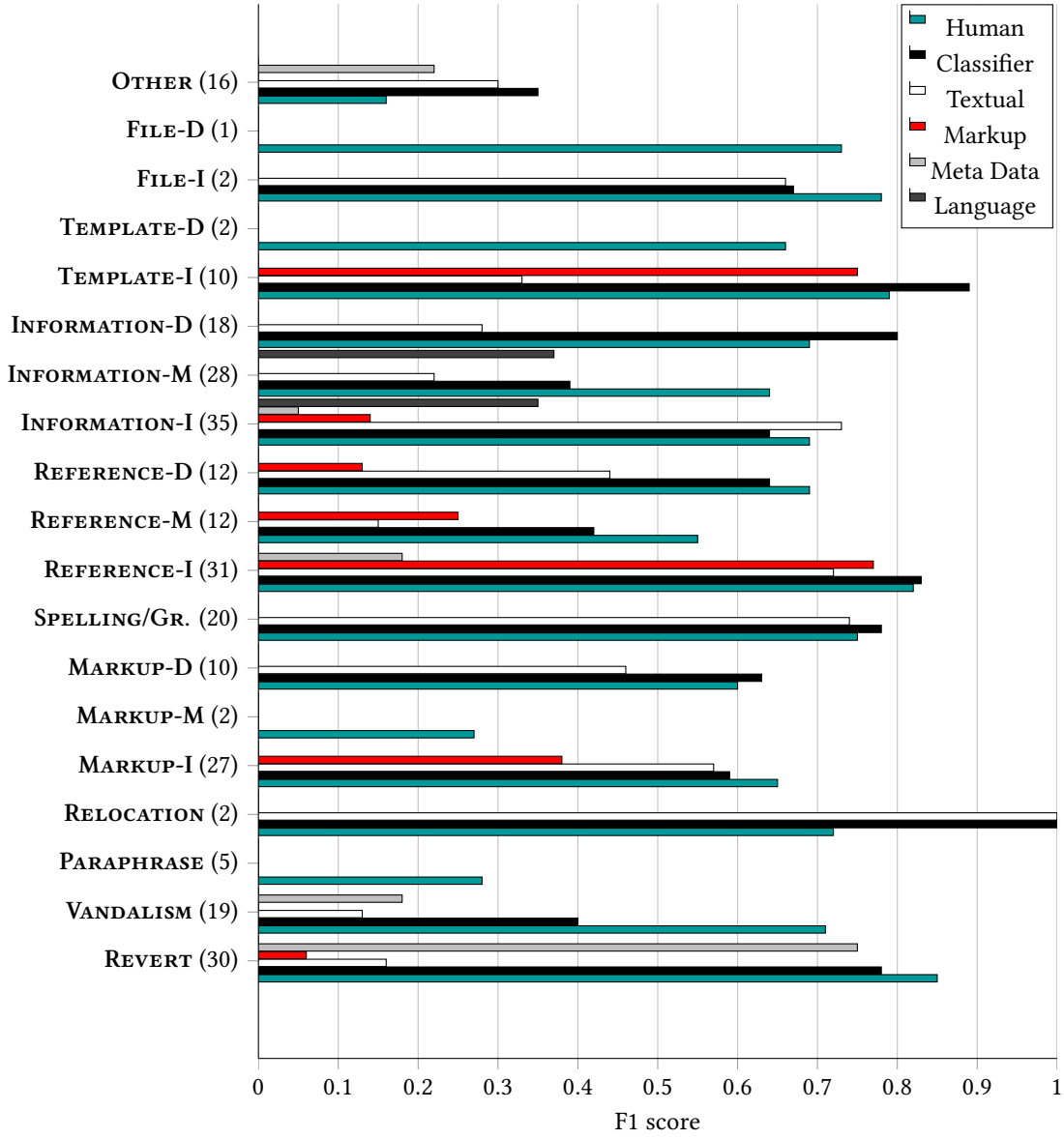


Figure 4.6: F1 scores of RAkEL with C4.5 as base classifier for individual categories on the test set of WPEC. We add human inter-annotator agreement as average pair-wise F1 scores as well as F1 scores for classifiers trained and tested on single feature groups, cf. table 4.7. The number of edits labeled with each category in the test set is given in brackets. The File-M and Template-M categories are omitted in this figure, as they had no examples in the development or test set.

For the label-based measures, we report macro- and micro-averaged F1 scores. As a ranking-based measure, we report one error, which is defined as

$$\frac{1}{|E|}\sum_{i=1}^{|E|}[\![[arg\ max_{c\in C} f(e_i,c)] \notin y_i]\!], \text{with } [\![expr]\!] = 1 \text{ if } expr \text{ is true and } [\![expr]\!] = 0 \text{ otherwise.}$$

$f(e_i, c)$ denotes the rank of category $c \in C$ as predicted by the classifier. The one error measure evaluates the number of edits where the highest ranked category in the predictions is not in the set of relevant categories. It becomes smaller when the performance of the classifier increases.

Table 4.9 shows the overall classification scores. We calculated a random baseline, which multi-labels edits at random considering the label powerset frequencies it has learned from the training set. Furthermore, we calculated a majority category baseline, which labels all edits with the most frequent edit category in the training set. In figure 4.6, we list the results for each category, together with the average pair-wise inter-rater agreement (F1 scores). The F1 scores are calculated based on the annotation study described in section 4.3.2.

Parameters and Feature Selection All parameters have been adjusted on the development set using the RAkEL classifier, aiming to optimize accuracy. With respect to the n-gram features, we tested values for n = 1, 2 and 3. For comment n-grams, unigrams turned out to yield the best overall performance, and bigrams for character and token n-grams. The word and character n-gram spaces are limited to the 500 most frequent items, the comment n-gram space is limited to the 1,500 most frequent items. To transform ranked output into bipartitions, it is necessary to set a threshold. This threshold is reported in table 4.9 and has been optimized for each classifier with respect to label cardinality (average number of labels assigned to edits) on the development set. Since most of the traditional feature selection methods cannot be applied directly to multi-labeled data, we used the label powerset approach to transform the multi-labeled data into single-labeled data and subsequently applied χ^2. Feature reduction to the highest-ranked features clearly improved the classifier performance on the development set. We therefore limited the feature space to the 150 highest-ranked features (see section 4.3.3.3) in our experiments.

For the RAkEL classifier, we set $l = 42$ (twice the size of the category set) and $k = 3$. In HOMER, we used BR as transformation method, random distribution of categories to the children nodes and $k = 3$. For all other classifier parameters, we used the default settings as configured in Meka respectively Mulan.

4.3.3.3 Feature Discussion and Error Analysis

The classifiers significantly outperformed both baselines. RAkEL shows best performance for almost all measures in table 4.9. The simpler BR approach, which assumes no dependencies between categories, still outperforms HOMER.

We trained and tested the classifier with different feature groups (see table 4.7), to analyze the importance of single types of features. As shown in figure 4.6, textual features had the highest impact on classification performance. To the opposite, language features played a minor role in our experiments. Among the highest ranked individual features

for the entire set of categories, we find textual (Levenshtein-distance, Simple-edit-type), markup (Diff-number-markup-elements) and metadata (Number-of-edits) features.

Bronner and Monz (2012) report an accuracy of .88 for their best performing system on the binary classification task of distinguishing fluency and factual edits. The best performing classifier in their study was a Random Forest classifier (Breiman, 2001). To compare our features with their approach, we mapped our 21 edit categories (cf. section 4.2.2) to the binary category set (factual vs. fluency) of Bronner and Monz (2012). Edits labeled as Spelling/Grammar, Markup, Relocation and Paraphrase are considered fluency edits, the remaining categories factual edits. We removed all edits labeled as Other, Revert or Vandalism from WPEC. After applying the category mapping, we deleted all edits which were labeled with both the fluency and factual category. The latter may happen due to multi-labeling. This resulted in 1262 edits labeled as either fluency or factual. On the 80% training split from table 4.8, we trained a Random Forest classifier with the optimized feature set and feature reduction as described in section 4.3.3.2. The number of trees was set to 100, with unlimited depth. On the remaining data (test and development split), we achieved an accuracy of .90. Although we did not use the same data set as Bronner and Monz (2012), this result suggests that our feature set is suited for related tasks such as fluency detection.

With respect to vandalism detection in Wikipedia, state-of-the-art systems have a performance of around .82 to .85 AUC-PR on the English Wikipedia (Adler et al., 2011). We hypothesize that one reason for the low performance of our system for Vandalism edits is the fact that we did not include features which inform about future actions (e.g. whether a revision is reverted). Doing so would make real-time classification of edit categories impossible.

Sparseness is a major problem for some of the 21 categories, as shown in Figure 4.6 by categories such as File-D, Template-D, Markup-M or Paraphrase which have only very few examples in training, development and test set. Categories with low inter-annotator agreement in WPEC such as Markup-M, Paraphrase or Other also yielded low classification accuracy. We analyzed frequent errors of the classifier with the help of a confusion matrix. Paraphrase edits have been confused with Information-M by the classifier. Furthermore, the classifier had problems to distinguish between Vandalism and Revert as well as Information-I. Generally, modifications as compared to insertions or deletions perform worse. All of the tested classifiers build their predictions by thresholding over a ranking, cf. table 4.9. This generates a source of errors, because the classifier is not able to make a prediction, if it does not have enough confidence for any of the categories. This results in so called empty predictions (Liu and Chen, 2015) and affected our results in 13 to 17 percent of the test examples. The imbalance of the data, because of the high skew in the category distribution, is another reason for classification errors. In ambiguous cases, the classifier will be biased towards the category with more examples in the training set.

4.4 Classifying Edits in the German Wikipedia

The model analyzed in section 4.3.3 is language-dependent. As it is trained on a subset of the English WPEC, certain features such as word n-grams will not help to classify edits in another language version of Wikipedia. To understand the influence of this dependency as well as to overcome this drawback, we created another corpus with revisions from the German Wikipedia. This corpus is slightly smaller as compared to WPEC and contains only articles from a particular category. To keep the additional manual annotation effort as small as possible, it is intended to be used in combination with the existing English data.

4.4.1 Annotation Study and Corpus

To make advantage of the existing annotated data, we used WPEC to bootstrap the annotation of the German revisions. In detail, we trained a model on WPEC (English) with *language-independent* features only. With respect to table 4.7, we consider all of the features in the *Language* feature group and all n-gram features all well as the Author-group feature as language-dependent; hence the remaining features are *language-independent*. We then picked a random sample of revisions from German Wikipedia articles.[69] This sample of revisions was automatically classified with the English model trained on the entire WPEC; and the optimized parameters which yielded the best results for English Wikipedia edits (cf. section 4.3). Obviously, this classifier is not tuned for German data. The rationale behind this approach is to minimize the additional manual annotation effort by increasing the number of edits which supposedly take different form in another language, and decreasing edits which have the same shape in another language. For example, the syntax of the wiki markup language is mostly identical across languages. In contrast, the correction of a grammatical mistake may take different forms across languages. For example, fixing a diacritic mark can only be a frequent instance of a spelling error correction in languages which make heavy use of diacritics (e.g. Spanish, but not English) (Mihalcea and Nastase, 2002). With this assumption in mind, we also divided the edit category set in language-dependent categories and language-independent categories. We consider the File, Reference, Template, Markup, Relocation, and Other categories to be language-independent, and the remaining categories as language-dependent. Based on the result from the above classification (English model, prediction on German edits), we picked a random set of 764 edits that were classified with language-dependent categories and another set of 562 edits from all categories (balancing the number of edits in each category). The former set is intended to be used for training in combination with language-independent training data from WPEC. The latter set is intended to be used for testing to ensure that a reliable estimate of classi-

[69] Wikipedia dump from March 2013; all articles are from the category “Wirtschaft”.

	N_e	N_r	**Cardinality**
Train	764	491	1.15
Test	382	322[a]	1.40
Dev	180		1.37
All	1326	813	1.25

[a]The split between test and development set is based on edits rather than revisions to ensure a more balanced category distribution.

Table 4.10: Absolute number of edits and revisions in train, test and development sets of WPEC-GER. Please note that these numbers only refer to the German data, while the results reported in table 4.12 are generated on a model trained with German and English data.

fication performance on German data is available for all categories. We collected a total of $N_r = 813$ revisions containing $N_e = 1326$ edits for the purpose of manual annotation.

For the annotation study itself, two native German speakers with working knowledge of Wikipedia policies and markup labeled all edits. They used the same edit category taxonomy, annotation guidelines and annotation tool as for the annotation of WPEC (cf. section 4.3.1).[70] The annotators were not made aware of the initial split of train and test data in order to maintain objectivity. The overall agreement in terms of Krippendorf's Alpha with MASI distance measure (Krippendorff, 2004; Passonneau, 2006) is $\alpha = .75$; this is almost 10 points more as compared to the WPEC annotation study. This improvement might be explained with a better training of the annotators and clarification updates in the annotation guidelines. In the course of the study, no increase in the number of segmentation errors as compared to the English Wikipedia revisions was observed.[71] We thus assume that the parameters for the segmentation routine discussed in section 4.1.2 are applicable to both English and German revisions. Finally, all cases of disagreement were re-annotated by one of the annotators, arriving at the final gold standard, referred to as WPEC-GER.

For the experiments explained in the following, we used the above outlined division of WPEC-GER into 764 mostly language-dependent edits for training and 562 edits from all edit categories for testing and development. Details can be found in tables 4.10 and 4.11. Label cardinality in WPEC-GER is $LC = 1.2$, label density $LD = .06$ (cf. section 4.3.2), equal to the respective values in WPEC. The difference of cardinality in the test, development and training set is quite high, cf. table 4.10. This is probably due to the controlled distribution of categories in the training and test/development sets, with a high number of language-dependent categories in the training set and vice versa in the test/development set, cf. table

[70]Based on the experiences from the WPEC annotation study, we slightly updated the guidelines for the annotation study on German.

[71]The fact that we tried to control the edit category distribution in WPEC-GER is unlikely to be the only reason for this, as segmentation errors might have happened for any edit category, cf. the low performance of the classifier for the Other category in figure 4.6.

Category	C_e			
	all	train	test	dev
Spelling/Grammar	229	182	32	15
Paraphrase	41	26	11	4
Vandalism	56	47	6	3
Revert	98	82	11	5
Information-I	208	135	52	21
Information-D	127	63	43	21
Information-M	235	160	50	25
Reference-I	134	40	64	30
Reference-D	82	23	39	20
Reference-M	88	26	41	21
Template-I	27	2	17	8
Template-D	21	1	14	6
Template-M	9	4	4	1
File-I	22	4	15	3
File-D	8	2	3	3
File-M	23	1	13	9
Markup-I	102	34	48	20
Markup-D	83	27	38	18
Markup-M	21	11	8	2
Relocation	24	0	16	8
Other	17	5	8	4
Overall count	1655	875	533	247

Table 4.11: Number of edits labeled with a certain category in train, test and development sets of WPEC-GER. Please note that these numbers only refer to the German data, while the results reported in table 4.12 are generated on a model trained with German and English data.

4.11. WPEC-GER does not contain any edits labeled with the Relocation category in the training set. Consequently, a classifier can only learn to classify edits in this category when adding cross-language training data from WPEC.

4.4.2 Cross-Language Learning on English and German Data

To measure the performance of the machine learning model proposed in section 4.3.3 on German revisions, we extracted all edits labeled only with language-independent categories from WPEC and added them to the training set from WPEC-GER. Together with the annotated German training data from WPEC-GER, this sums up to $N_e = 1484$ edits in the cross-language training set. On this data, we trained a model with language-independent features, cf. section 4.4.1. The pipeline used to extract consecutive revisions, segment them into edits, preprocess the edits and extract the features is very similar to the pipeline used

		Random	Majority	BR	HOMER	RAkEL
	Threshold[a]	–	–	.07	.40	.15
Example	**Accuracy**	.05	.08	.46	.45	**.51**
	Exact Match	.03	.08	.35	**.40**	**.40**
	F1	.06	.08	.51	.47	**.55**
	Precision	.07	.08	.51	.50	**.56**
	Recall	.06	.08	.55	.48	**.59**
Label	**Macro-F1**	.04	.01	.39	.34	**.43**
	Micro-F1	.07	.07	.51	.40	**.55**
Ranking	**One Error**	.94	.92	.45	.50	**.40**

[a] Used for creating bipartitions when the classifier outputs a ranking.

Table 4.12: Evaluation on the test set of WPEC-GER.

in section 4.3.3.2, but was fully integrated into the DKPro TC framework. Together with this step, we also remodeled the representation of edits in the pipeline, making use of the *pair classification* mode in DKPro TC. More architectural details can be found in appendix B.3. Furthermore, we made minor modifications to the feature set.[72] This model was evaluated on the test set of WPEC-GER. We used the classifier and feature parameters which yielded the best performance on WPEC (section 4.3.3). All dictionaries and word lists (vulgarism and revert words) were adapted to contain both German and English entries. As the number of features was below 100 (language-independent features only), we did not apply feature selection to further limit the feature set. Equally to the experiments on WPEC, we tested a C4.5 decision tree as base classifier and RAkEL (Tsoumakas et al., 2011), HOMER (Tsoumakas et al., 2008), and BR for multi-label classification. As baselines, we used a random multi-label classifier and most frequent category (cf. section 4.3.3.2).

Overall, the performance of the model trained on the training set of WPEC-GER combined with the language-independent categories from WPEC and tested on the test set of WPEC-GER is worse as compared to the experiments on WPEC (cf. section 4.3.3), especially with respect to the label-based measures. The example-based scores, which weight each edit equally, do not show a substantial decrease in performance.

We assume that the main reason behind the performance drop as measured by the label-based scores is the reduced set of features. Due to the cross-language training set, we could

[72] In addition to the language-dependent features we removed Diff-Type-Context and Is-Covered-By. The latter features become obsolete due to the changes in the architecture of the edit representation in the pipeline.

not use the entire feature set as we did for the experiments on WPEC, but only the language-independent features. Important features such as the ones based on n-grams could thus not be used. Second, as opposed to the WPEC-based experiments and as a result of the cross-language approach, the category distribution in the training and development/test data in WPEC-GER is not balanced. However, a model trained on cross-language training data still outperforms a model trained on the German training set from WPEC-GER only. The best micro-averaged F1 score in a German-only scenario following the above pipeline (all parameters tuned on the development set of WPEC-GER) is .52; three points worse as compared to the best cross-language model (cf. table 4.12). The same is true for a German-only experiment with all features (i.e. language-dependent and language-independent), where the micro-averaged F1 score drops even further to .49. This might seem surprising at first sight. It can however be explained with the fact that the training set of WPEC-GER (see table 4.11) has been designed to be complimented with language-independent training data from the English WPEC and thus contains little to no edits in the language-independent categories. Consequently, WPEC-GER should in practice always be used in combination with language-independent training data from WPEC.

To understand the errors of the classifier on the data level, we inspected the results in more detail. The results on the individual edit categories are shown in table 4.7, in comparison to the human agreement. Unsurprisingly, most of the edit categories which performed weakly in the experiment on WPEC also have low scores on WPEC-GER, among them Markup-M, Paraphrase, Vandalism, File-D, and Other (see figure 4.6). For some of the weak categories, the human agreement was also low (e.g. Paraphrase, Other, File-M, and Markup-M. Other categories (e.g. Vandalism, Template-M, and File-D) are underrepresented in the development and test set and might thus not have been properly accounted for by the model. We cannot identify systematic differences in the performance of the classifier on language-dependent and language-independent categories. Rather, the German model has similar weaknesses as compared to the English model.

In conclusion, we have shown how the English WPEC can be used in combination with the German WPEC-GER to train a better model for classifying German Wikipedia edits. Although the resulting model does not work equally well across all categories, it is a first step towards cross-language classification of Wikipedia edits.

4.5 Classifying Revisions in the English Wikipedia

Having analyzed the performance of our classifier model on Wikipedia edits in a language different from English, we will now turn back to English data, but test the model on a completely different dataset which has been annotated with a different taxonomy and granularity, namely on the level of revision (rather than edits). Like this, we will show that our proposed feature set can well generalize to datasets and tasks with different properties. A

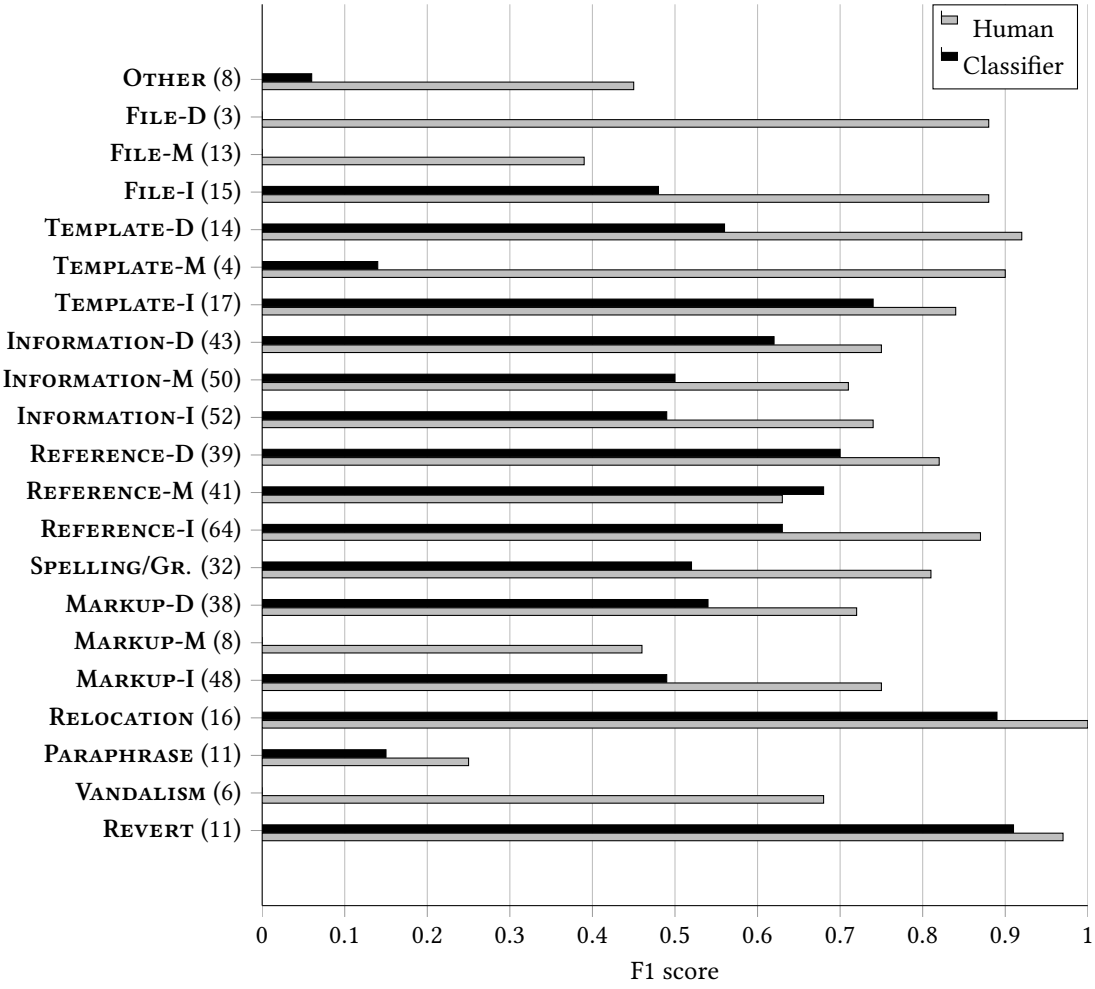


Figure 4.7: F1 scores of RAkEL with C4.5 as base classifier for individual categories on the WPEC-GER test set. We add the inter-annotator agreement between the two annotators on WPEC-GER as F1 scores. The number of edits labeled with each category in the test set is given in brackets.

drawback of WPEC and WPEC-GER are their relatively small sizes, which limit the performance of models trained on them. However, there are larger corpora available, and they can also be used with the model explained in section 4.3.3 as we will show in the following.

4.5.1 Annotating Wikipedia Revisions

For the following experiments, we employed the sample of Wikipedia articles used in Arazy et al. (2011, 2013). This set was created using a stratified sampling approach, based on articles' topical category. Such an approach is necessary given that the articles' topics affect editing patterns (Kittur et al., 2009b). The article sample is organized into six mutually exclusive and collectively exhaustive classes: (a) culture, art, and religion; (b) math, science, and technology; (c) geography and places; (d) people and self; (e) society; and (f) history and

events. The sampling procedure for this set required that the article's length be between 200 and 3,500 tokens, thus excluding undeveloped articles and extremely long outliers. Altogether, the sample contains 93 articles.

The annotation of the revisions was based on a taxonomy developed by Kriplean et al. (2008), which was already employed as a basis for the large-scale manual annotation study in Antin et al. (2012).[73] Kriplean's original taxonomy of "Editing Work" categories included ten categories, which were refined by Antin et al. (2012) after some pilot testing. This taxonomy was further refined through discussion among the authors, resulting in a comprehensive list of twelve meaningful categories that could be understood and identified by the annotators. The unit of analysis for the annotation was at revision level, where multi-labeling was allowed, i.e. each r_{v-1}, r_v-pair could be annotated with one or more revision categories. Each r_{v-1}, r_v-pair was annotated by at least two annotators. Table 4.13 lists all categories with short explanations.

For the sake of this study, we only used a subset of revisions labeled in the annotation study. To create a gold standard based on the outcome of the annotation study, we took a conservative approach and excluded all revisions which did not have full agreement on all categories. Given the high number of each r_{v-1}, r_v-pairs labeled during the annotation study, we aimed to maximize the quality of the training data and therefore wanted to exclude all cases with disagreement. The data used in the experiments and described in the following have been extracted from a Wikipedia dump from January 2012, which contains only revisions up to January 4th, 2012. As a result, we had to exclude further revisions after this cutoff date. The number of r_{v-1}, r_v-pairs in the final gold standard is 13,592. We refer to this data as WPRC.

Table 4.14 lists the distribution of categories in WPRC. Similar to WPEC, the addition of information is one of the most frequent changes. The Add or Change Wiki Markup category in WPRC includes various categories which are separated in our edit category taxonomy, e.g. Reference-I which is the second most frequent category in WPEC, and Markup-I, which is also very frequent in WPEC.

The label cardinality (labels per revision on average) in WPRC is 1.5. On the overall annotated data, we calculated the inter-rater agreement. Krippendorf's Alpha (Krippendorff, 2004), with MASI function (Passonneau, 2006), is $\alpha = .71$, reflecting sufficient agreement (Krippendorff, 2004).

4.5.2 Automatic Classification of Revisions

The model we use to train and test our automatic revision classifier is very similar to the model described in section 4.3.3. We also use the segmentation algorithm explained in

[73] The annotation study was not carried out as part of this work and is therefore not covered in detail here. The permission to use the annotated data was granted by the owners of the data, represented by Ofer Arazy.

Category	Description
Add Substantive New Content	New information is added, changing the meaning of the article
Delete Substantive Content	Existing information is removed, changing the meaning of the article
Move or Create New Article	An article is created or moved
Fix Typos and Grammatical Errors	Grammatical, spelling and/or minor formatting errors are corrected
Rephrase Existing Text	Sentences are re-structured for clarity, not changing the meaning of the article
Reorganize Existing Text	One or more bodies of text are moved from one location to another; headings or categories are added or deleted, changing the overall structure of the article
Insert Vandalism	Malicious content is added, text is deleted without any obvious reason
Delete Vandalism (revert)	Damage done by a vandal is reverted
Add or Change Wiki Markup	A body of text containing wiki markup characters is added, deleted or changed
References (to external sources)	References to external sources are added, deleted or changed
Hyperlinks (to other Wikipedia pages)	The target of a link is changed; a new link is added; an existing link is deleted
Miscellaneous	A change which does not fall under any of the other categories is performed

Table 4.13: The 12 categories we used to annotate revisions in Wikipedia.

Category Label	Revisions	Percent.
Add or Change Wiki Markup	5600	27.5
Add Substantive New Content	4174	20.5
Fix Typos and Grammatical Errors	2354	11.6
Insert Vandalism	1750	8.6
Delete Substantive Content	1452	7.1
Delete Vandalism	1430	7.0
Reorganize Existing Text	1042	5.1
Rephrase Existing Text	800	3.9
Hyperlinks (to other Wikipedia pages)	667	3.3
References (to external sources)	680	3.3
Miscellaneous	312	1.5
Move or Create New Article	74	0.4

Table 4.14: Number and percentage of revisions in WPRC labeled with a certain edit category.

section 4.1.2 to split r_{v-1}, r_v-pairs into edits. Although we only have information about categories on revision level in WPRC, the segmentation into edits has the advantage that we can reuse a lot of the existing features from section 4.3.3, which extract edit information on a fine-granular level. To map the extracted information from edit to revision level, we recombine the edited text from all n edits $e^k_{v-1,v}$ in each r_{v-1}, r_v-pair. The edited text span corresponding to e^k_{v-1} is labeled as t^k_{v-1} and the edited text span corresponding to e^k_v as t^k_v. In edits which are insertions, t^k_{v-1} is considered to be empty, while t^k_v is considered empty for deletions. For Relocations, $t^k_{v-1} = t^k_v$.

Recombining all edits $e^k_{v-1,v}$ from a r_{v-1}, r_v-pair is done by joining all t^k_{v-1} from all e^k_{v-1} and likewise for all t^k_v. The recombined text spans are labeled with t_{v-1} and t_v, respectively.

In section 4.3.3, we divide the feature set into features based on metadata, textual features, wiki markup features and language features. For the experiments in the following, we used the metadata features as-is, since features in this group act on revision level only.[74] The metadata features account for information extracted from the revision comment, author name, time stamp or other flags. They do not consider the actual textual change within the main text of the revision.

The features in the other groups are calculated based on the textual change, i.e. differences between t_{v-1} and t_v. We use all features from the markup and language group, cf. table 4.7.[75] From the textual features, we used all features but Diff-repeated-tokens, Ratio-diff-to-revision-tokens, Character-n-grams, and Token-n-grams.[76] The Simple-edit-type feature was modified to extract the type of edit which occurred most frequently in the r_{v-1}, r_v-pairs.[77] We added two new metadata features: a simple feature indicating whether the author left a comment at all, and another feature specifying whether the name of the author indicates that this user is a bot (whether the name contains a suffix or prefix corresponding to "bot").

All parameters of the features were set as for the experiments described in section 4.3.3. The feature space was limited to the 100 highest-ranked features, using the label powerset approach to transform the multi-labeled data into single-labeled data and subsequently applying the information gain algorithm. As for the machine learning algorithm, we again used RAkEL. In addition to the C4.5 decision tree classifier (Quinlan, 1993), we tested Random Forest (Breiman, 2001) as base classifier on WPRC. Therefore, we randomly split the 13,592 revisions into 80% training and 20% test set. Classifier hyperparameters were optimized according to the Macro-F1 score on the test data.

Table 4.15a shows the results of the best classifier on the test data. Overall, the micro-averaged F1 score score of .78 is very satisfying. The classifier performs close to human

[74] Except for the Number-of-edits feature.
[75] Except for the Diff-type-POS-tags feature.
[76] The exclusion of the latter two slightly speeds up the classification process.
[77] We return a separate value to mark cases where multiple edit categories have the same (highest) frequency.

	BL	RF	C4.5
Exact Match	.16	**.66**	.60
Example F1	.36	**.75**	.73
Macro-F1	.09	**.68**	**.68**
Micro-F1	.41	**.78**	.77
One Error	.70	**.22**	**.22**
Macro-F1 Human		.73	

(a) Performance across all categories, compared to the majority baseline BL and human agreement (macro F1).

	RF	C4.5	Human
Add New Content	**.81**	.79	.80
Delete Content	**.57**	.56	.68
New Article	.81	**.85**	.90
Fix Typo	**.73**	.72	.69
Rephrase Text	**.38**	.35	.61
Reorganize Text	**.70**	.69	.79
Insert Vandalism	**.64**	.59	.81
Delete Vandalism	**.78**	**.78**	.85
Add Wiki Markup	**.92**	**.92**	.80
References	.78	**.81**	.66
Hyperlinks	.56	**.59**	.64
Miscellaneous	.49	**.51**	.52

(b) Performance per category, compared to human agreement (using the F1 metric).

Table 4.15: Performance of classifiers (RAkEL with Random Forest and C4.5 base classifiers), on the test set of WPRC.

agreement, as shown by the macro-averaged F1 score of .68 (as compared to human agreement of .73). It clearly outperforms the majority baseline, which labels all revisions with the most frequent category set in the training data (Add or Change Wiki Markup and Add Substantive New Content). For two thirds of all revisions in the test data, our model found the exact set of categories. As indicated by the One Error measure, for about every fifth revision, the category predicted with the highest confidence was not in the set of true categories. Among the individual categories, we find the Rephrase Text and Miscellaneous categories with a particularly low F1 score below .5. This is likely to be due to the agreement of the human annotators on these categories which is the lowest among all categories, cf. table 4.15b. Rephrase Text revisions were frequently classified as Fix Typo, whereas Miscellaneous was confused with many different categories. Furthermore, the classifier has difficulties with the categories Hyperlinks and Delete Substantive Content. Hyperlinks typically co-occurred with Add Wiki Markup in the training data, so that the classified also assigned those two categories to test instances which were only annotated with Hyperlinks by the human annotators. The lower performance of the latter categories is also to be explained with the comparably low human agreement on these categories. The classifier also confused Insert Vandalism with Add New Content, a problem we already noted for WPEC (cf. section 4.3.3.3).

4.5.3 Insights from Classifying Revisions

The performance of the above classifier is better than the performance of the classifier models trained on WPEC and WPEC-GER. There are two main reasons for that: (a) the higher amount of training data in WPRC (5 to 10 times more as compared to WPEC resp. WPEC-GER), and (b) the different revision category scheme, which contains a lower number of categories than our proposed edit category taxonomy (12 vs. 21, cf. section 4.2.2), and which labels revisions and not edits. The effect of the lower number of categories can also be observed by the increased performance of the majority category baseline on WPRC as compared to WPEC and WPEC-GER. The improvement in performance comes at the price of the degree of information about edits. Since the classifier explained in section 4.5.2 is only able to classify revisions and not edits, it cannot give detailed information about: (a) the number of individual edits (e.g. it is not clear whether a revision labeled as Fix Typos and Grammatical Errors contains one correction of a typo or ten), and (b) the impact of individual edits (e.g. in a revision classified as Add Substantive New Content and Delete Substantive Content it is not clear if one edit had a higher impact in terms of its size and if so which one). Second, the smaller set of categories used to label WPRC (section 4.5.1) obviously limits the explanatory power of the classification. For example, the changes affecting templates are not explicitly covered in the WPRC category set.

In Figure 4.8, we outline a mapping between the taxonomies used to annotate edits and revisions in WPEC/WPEC-GER and WPRC, respectively. The mapping is based on the overlap between the explanations in the annotation guidelines. We have not carried out an empirical investigation on the actual data to confirm whether this correspondences hold. As explained above, for all of these mappings the intended level of edit granularity needs to be considered. For example, while in WPEC the Paraphrase category is used to describe a single edit (typically a small portion of text, e.g. a word or a sentence), in WPRC the Rephrase Existing Text category is applied to the entire revision, possibly in combination with other categories. Only four categories have a clear overlap in their definitions. The remaining mappings are either many-to-one (several categories in WPEC are mapped to one category in WPRC), or only partially correspond to each other. It is also noteworthy that while moving or creating an article cannot be expressed explicitly in WPEC, there is no direct counterpart for WPEC's Information-M category in WPRC. Those missing categories need to be expressed using other edit categories from the respective taxonomy, and it is not clear which category will be used in which context.

The choice of the edit classification model depends on the task at hand. If a fine-granular classification of edits at the textual level is important (e.g. for analyzing a certain kind of edit such as spelling correction or paraphrases), WPEC and WPEC-GER are to be preferred as sources for training data. For large-scale studies analyzing all kinds of edits across many revisions and articles (e.g. tracking the edits of users across Wikipedia), WPRC might be the

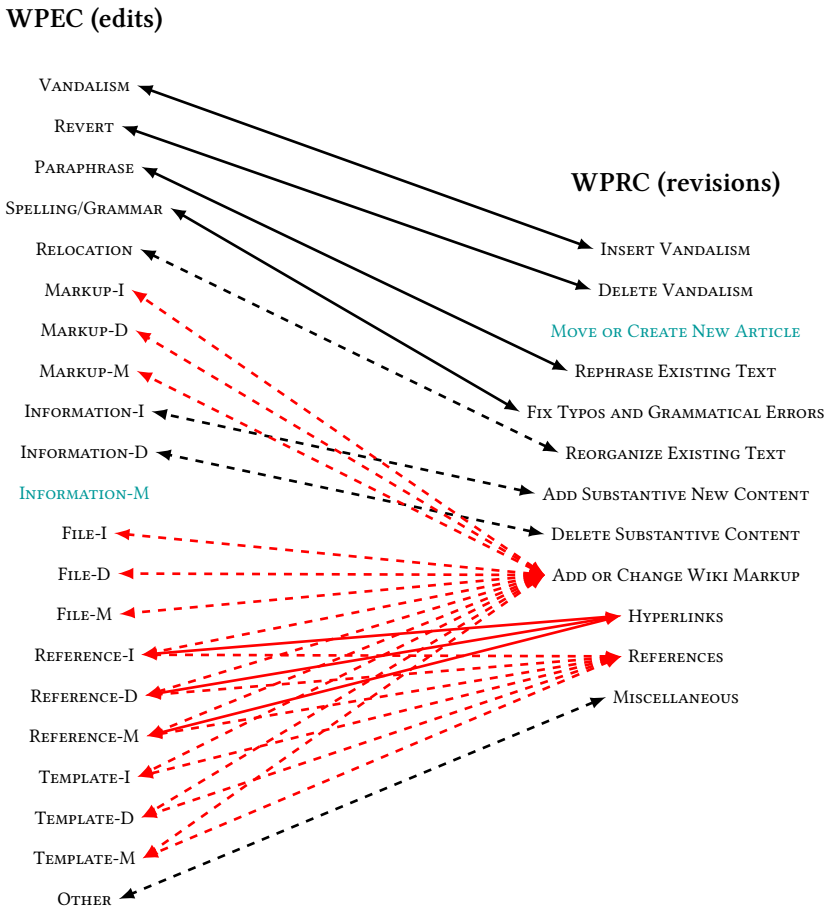


Figure 4.8: A mapping between edit categories in WPEC (annotated on edit level) and WPRC (annotated on revision level). Dashed lines indicate partial mapping, whereas solid lines represent substantial overlap between the categories. Black lines connect one-to-one relation, red lines many-to-one relations. Categories without an appropriate mapping are additionally highlighted in color.

better suited, since higher-level information is likely to reveal a higher number of recurring patterns. We apply WPRC in chapter 5 to detect roles of users based on edit behavior. In the next section, we turn back to WPEC and show how to use it to analyze differences between featured article and non-featured article.

4.6 Wikipedia Revisions and Aspects of Article Quality

In this section, we will study the relationship between the revision history of a Wikipedia article and the (information) quality of this article. To this end, we rely on the Wikipedia internal measures for article quality presented in section 3.2.7, namely FA and NFA. The corpora introduced in section 4.3.2, WPQAC and WPEC, are targeted towards this analysis.

Group	r (all)	r (Top-16)	r (Jones, 2008)	Correlation criteria
All	.87*	.80*	.91*	FA/NFA
All	.90*	.84*	—	pre/post
FA	.72*	.57	.68	pre/post
NFA	.87*	.81*	—	pre/post
pre	.86*	.80*	—	FA/NFA
post	.68*	.52	—	FA/NFA

Table 4.16: Pearson correlation r between frequency distributions of edit categories by revision group for all and for the 16 largest categories in WPEC. For comparison, we added the corresponding numbers for Jones's (2008) study. Values marked with * are statistically significant for $p < 0.01$.

In section 4.6.1, we compare the distribution of edit categories in different stages of FAs and NFAs. Subsequently, in section 4.6.2, we mine patterns of edit category sequences and compare them across FAs and NFAs.

4.6.1 Edit Category Distribution in Featured and Non-Featured Articles

We designed WPEC 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 4.4). Table 4.16 shows the Pearson correlations over category distributions between relevant groups. These calculations are based on the category frequencies of multi-labeled edits (table 4.5, 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 4.16, the correlations between *pre-FA* and *post-FA* as well as *post-FA* and *post-NFA* are not statistically significant when calculated for the top 16 categories. We can thus assume that the revisions in the *post-FA* group show distinctive properties with respect to the way in which they are being edited.

For the Spelling/Grammar and Reference categories, the deviation between the absolute number of edits in FAs and NFAs is particularly high (see table 4.5). 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

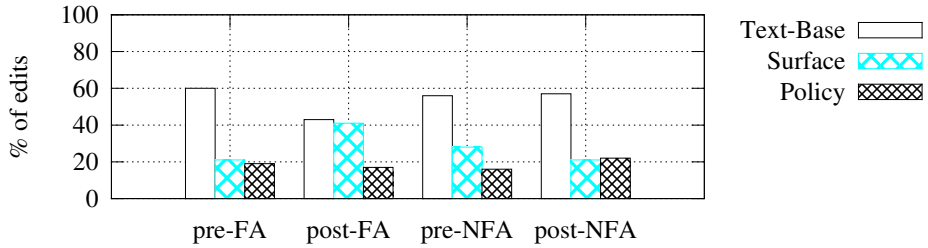


Figure 4.9: Percentage of edits, based on C_e, for layers in revision groups in WPEC.

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 (cf. section 3.2.7) 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.

It is not possible to verify the distinction between experienced and inexperienced authors as explained by Sommers (1980) for the CW process in Wikipedia. As can be seen in table 4.5, 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 4.16. 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 comparing *post-FA* revisions to *pre-FA*, *pre-NFA*, and *post-FA* revisions, cf. table 4.5. Hence, even if we cannot find significant differences in the editing history of FAs and NFAs, there is a deviation in the CW 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 CW 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 4.9 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.

To find out whether these findings hold over a larger sample of Wikipedia edits, we used the best-performing model described in section 4.3.3 to automatically classify the en-

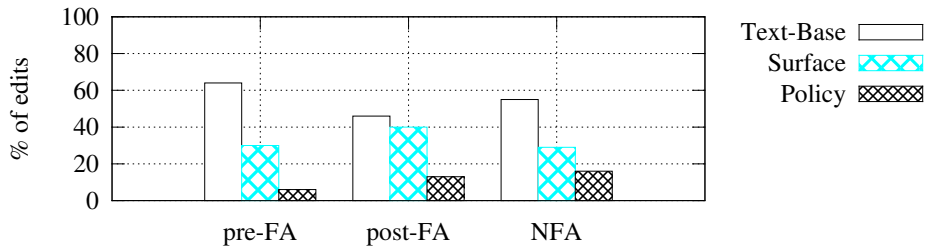


Figure 4.10: Edit category distribution (percentages) as classified by our model over revision groups in WPQAC.

tire revision history of the articles contained in WPQAC (see section 4.3.2). The set of articles covered in WPQAC are 10 FA and 10 NFA, containing an overall number of 21,578 revisions (9,986 revisions from FA and 11,592 from NFA), extracted from the April 2011 English Wikipedia dump. During the classification process, we discarded revisions where the classifier could not assign any of the 21 edit categories with a confidence higher than the threshold, cf. table 4.9. This resulted in 17,640 remaining revisions.

Within WPQAC, we divided the revision history of FAs into pre- and post-revision groups and calculated the edit category distribution for each group, following the process described in section 4.3.2. Given that the pre- and post-revision groups for NFAs did not show significant differences in WPEC (cf. table 4.16), we calculate the distribution of edit categories for NFA on the entire revision history. The result is summarized in table 4.10. The distribution of edit categories in WPQAC backs our findings on the smaller WPEC in that revisions of FAs after they were featured have a higher number of surface edits and a lower number of text-base edits as compared to before the articles were featured. We conclude that there are indeed strong indications that FAs become more stable in terms of the categories of edits they receive after they were featured. Nevertheless, we have to say, that – although WPQAC covers different types of articles in the English Wikipedia– these results are drawn from a relatively small sample of 20 articles so that inferences regarding the entire Wikipedia need to be considered carefully.

4.6.2 Mining Collaboration Patterns in Featured and Non-Featured Articles

In addition to the mere distribution of edits, we analyzed sequences of edits across time. As a recent study by Wang et al. (2014) suggests for the case of ontology development, the order of edits in a sequence is typically not random, but follows more or less predictable patterns. We make use of this assumption, as we analyze connections between edit sequences and article quality in Wikipedia. For this type of analysis, we expected meaningful results only from longer edit sequences and therefore carried out our analysis on the automatically

classified revisions from WPQAC. On the 17,640 revisions, divided into sequences for each of the 20 articles contained in WPQAC, we applied a sequential pattern mining algorithm with time constraints (Hirate and Yamana, 2006; Philippe Fournier-Viger et al., 2008). The latter is based on the PrefixSpan algorithm (Pei et al., 2004). The calculations have been carried out within the open-source SPMF Java data mining platform.[78]

We created one time-extended sequence database for the ten FAs and one for the ten NFAs. The sequence databases consist of one row per article. Each row is a chronologically ordered list of revisions. Each revision is represented by the itemset of all edit categories for all edits in that revision (in alphabetical order).

The output of the algorithm are sequential patterns with time constraints. To obtain meaningful results, we constrained the output with the following parameters:

- Minimum support: 1 (the patterns have to be present in each article)
- Time interval allowed between two successive itemsets in the patterns: 1 (patterns are extracted only from adjacent revisions)
- Minimum time interval between the first itemset and the last itemset in the patterns: 1 (the length of the patterns is 2 or higher)

As this output reflects recurring sequences of adjacent revisions labeled with edit categories, we refer to it as *collaboration patterns*. With these parameters, the algorithm discovered 1,358 sequential patterns for FAs and 968 for NFAs. The number of shared patterns in FAs and NFAs is 427, this corresponds to the number of frequent patterns in a sequence database which contains all 20 FAs and NFAs. The maximum length of patterns which were found was 6 for FAs, and 5 for NFAs. These numbers show that the defined collaboration patterns seem to have discriminative power for different kinds of articles. FAs can be characterized by a higher degree of homogeneity with respect to their collaborative patterns due to a higher number and length of frequent sequential patterns in FAs as compared to NFAs.

In table 4.17, we list some examples of collaboration patterns with a minimum support of 1 which we found in featured, but not NFAs, or vice verse. Unsurprisingly, patterns which contain combinations of the most frequent categories (Information-I, Reference-I), have a high overall frequency. The diversity inside collaboration patterns measured by the number of different edit categories was higher in NFAs. For example, the Vandalism - Revert pattern was only found in NFAs. Patterns in FAs tended to be more homogeneous, as shown by the first pattern in table 4.17, a repetition of additions of information. We conclude that distinguished, high-quality articles, show a higher degree of homogeneity as compared to a subset of NFAs and the overall corpus.

[78] `http://www.philippe-fournier-viger.com/spmf`, accessed May 25, 2015

Featured articles

❶	❷	❸	❹	❺
Information-I	Information-I	Information-I	Information-I	Information-I
Information-D, Information-I	Information-I	Information-I	Reference-I	
Template-D	Reference-I			

Non-featured articles

❶	❷	❸	❹	❺
Information-I	Information-I, Reference-I	Information-I	Reference-I	Markup-I
Markup-I	Reference-D	Markup-I		
Vandalism	Revert			

Table 4.17: Examples of collaboration patterns with different pattern length which have been found in either all FAs or all NFAs in WPQAC.

4.7 Implications beyond Wikipedia

In this chapter, we have presented a detailed investigation of indirect user interaction in online mass collaboration. The data used for our analysis was extracted from the English and the German Wikipedia, however, we believe that our findings also yield implications for online mass collaboration settings outside Wikipedia.

The data generated by indirect user interaction in Wikipedia is used to support a very broad range of applications (Nelken and Yamangil, 2008; Yamangil and Nelken, 2008; Max and Wisniewski, 2010; Zesch, 2012; Recasens et al., 2013). The ability to automatically classify edits in Wikipedia helps to produce training data for applications outside Wikipedia. As we have shown in this chapter, the bandwidth of edit categories in Wikipedia is broad and ranges from simple spelling correction to vandalism. To make this data useful for external applications, it is necessary to preprocess and filter the edits according to the intended usage. We have produced and analyzed two corpora (WPEC and WPEC-GER) which can be used for to train models able to classify Wikipedia edits in English and German.

Second, the experiments carried out in section 4.3 through 4.5, although certainly biased towards editing behavior in Wikipedia, give general insights about the particularities of the CW process in online mass collaboration projects. The impact of vandalism and countermeasures (e.g. reverts) in open online collaboration is impressive. As shown in our data, in Wikipedia, approximately every tenth revision is a malicious edit. Surface edits (e.g. for-

matting and fixing spelling errors) also require a high effort. Based on our data, we estimate that about every forth revision is concerned with rather cosmetic changes.

Third, we have analyzed potential correlations between indirect user interaction through editing and article quality. An analysis of collaboration patterns in Wikipedia suggests that the editing behavior and the chronological sequence of edits during CW has an impact on information quality. As we have shown, collaboration patterns in high-quality articles are more homogeneous than the average editing pattern. Our findings further indicate that different phases in the development of a document under mass collaboration require different categories of edits. Our analysis of FAs and NFAs in Wikipedia showed that high-quality articles tend to receive a higher proportion of edits to the text-base (mostly dealing with the information in the article) in their earlier stages and later on, once the articles have matured, they tend to receive more surface edits (concerned with formatting etc.). These findings roughly resemble the activities involved in CW (discussed in section 2.3.1), and indicate that the organization of the overall writing strategy (e.g. drafting the information first and later on improving the style) is likely to have a positive impact on the quality of the collaboratively written document.

4.8 Conclusion

In this chapter, we presented a detailed study of indirect user interaction in Wikipedia. To this end, we have developed and discussed a taxonomy to categorize edits in Wikipedia. We tested the edit category system on two corpora in different languages. Furthermore, we have developed a machine learning model to automatically classify Wikipedia edits and revisions, which we tested on three corpora and two different edit/revision category systems.

In detail, we answered the following questions. To understand the content of and intention behind edits in Wikipedia (first research question), we have developed 21-category taxonomy for Wikipedia edit categories. Our taxonomy distinguishes surface edits which do not change the meaning of the edited document and text-base edits which have an impact on the meaning. In the second research question, we asked how to automatically categorize edits in Wikipedia. To answer this question, we first manually annotated two corpora of English and German Wikipedia edits. Second, we present a machine learning system which we trained on both annotated corpora and which is able to automatically classify English edits with a micro-averaged F1 score of .62 and German edits with .55 F1 score. Furthermore, we tested our system on a third corpus of Wikipedia revisions, and show that (a) it can also be used to classify revisions as opposed to edits, (b) it also works with a different category taxonomy, and (c) it is able to classify revisions with a macro-averaged F1 score of .78, when trained on a larger set of training data. We also wanted to know whether there is a relationship between article quality and indirect interaction in Wikipedia (third research

question). Addressing this question, we have shown that the information content in FAs tends to become more stable after their promotion and that FAs show a higher degree of homogeneity with respect to their collaboration patterns as compared to random articles. We conclude that documents require different writing strategies in different stages of the writing process in CW projects. This finding could be helpful to point authors with certain editing interests or behaviors to the right places. For example, an author who appears to be an expert in spelling correction or applying layout conventions could be pointed to a document which has just received a high number of text-base edits and thus might need refurbishing in terms of surface edits. We will discuss the matter of editing behavior in much more detail in chapter 5.

It should be noted that, in the specific context of Wikipedia, the internal quality rating system (e.g. FAs vs. NFAs) does not ensure an impartial decision making process. The same authors who heavily edit an article might be the ones who up-vote or down-vote this article. While it is to be assumed that authors who actively participate in the CW process should have an idea about the quality of their product, this factor may also introduce potential bias in a competitive process as an article might be supported by one or more of its authors who appear to be nominators or reviewers in the promotion process. This limitation should be considered when drawing conclusions about the information quality of articles (Ferschke, 2014), but it does not harm our findings about different stages in the CW process of Wikipedia articles.

In this chapter, we have laid the technical and experimental foundations for the approaches presented in the subsequent chapters. In chapter 5, we further elaborate on the matter of CW organization by analyzing the roles of the writers in online mass collaboration. In chapter 6, we establish ties from indirect user interaction to direct user interaction.

Chapter 5

Activity-Based Roles in Wikipedia

We have analyzed edit categories and how these can be used to understand the CW process in Wikipedia. In this chapter, we will abstract away from the textual level of edits and turn our focus to authors and their edit behavior. As discussed in section 3.1, coordination is a crucial matter for the success of online CW projects. In Wikipedia and other online CW communities, an important attempt to address this issue are roles which are assigned to users, based on different criteria (cf. section 2.3.2). As explained in section 3.2.1.1, there are two main dimensions which are used to determine the role of a user, namely the formal dimension and the activity-based dimension. The formal dimension is based on administrative functions of authors, involving a defined set of responsibilities and rights, whereas activity-based roles are assigned based on the (edit) behavior or preferences of authors. A number of studies from several research areas have analyzed the formal dimension as an important factor for coordination in online mass collaboration (Arazy et al., 2015, 2014; Stvilia et al., 2008; Butler and Sproull, 2007). In this chapter, we will turn to the less studied dimension of activity-based roles, based on the findings about indirect user collaboration in Wikipedia from chapter 4.

Not much is known about activity-based roles in online mass collaboration, about their impact and their nature. Particularly, the stability of activity-based roles is unclear, i.e. whether they can reliably be detected in a large online mass collaboration project such as Wikipedia and whether authors keep or change activity-based roles during the entire CW process. We therefore seek to answer the following questions:

1. Does the CW process in Wikipedia produce prototypical activity-based roles of authors?
2. If so, what is the nature of activity-based roles?
3. How stable are activity-based roles across time?

To answer these questions, we first present a suitable selection of English Wikipedia articles for our analysis (section 5.2) and classify all revisions in this sample with the classifier presented in section 4.5. Using an unsupervised learning approach, we then cluster users to detect activity-based roles in Wikipedia. We analyze the nature of these roles and show that they are quite stable in Wikipedia (section 5.3, first and second research question). In a more detailed analysis, we compare the roles across time in Wikipedia (section 5.4, third research question).

5.1 The Concept of Emergent Roles

Organizational research recently began to develop theories of knowledge collaboration, calling into attention the emergent nature of roles in online collaboration (Faraj and Johnson, 2011; Kane et al., 2014). These *emergent roles* are based on authors' activity and enacted at the moment, and typically respond to previous contributions by co-authors. An emergent role is defined in terms of a set of activities, and authors can play different roles at different times. We will apply the concept of emergent roles to edit activity in Wikipedia and empirically investigate their nature, based on the findings and concepts presented in chapter 4.

In the context of online mass collaboration, there is typically a core group of users who feel responsible for the main development of content, while the vast majority of users (the so called *long tail*) performs rather simple changes (like spelling corrections) and edits infrequently (Priedhorsky et al., 2007; Wilkinson and Huberman, 2007). Wikipedia is no exception in this regard. Despite the fact that the core users carry out the bulk of the work (in terms of their number of edits), users who participate only sporadically in the development of an article, play an important role in maintaining and improving the overall quality and thereby the success of Wikipedia (Kittur et al., 2007a). However, little is known about the categories of edits users perform and whether they tend to concentrate on certain kinds of edits or whether they perform all kinds of edits. Majchrzak et al. (2013) and Yates et al. (2010) suggest two quite broad groups of users which can be distinguished based on their editing behavior: those adding new content and those who rewrite, reorganize, and integrate existing content (called *shapers*). However, this rather superficial categorization is too coarse to properly explain emergent roles. Furthermore, the survey-based categorization of Majchrzak et al. (2013) and Yates et al. (2010) fails to account for the intricacies of wikis (cf. section 3.2.1.1).

In this chapter, we will substantially extend previous studies which have analyzed the collaborative structures that have been driving the success of Wikipedia. Based on the notion of "social roles", Welser et al. (2011) identified four roles of Wikipedia users, which are partly defined in terms of the activity-based dimension. The roles are motivated based on the edit distribution of users across Wikipedia namespaces and on the explicit recogni-

tion of other users. Their roles include Substantive Experts (mainly involved in providing their expertise to expand articles), Technical Editors (mainly involved in the maintenance of content), Counter Vandalism Editors (search and revert vandalistic changes), and Social Networkers (mainly interact with other users through networking and communication). While Welser et al.'s (2011) study yields interesting results about roles in Wikipedia, the motivation of their proposed four roles remains vague. It is not clear how well the roles they suggest actually model the Wikipedia community, whether there are more or fewer than four roles, and whether the role definitions hold over a larger sample of users. As Welser et al.'s (2011) roles are not purely defined within the activity-based dimension, it is hard to empirically prove and motivate their existence.

To the best of our knowledge, Liu and Ram (2011) have carried out the only study to date which empirically tried to detect emergent roles in Wikipedia.[79] The edit category taxonomy on which their analysis is based (see section 3.2.4) contains ten categories, which can be detected using hand-written rules rather than machine learning techniques. Their categories are Sentence Insertion, Sentence Modification, Sentence Deletion, Link Insertion, Link Modification, Link Deletion, Reference Insertion, Reference Modification, Reference Deletion, and Revert. For their analysis, Liu and Ram (2011) used a set of 1600 English Wikipedia articles, composed of different Wikipedia-internal quality classes. Their set includes each 400 featured, A-class, B-class and C-class articles (cf. section 3.2.7 for an explanation of these quality classes). They calculated sentence-based edits from the revision history of the 1600 English Wikipedia articles and automatically labeled each with one of their categories. By aggregating the edit categories performed by users, they create *activity-based profiles*. These profiles were then grouped using a clustering algorithm. The resulting six clusters represent prototypical activity-based profiles, which they refer to as (cluster size is given in brackets):

- Content Justifiers: users who mainly performed Link Insertions and Sentence Insertions (29%)
- Copy Editors: users who have performed edits in several categories, but more than half of all edits are Sentence Modifications (26%)
- All-round Contributors: users who have performed edits across all ten categories, with Sentence Insertions, Modifications, Deletions, and Link Insertions and Deletions predominating (22%)
- Cleaners: users who mainly performed Link Deletions and to a lower degree Sentence Deletions (10%)

[79] Liu and Ram (2011) did not call the activity based rules "emergent", but their concept is essentially very similar to ours.

- Starters: users who mainly performed Sentence Insertions and to a much lower degree Link Insertions (9%)
- Watchdogs: users who have performed almost exclusively Reverts (5%)

While the work by Liu and Ram (2011) provides a good starting point for exploring emergent roles in Wikipedia, it suffers from several limitations. Most important, the edit category taxonomy they propose is clearly targeted towards the rule-based identification algorithm for edit categories. While their approach can be applied to a very large number of revisions without further effort (once the rules are pre-defined), its main drawback is that it disregards substantial information about the meaning of changes made at each revision. As a result, text-base edits (meaning-changing edits, e.g. modifying an important word to change the proposition of a sentence) and surface edits (meaning-preserving edits, e.g. correcting a spelling error) cannot be properly distinguished. Furthermore, the taxonomy does not include the action of inserting vandalism (although it does include vandalism removal, given it is easy to detect). It also ignores the important activity of *shaping*, which refers to rewriting, reorganizing, and integrating existing content. Previous research considers this to be the most important CW activity within Wikipedia (Kane et al., 2014; Majchrzak et al., 2013; Yates et al., 2010).

Second, Liu and Ram (2011) exclude casual users who made up over 80% of their sample. While excluding users who made few edits may reduce computational complexity, it ignores a very important section of the users and may obscure the analysis. For example, vandals – who tend to make only few edits – are largely excluded from the analysis. From the overall CW perspective, it does not make sense to exclude authors simply because they perform few edits - nor does it make sense to exclude authors with bad intentions (vandals) when at the same time including authors who repair the damage (mainly) caused by vandals.

Third, the analysis by Liu and Ram (2011) is based on a set of articles labeled with different quality classes, but these articles do not necessarily yield a set of representative Wikipedia articles. Within the context of Liu and Ram's (2011) study, this is feasible as they analyze the relationship between emergent roles and article quality. However, for a general analysis of emergent roles in the CW process of Wikipedia, this set of articles might be biased towards a non-representative set of articles which have received above average attention (cf. table 3.6 for the distribution and explanation of the quality classes). Forth, Liu and Ram (2011) performed a high-level analysis, but did not study the stability of roles produced by the clustering of the activity-based profiles. Furthermore, they do not analyze whether a user's role is stable across time.

In our study of emergent roles produced by the CW process in Wikipedia, we address the above limitations of Liu and Ram's (2011) work. As explained in section 5.2, our findings are based on a full set of user profiles from a representative sample of Wikipedia articles which accounts for both the topical as well as the maturity dimensions of articles. Next,

in section 5.3, we show how the edit classification approaches we presented in chapter 4 can be used to overcome the drawbacks of Liu and Ram's (2011) edit category taxonomy and cluster activity-based profiles. Finally, we both analyze the stability of the resulting emergent roles themselves as well as the stability of emergent roles across the temporal development of articles.

5.2 Creating a Representative Sample of Wikipedia Articles

To ground our analysis in representative data, we employed a double-stratified sampling procedure, randomly selecting 1000 articles from the January 2012 dump of the English Wikipedia. Our strata were based on: (a) the maturity of articles (in terms of the number of revisions), and (b) the articles' topical domains. This is important given that CW patterns could differ across articles in different phases of their life cycle and across topical domains (Arazy et al., 2011; Kittur et al., 2009a). Our sampling approach is in line with prior studies of Wikipedia (Arazy et al., 2011, 2013). Given the power law distribution in the number of articles' revisions (Ortega et al., 2008), we used the following four maturity strata: (a) 1-10 revisions, (b) 11-100 revisions, (c) 101-1,000 revisions, and (d) more than 1,000 revisions. The topical strata are based on Wikipedia's categorization system, using the main topics scheme.[80] The 25 categories contained in the scheme are: Agriculture, Arts, Business, Chronology, Concepts, Culture, Education, Environment, Geography, Health, History, Humanities, Humans, Language, Law, Life, Mathematics, Medicine, Nature, People, Politics, Science, Society, Sports, and Technology.

With four maturity strata and 25 topical categories, we have 100 cells with ten randomly selected articles in each stratum (i.e. 250 articles in each maturity stratum and 40 articles in each topical category). The aim of this sampling approach is to model a realistic (representative) distribution of edits within the CW process of the article namespace in the English Wikipedia.[81] Altogether, our sample contains 721,806 revisions, authored by 222,119 users. We refer to this sample as WPREP.

More than half of the users in WPREP performed only a single edit, whereas the most active user has performed 3815 edits across the entire article sample. Roughly 12% of all users in WPREP have been active four times or more, and 10% of all users were active in more than one article. Figure 5.1 shows the number of users plotted against the number of

[80] The English Wikipedia main topic categorization scheme can be found at `http://en.wikipedia.org/wiki/Category:Main_topic_classifications`, accessed May 25, 2015. Please note that this categorization is not static, as it is based on the category assignments of individual pages from the entire article space. The latter is subject to frequent change; in this analysis, we used a version from January 2014.

[81] This is not the same as modeling a representative sample of users and tracing their actions, as a representative sample of users should additionally include activity from other Wikipedia namespaces (incl. discussion, user etc., cf. section 3.2.1.1).

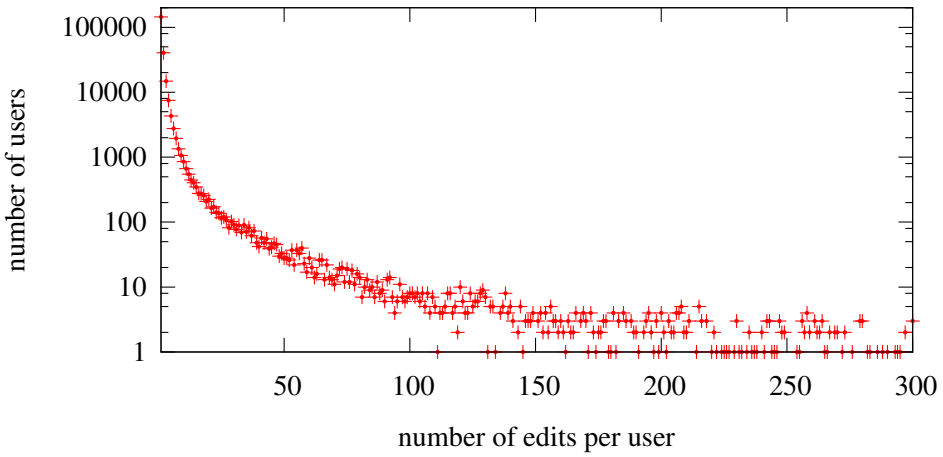


Figure 5.1: How many users performed how many edits across all articles in WPREP.

edits per user in WPREP. It is clearly visible that the distribution follows a power law, with a lot of users performing few edits and few users performing many edits.

5.3 Analysis of Emergent Roles in Wikipedia

We used the model explained in section 4.5 to classify all 721,806 revisions in WPREP. Since this is a large-scale study, where the focus is on user behavior rather than fine-granular textual changes, we decided to use the revision-based classification model rather than the edit-based model (cf. the discussion in section 4.5.3). The model used to predict the categories of all revisions in WPREP has been trained on the entire WPRC, which comprises 13,592 manually annotated revisions. As shown in section 4.5.2, the model performs quite well (micro-averaged F1 score of .78 on our test set), so that we expect reliable predictions. Roughly 4% of the revisions from our article set could not be classified with any category (the classifier had a confidence lower than the threshold), leaving us with 689,514 revisions with a valid category, contributed by 222,119 distinct users.[82] In contrast to prior studies that have focused on active users (Liu and Ram, 2011), we decided to include all users, even those with very few editing activities assuming that such activities are intentional rather than random and thus an important part of the CW process in Wikipedia. This inclusive approach enables us to model vandals and other types of occasional users (note that users with only one activity make up more than half of all users in our sample, as illustrated in

[82]Our sample comprised all types of users, including unregistered anonymous users and bots (cf. section 3.2.2.2). We acknowledge that anonymous editors identified by their IP address do not necessarily correspond to a single user. However, this has a rather low impact, and it is common practice for studies of Wikipedia to associate an IP address with a user (Arazy et al., 2011).

Category Label	Revisions	% rev.	% categ.
Add or Change Wiki Markup	294,444	42.7	29.3
Add Substantive New Content	203,501	29.5	20.3
Delete Vandalism	110,960	16.1	11.1
Insert Vandalism	98,445	14.3	9.8
Fix Typos and Grammatical Errors	95,182	13.8	9.5
Delete Substantive Content	51,304	7.4	5.1
References (to external sources)	44,247	6.4	4.4
Reorganize Existing Text	40,845	5.9	4.1
Rephrase Existing Text	39,219	5.7	3.9
Hyperlinks (to other Wikipedia pages)	12,435	1.8	1.2
Miscellaneous	11,533	1.7	1.1
Move or Create New Article	1,126	0.2	0.1
All	1,003,241	145.5	100.0

Table 5.1: Number and percentage of revisions/categories in WPREP labeled with a certain category, after automatic classification.

figure 5.1). Table 5.1 shows the distribution of edit categories in WPREP after the classification.

5.3.1 Clustering Users Based on Activity Profiles

Each of the users in our sample was represented through a vector listing the number of revisions performed by this user for each of the 12 categories of WPRC. Following Liu and Ram (2011), we initially assumed that a user may enact different roles in different articles and created several activity profiles for each user, one for each article he or she contributed to. Like this, we calculated 325,417 activity vectors. Given our goal of modeling emergent roles rather than individual user profiles, we normalized the activity profiles, dividing the count of revisions in each category by the overall number of revisions made by the particular user on the article at hand. We then employed a clustering algorithm to group users' activity profiles, referring to each cluster's centroid as the prototypical activity profiles. These prototypical profiles are interpreted as emergent roles.

The input to clustering are the users' activity profiles, a profile for each article $p_i \in P$ the user was active on. Let $e^1_{u_m,p_i}, e^2_{u_m,p_i}, \ldots e^{12}_{u_m,p_i}$ denote the number of each of the 12 categories (cf. table 5.1) performed by user u_m to the article p_i, where $e^T_{u_m,p_i}$ denotes the total number of revisions by user u_m to article p_i. Then, we define the activity profile vector of user u_m to article p_i as

$$\overrightarrow{prof}_{u_m,p_i} = \langle \frac{e^1_{u_m,p_i}}{e^T_{u_m,p_i}}, \frac{e^2_{u_m,p_i}}{e^T_{u_m,p_i}}, \ldots \frac{e^{12}_{u_m,p_i}}{e^T_{u_m,p_i}} \rangle.$$

We refer to this clustering as *article-dependent*, because profiles are bound to articles - as opposed to *article-independent* clustering, where users' profiles are generated across all articles they were active on.

Following Liu and Ram (2011), we employed the K-means clustering algorithm with Euclidean distance measure, which divides the input space (in our case, a user's activity profile) into k clusters. The algorithm groups activity profiles based on their distance to the closest centroid; the centroid itself represents the prototypical vector of a cluster. We iteratively tested K-means clustering for k clusters, where $k \in [2, 10]$; assuming that a clustering with $k > 10$ would result in clusters which are hard to interpret based on their centroids. In order to determine the optimal number of clusters, for each value of k, we calculated the cluster Compactness and Separation metrics for the results of K-means clustering (He et al., 2004). Compactness is based on the homogeneity of vectors in each cluster (smaller values indicate higher average compactness), Separation measures the overall dissimilarity between the clusters (smaller values indicate higher average separation). Following He et al. (2004), we define *Compactness* as

$$Cmp = \frac{1}{k}\sum_{i}^{k}\frac{v(c_i)}{v(W)},$$

where c_i is the ith cluster, W is the input space of all vectors, with $|W| = N$. v is the variety of input space X, defined as

$$v(X) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d^2(x_i, \bar{x})},$$

where $d(x_i, x_j)$ is the distance metric (here: Euclidean distance) between the vectors x_i and x_j, and $\bar{x} = \frac{1}{N}\sum_i x_i$. *Separation* is defined as

$$Sep = \frac{1}{k(k-1)}\sum_{i=1}^{k}\sum_{j=1, j\neq i} exp(-\frac{d^2(x_{c_i}, x_{c_j})}{2\sigma^2}),$$

where x_{c_i} is the centroid of cluster c_i, and σ is a Gaussian constant.[83] We combined the two metrics using the *Optimal Cluster Quality* (OCQ) measure (He et al., 2004). OCQ is defined as

$$OCQ(\beta) = \beta \cdot Cmp + (1 - \beta) \cdot Sep.$$

We give equal weight to *Cmp* and *Sep*, setting β to 0.5. Given that clustering results depend on the selection of initial random seeds, we instantiated the seeds using the K-means++ method (Arthur and Vassilvitskii, 2007), and iteratively tested a range of values for the initial seed.[84] The lowest OCQ score (indicating the best clustering quality) was obtained for $k = 7$.

[83] We set $\sigma = 1$ for all experiments.

[84] The experiments were carried out with the help of the Weka Data Mining Software (Hall et al., 2009).

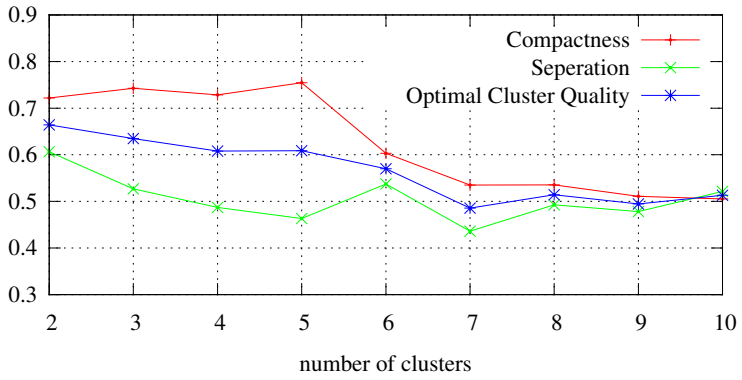


Figure 5.2: Compactness, Separation and Optimal Cluster Quality for $k \in [2, 10]$ (K-means clustering).

Figure 5.2 presents the Compactness, Separation, and OCQ metrics for the various cluster numbers. It should be noted that, as *Cmp* and *Sep* are not measured in the same dimension, OCQ cannot serve as a reliable measure for the overall quality or stability of a clustering solution. Rather, it should be used to compare a number of clustering solutions, as in our case, solutions for various k. In section 5.3.2, we will discuss a better way to analyze the stability of our clustering solution.

The centroids of the resulting seven clusters are displayed in figure 5.3. We titled each cluster with an intuitive name (names initially suggested by Liu and Ram (2011) are marked by †): All-Round Contributors†, Quick and Dirty Editors, Copy-Editors†, Content Shapers, Layout Shapers, Watchdogs†, and Vandals. The All-Round Contributors cluster has the highest percentage of users, 41% of all users' profiles in our sample are assigned to this cluster. As shown by its centroid, users with this role are active in many categories, with a slight tendency towards adding content and wiki markup. The Quick and Dirty Editors cluster (11%) represents users with a relatively clear focus on adding new content. However, some of their contributions were labeled as vandalism. Different from the Vandals cluster which has a clear focus on vandalism activities, the Quick and Dirty Editors cluster couples edits labeled as vandalism with the addition of new content. We assume that these contributions were mostly made in good faith, but did not comply with Wikipedia's extensive policies (e.g. neutral point of view, supporting claims by references, etc.), and thus might have been reverted later on. Copy-Editors show a clear tendency towards one activity category, namely fixing grammar and spelling errors. Rather few profiles have been labeled with the two shaping activities clusters: Content-Shapers (4%) concentrate on edits associated with the (re)organization of content; whereas Layout Shapers (6%) focus almost entirely on adding markup to an article. The Watchdogs and Vandals clusters both have equal size (13% of profiles) and contain users with a clear focus

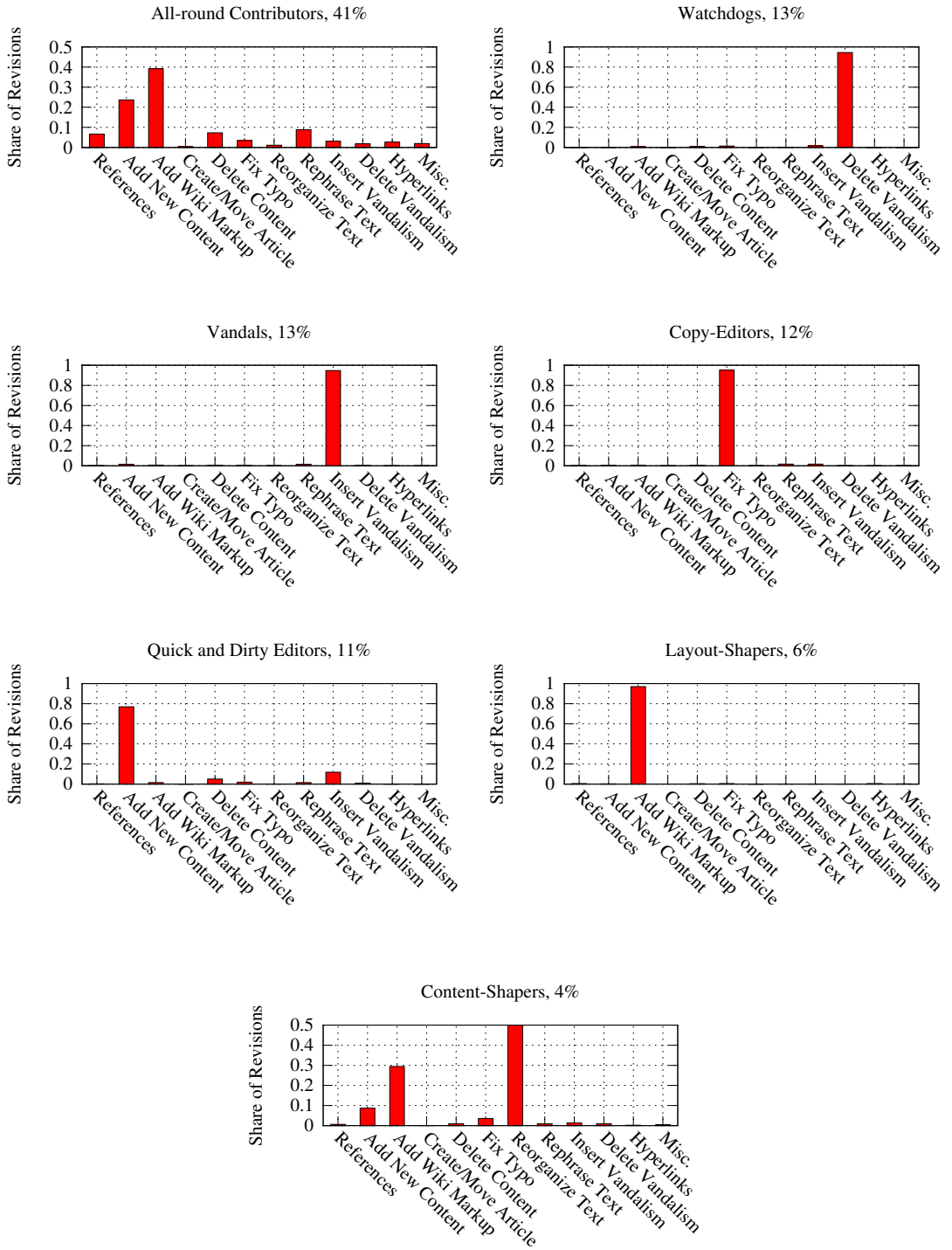


Figure 5.3: Centroids of the seven clusters, based on an analysis of the 1000 articles in WPREP; article-dependent setting.

on a single revision category, namely inserting or removing vandalism (reverting) respectively.

We compared our findings to the study of Liu and Ram (2011), who applied a different edit category taxonomy, a different labeling technique and a very different sample of activity profiles (different articles and users). Given these variations, we find a surprisingly high agreement in terms of the nature of emergent roles. Namely, our study confirms the existence of roles that focus on: adding new content (Quick and Dirty Editors, labeled "Starters" in Liu and Ram (2011)); making small corrections (Copy Editors); reverting vandalism (Watchdogs); or the general role of making various types of activities (All-Round Contributors). Nonetheless, our findings suggest that there are additional prototypical activity profiles that have not been detected previously: Content Shapers, Layout Shapers, and Vandals. The novel findings are to be attributed to the edit category taxonomy we have employed, allowing us to record shaping and vandalism activities. That is, we believe that these additional emergent roles have always existed, only that earlier studies did not have appropriate tools to detect them.

5.3.2 The Stability of Activity Profile Clusters

In section 5.3.1, we have analyzed the nature of emergent roles in Wikipedia, showing that a clustering solution with seven activity profile based roles yields best results. However, the degree to which these roles are stable, i.e. whether they can be generalized and are thus, meaningful, remains unclear. Evaluating the validity or quality of clustering solutions remains a challenging problem (Jain and Dubes, 1988). Although the concept of emergent roles entails certain fluidity, their existence as part of the CW process in online mass collaboration needs to be verified (cf. the first research question of this chapter).

In the following, our goal is to show that the clustering solution presented in section 5.3.1 yields a natural grouping of activity profiles based on the indirect interaction of users into emergent roles. To this end, we analyze the *stability* of our clustering solution. Clustering results may be unstable in the sense that different clusterings may claim to summarize a given data sample equally well, and we cannot tell which ones better reflect the intrinsic structure of the data (Bayá and Granitto, 2013). The metrics described above (Compactness, Separation and OCQ) are useful in determining the best clustering solution for a K-means algorithm on a given solution space, but cannot generalize to compare clustering solutions across algorithms and different input data spaces. Thus, to assess clustering stability, a much more general approach is required.

Lange et al. (2004) propose a validation method able to detect the number of arbitrary shaped clusters through training a classifier that learns the structure that was found by a clustering algorithm (using the "natural groups" produced by the clustering algorithm as labels of the input to the classifier). *Cluster stability* quantifies the reproducibility of a clustering solution by measuring the performance of a classifier trained and tested on the

labels produced by the clustering. Following Lange et al. (2004), we calculated the cluster stability, $\bar{S}(A_k)$, where A is the clustering algorithm and k is the number of clusters. In several rounds, we split the full data sample randomly in two halves X and Y. The average 0-1 loss between $A_k(Y)$ and a classifier prediction $\phi(Y)$ (ϕ is trained on $A_k(X)$) corresponds to the average dissimilarity of clustering solutions. After normalizing this value by the misclassification rate of a random labeling, we arrive at the cluster stability value, $\bar{S}(A_k)$. Lower values of $\bar{S}(A_k)$ suggest higher stability.

We calculated $\bar{S}(A_k)$ on the clustering solution presented in section 5.3.1 using K-means clustering, for both the article-dependent and the article-independent setting. As a first result, we found that cluster stability $\bar{S}(A_k)$, for values of $k \in [2, 10]$, reached a local minimum at $k = 7$, further corroborating our earlier findings regarding the optimal number of clusters. Second, the clustering stability values for the article-dependent setting, $\bar{S}(A_7)^d$, and the article-independent setting, $\bar{S}(A_7)^i$, were 0.31 and 0.33 respectively.[85] The average 0-1 loss between $A_7(Y)$ and a classifier prediction $\phi(Y)$ in the article-dependent setting is 0.27, indicating that the risk of instability in our clustering solution is not high. These numbers show that, at least to a certain degree, activity profile based emergent roles in Wikipedia are stable and thus, meaningful.

5.3.3 The Relationship Between Users and Emergent Roles

To analyze the relationship between users and roles, we first explored whether a user is typically associated with a single role (article-independent), or whether users take multiple roles across articles in Wikipedia (article-dependent). The intuition behind the article-independent assumption is that average users have general editing preferences which they apply across all articles they are active on. In the article-dependent setting, we assume that users show a different edit behavior depending on the content of the article they are active on. For example, a user might perform information-related edits only where he or she feels to have enough expertise with the content of an article, and otherwise remain with surface level edits such as spelling corrections or shaping edits.

We first analyzed the clustering solution produced by the article-independent assumption. Our results show that when a user has a single profile across all articles he or she edited, the profiles of the cluster centroids remain similar, as illustrated in figure 5.4. This finding is backed by the result from the stability analysis in chapter 5.3.2, where the values for $\bar{S}(A_7)^i$ and $\bar{S}(A_7)^d$ did not differ too much. It must be pointed out however, that only 10% of all users in WPREP are active across more than one article (cf. section 5.2), so that their potential influence on the overall clustering is naturally limited.

[85] Further parameters as listed in Lange et al. (2004): $r = s = 20$ (number of splits/iterations), classifiers tested for ϕ were SMO (Platt, 1998) and C4.5 (Quinlan, 1993).

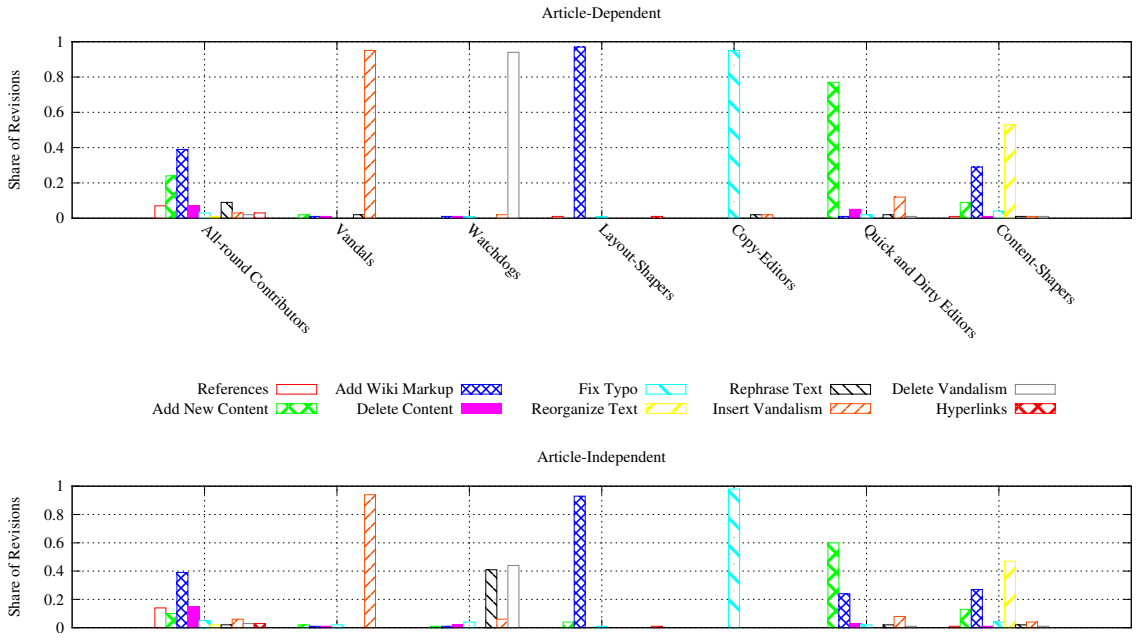


Figure 5.4: Centroids of the seven clusters, comparing article-dependent and article-independent clustering, based on WPREP.

As for the article-dependent assumption, we found that the set of activity profiles representing a user is on average associated with 1.1 roles. The vast majority of users were linked to only a single role and only relatively few users (7%) are associated with multiple clusters. However, we also discovered eight users who were assigned to all seven roles. All of these are associated with accounts which up to May 2015 have been used to create more than 10,000 revisions in the English Wikipedia, including one IP account which may have been used by more than 300,000 people in Singapore in the years 2005 and 2006. Hence, these users are either very active and experienced experts who have played different roles in different articles, or accounts that have been used for editing Wikipedia by more than one person. Further analysis reveals that predominantly users associated with the Vandals and Quick and Dirty Editors clusters do not play other roles. In contrast, users linked to the Watchdogs cluster often take on different roles in other articles they are active on.

We also analyzed the relationship between formal roles and emergent roles. To this end, we calculated the percentage of users associated with certain privileges (anonymous users, bots and administrators) for each of our clusters (Arazy et al., 2015, 2014). The results are summarized in table 5.2. For an overview of formal roles in Wikipedia, please refer to section 3.2.2.2. Anonymous users are identified by the IP address they are editing from, rather than their Wikipedia user name. Bots are flagged as such by their user access level, cf. section 3.2.2.2. Administrators are users associated with any of the `sysop` and `bu-`

Emergent Role	Cluster Size (%)	Anonymous (%)	Bots (%)	Admins (%)
All-round Contributors	41	62.6	0.3	1.2
Watchdogs	13	39.4	0.1	6.7
Vandals	13	90.2	0	0.1
Copy-Editors	12	69.0	0.6	1.8
Quick and Dirty Editors	11	79.9	0.3	0.5
Layout-Shapers	6	49.7	0.6	4.4
Content-Shapers	4	42.8	1.9	5.3
all (micro-avg.)	–	67.7	0.3	1.7

Table 5.2: Percentages of anonymous users, bot and administrators for each of our clusters (based on the article-dependent setting).

`reaucrat` access levels, as suggested by Arazy et al. (2014). Our analysis reveals that there are indeed some correlations between formal roles (as determined by user access level) and emergent roles. The Watchdogs and the two Shapers clusters all show a share of administrators above average. For the Watchdogs cluster, this is to be expected, as administrators have special privileges which allow them to quickly revert changes, e.g. to fight vandalism. Furthermore, administrators seem to be interested more than others in rewriting and reorganizing, as expressed by their increased share in the Content- and Layout-Shapers roles. Some of the tasks associated with the Content-Shapers role, seem to be carried out by bots, as shown by their above-average share in this cluster. Most of the users in the Vandals cluster are anonymous, and almost none of them are associated with special privileges.

As we explained earlier, in cases where a user is associated with more than one cluster it must not necessarily indicate that he or she *concurrently* plays multiple roles. Alternatively, this could also reflect changes in the editing behavior of a user over time (e.g. due to a user's progression in the Wikipedia community, typically reflected by the formal role(s) issued to this user). In order to test this assumption, in section 5.4, we analyzed the (possible) time-dependency of activity profiles.

5.4 Emergent Roles Across Time

In order to test whether the prototypical activity profiles are stable over time, we divided the timespan covered by the processed revisions (year 2001 until early 2012) into two phases: from 2001 to end of 2006, and from 2007 to early 2012. The years from 2001 until end of 2006 are the early phase and phase of growth in the English Wikipedia according to Halfaker et al. (2012). As discussed in section 3.2.2.1, around the year 2007, there has been a shift in the growth of the number of active users, and numbers have been declining rather

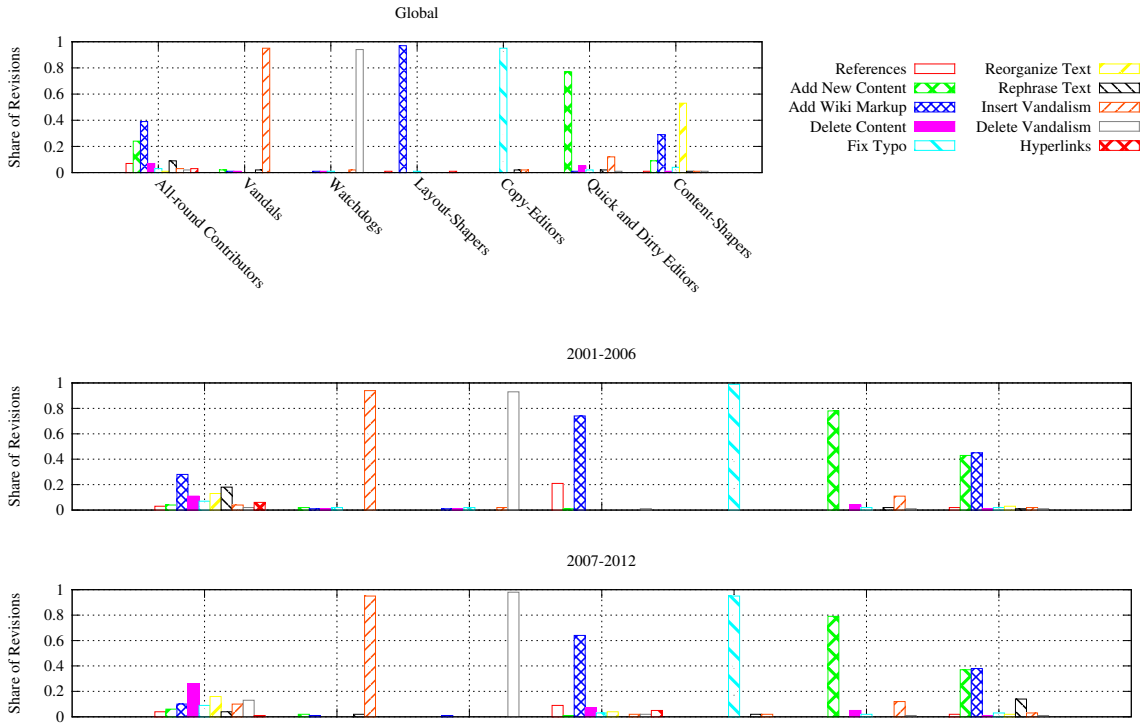


Figure 5.5: Centroids of the seven clusters, comparing two time periods. For comparison, we added the centroids of the article-dependent clustering on top of the figure.

than increasing since then. As this shift apparently marks a significant change in the CW community of Wikipedia, we decided to use it for the division into time periods. For each period, we created profiles of users' activity. We applied the article-dependent clustering procedure described earlier. There were 96,757 activity vectors in the early period and 233,687 in the later period. Based on the Euclidean distance between centroids, we mapped the most similar clusters from the early and the later phases. Figure 5.5 shows the clustering solutions for the two periods, compared to the clustering for the entire article history ("Global"), indicating the extent to which prototypical activity profiles changed over time.

A visual inspection shows that the centroids of the aligned clusters are very similar. Overall, this is a strong indication that the nature of emergent roles (defined in terms of centroids' activity profile) changed very little between the two time periods.

As the nature of emergent roles at both periods appears to be very similar, we established a mapping between the clusters in one period and those in the other period. To do so, each centroids from one period was mapped to the closest centroid of the other period (based on Euclidian distances). Like this, we were able to analyze whether a user is associated with different emergent roles across the two periods. We found that among the 5027 users active within the same article at both periods, more than 50% changed their role over time. A closer look at the role transitions reveals that users tend to move towards the Lay-

out Shaper role (incoming: 956; leaving: 346), and to a lesser extent to the Watchdogs role (incoming: 482; leaving: 187). In contrast, users tend to leave the All-Round Contributors role (incoming: 378; leaving: 1236). These role transitions suggest that while the nature of emergent roles is quite stable across time, users do change their role (within the same article). More complex roles (Shapers, Watchdogs) seem to be preferred targets. The transitions also reflect the changes in the structure of the Wikipedia community, which according to Halfaker et al. (2012) resulted in a decline of the open collaboration system. The introduction of a more complex formal role system, intended to improve mechanisms for quality management, effectively caused a higher rejection of newcomers' edits, which consequently kept new users from joining the community.

5.5 Implications beyond Wikipedia

In this chapter, we have analyzed activity-based roles on a corpus of representative Wikipedia articles. We model activity-based roles as emergent roles, a concept that has arisen in organizational research, based on indirect user interaction. We have shown that participants of the CW process in Wikipedia indeed take emergent roles, and that these roles are quite stable; although users might change their role over time. In the following, we will show the relevance of our findings about emergent roles in Wikipedia beyond the special use case of Wikipedia. We first present some theoretical considerations, before we turn to the practical implications of our findings.

The experiments in this chapter yield strong indications that the CW process in online mass collaboration produces emergent roles. While this finding was already backed in previous work (Liu and Ram, 2011), our experiments and empirical findings go several steps further and investigate the concept of emergent roles in depth. The nature of activity-based roles obviously depends on the editing behavior of authors in the given community. For example, vandalism fighting will only influence the nature of roles in those CW projects in which vandalism is a serious problem. On the other hand, the emergent roles which we identified for the CW process in Wikipedia indicate that a distinction between core authors and peripheral authors involved solely in superficial editing such as spelling corrections, is too coarse-grained. Our study revealed that, beyond the core group of authors who perform all kinds of edits and are mainly involved in the development of documents (All-round Contributors), there are at least three groups of peripheral authors: Copy-Editors who care about the linguistic quality of documents and the two Shapers roles, who deal with surface edits on the layout and content level, respectively.

Furthermore, we have shown that the implementation of formal roles in online CW systems is likely to have an influence on the editing behavior of authors and thus, emergent roles. For example, authors with advanced privileges are more likely to be involved in vandalism fighting and shaping activities. Beyond the influence of formal roles on emergent

roles, our findings also reveal potential correlations between changes in CW community policies and emergent roles (Halfaker et al., 2012). As explained in section 3.2.2, the introduction of quality and consistency management tools around the year 2007 produced a drop in the number of new authors. In section 5.4, we have shown that, although the nature of emergent roles did not significantly change with the introduction of new policies, more than half of all authors active before and after 2007 were associated with different roles in the two periods. There is a clear tendency to take more complex roles later on, with authors giving up their All-round Contributors role in favor of the Watchdogs and Layout Shapers roles. The latter both represent roles dealing with quality and consistency of documents, i.e. roles which at best ignore, or, more likely, reject the edits of newcomers.

There are indications that the above consequences are not unique to Wikipedia, but that other open online collaboration systems as well suffer newcomer decline after an initial phase of growth.[86] Although the circumstances might be different across CW platforms, we assume that some of the underlying factors which lead to this development are universal. It should be noted, though, that the decline in community structure can be interpreted in different ways. Low retention of newcomers does not necessarily weaken the quality of content. Some might argue that "mature" CW projects do not need to constantly increase participation. On the other hand, low newcomer retention certainly does not help to address the systematic bias explained in section 3.2.2. It might be too early to judge on this matter, but a constantly low newcomer retention can become a serious challenge for the future development of Wikipedia and other online CW projects.

Beyond our contributions to the theory in the area, our findings also have important practical implications for designers and administrators of online CW communities. Understanding emergent roles of authors can help in the allocation of tasks, such that authors with different skill sets and interests could find suitable ways to spend their time, knowledge and energy. Specific administrative implications include: (a) designing task forces to tackle a certain job, bringing together authors with relevant activity profiles, (b) providing a shared space for people with similar roles to share their experiences and learn from one another, (c) placing more emphasis on the early detection and direction of authors who seem to be suitable (as indicated by their activity profiles) for particular roles, and (d) recommendations for authors to change their role from time to time, e.g. to limit the rejection of newcomers' edits. A design implication would be to develop tools that track the editing

[86] A 2014 discussion on the Wikipedia Research Mailing List has addressed low newcomer retention in Wikipedia and other online platforms extensively, see http://lists.wikimedia.org/pipermail/wiki-research-l/2014-December/thread.html#3983, accessed May 25, 2015. Furthermore, see e.g. http://meta.serverfault.com/questions/6701/server-fault-needs-professional-quality-questions-not-just-questions-from-profe, accessed May 25, 2015 and https://medium.com/technology-musings/on-the-future-of-metafilter-941d15ec96f0, accessed May 25, 2015.

behavior of authors and identify tasks of interest for them (Zhang et al., 2014). This could include offering "career guidance" (Cosley et al., 2006), including suggestions for roles that best match their profile.

5.6 Conclusion

In this chapter, we have proposed and analyzed activity-based roles in Wikipedia. We instantiated the concept of emergent roles in Wikipedia, which are based on activity profile vectors of users in a representative corpus of articles from the English Wikipedia. Beyond the analysis of the emergent roles themselves, we have also shown that the nature of these roles is quite stable across different input models and also across time. Based on the questions raised in the beginning of this chapter, we have come to the following answers.

The CW process in Wikipedia indeed produces prototypical activity-based roles, that can be modeled as emergent roles (first research question). Based on a clustering approach, we have detected seven such roles, which we interpreted based on the centroids of the clusters. A stability analysis showed that our clustering solution is stable.

With respect to the nature of emergent roles (second research question), we discovered the following. Despite substantial differences in our methodology to produce clusters of activity profiles (e.g. different input data, different taxonomy), our findings confirm several of the roles discovered by the earlier study of Liu and Ram (2011). However, beyond the All-round Contributors, Copy-Editors, Quick and Dirty Editors, and Watchdogs roles which were already discovered by them, we additionally identified Content- and Layout-Shaping roles, and Vandals. We attribute these novel findings to the detailed activity taxonomy we have employed, allowing us to record shaping and vandalism edits. That is, we believe that these additional emergent roles have always existed, only that earlier studies did not have appropriate tools to detect them.

To answer the third research question, we divided our input space into revisions performed before and after 2007. For each of the two periods, we created a clustering of user activity profiles and mapped the resulting clusters from the earlier to later time period. To find out whether users change or remain within certain roles, we analyzed the affiliation of users to emergent roles across two time periods and found that there is considerable variation. More than half of all users active both before and after 2007 changed roles over time, even within the same article. This behavior is likely to be related to changes in Wikipedia's community structure, namely an increased focus on content quality and consistency.

Emergent roles based on indirect user interaction are an important component of the CW process. We have analyzed their nature and potential influence on CW community structure. As argued in section 5.5, researchers, users and developers of online CW systems should be aware of the existence of emergent roles and carefully consider their influence on the CW process. Future work on emergent roles in Wikipedia should analyze whether

and how activity-based roles become manifested in other dimensions, e.g. in the direct interaction with other users. Like that, the social nature of emergent roles (Welser et al., 2011) could be analyzed in more detail. For example, it would be very interesting to find out whether users who change to more complex emergent roles over time (cf. section 5.4) also change their relations to co-authors. This effect could be studied either via a co-author network analysis (section 3.1.1.3) or by analyzing the direct interaction of these users on, e.g. discussion pages.

Chapter 6

Corresponding Edit-Turn-Pairs in Wikipedia

In chapter 3 we have introduced the concepts of direct and indirect interaction in CW. In chapters 4 and 5, we have focused on indirect interaction taking place when articles in Wikipedia are revised. In this chapter, we present a framework to analyze the relationship between direct and indirect user interaction in Wikipedia. We build on our findings about revision in Wikipedia and connect these to the activity of Wikipedians on discussion pages. We introduce the concept of edit-turn-pairs, which model potential connections between Wikipedia article edits and discussion pages. The following questions will be addressed:

1. What is the nature of correspondence between Wikipedia article edits and discussion page turns?
2. What are the distinctive properties of corresponding edit-turn-pairs and how to use these to automatically detect corresponding pairs?
3. What is the impact of corresponding edit-turn-pairs in Wikipedia and what do they tell us about the relationship between direct and indirect user interaction?

To answer the first research question, we define the notion of corresponding and non-corresponding edit-turn-pairs (chapter 6.1). Building upon this definition, we create a corpus of corresponding and non-corresponding edit-turn-pairs, which was annotated with the help of crowdsourcing (chapter 6.2). We turn to the second research question using the annotated corpus in a machine learning setting with the goal to learn a model which can automatically detect corresponding edit-turn-pairs (chapter 6.3). With the help of the resulting model, we analyze the impact of edit-turn-pairs across various articles in the English Wikipedia (third research question, chapter 6.4).

6.1 A Framework to Extract Edit-Turn-Pairs from Wikipedia

As noted by Marttunen and Laurinen (2012), discussing and revising are the most prominent activities in CW. With the notion of edit-turn-pairs, we create a means to analyze the connections between these two activities. We propose a framework to extract segments from discussion pages (called turns) and edits from the respective article. Edits have already been introduced and discussed in detail in chapter 4. In the following, we will additionally introduce the concept of *turns* extracted from Wikipedia discussion pages. We then define a number of constraints for *corresponding edit-turn-pairs*, i.e. pairs of an edit and a turn which create a link between an article's edit history and its discussion page. Consider the following snippet from the discussion page of the article "Boron" from the English Wikipedia. On February 16th of 2011, user `JCM83` added the turn:

Shouldn't borax be wikilinked in the "etymology" paragraph?

Roughly 5 hours after that turn was issued on the discussion page, user `Sbharris` added a wikilink in the "History and etymology" section of the article by performing the following edit:

'' borax'' → [[borax]]

This pair of an edit and a turn is what we define as a corresponding edit-turn-pair. Before we turn to the detailed definitions, we give a short motivation for our research on edit-turn-pairs.

6.1.1 Motivation

In online mass CW, direct user interaction is frequently happening in dedicated discussion spaces, as the coordination of writing cannot take place in face-to-face meetings (cf. section 3.1). The implicit knowledge which is created by this kind of interaction is typically hidden from the explicit writing process. Wikipedia offers discussion pages for direct interaction, but there is no technical support to map the results (i.e. the generated knowledge) from discussions to "tasks" which have been carried out in the form of edits to the article itself. Due to that reason, there is no explicit task management system (as commonly used in e.g. software development) available in Wikipedia. Apart from the organizational advantages of a task management system, analyzing the amount of knowledge in the article that was actually generated within discourse on the discussion page, poses an interesting problem (Cress and Kimmerle, 2008). To understand the influence of a discussion on the article text, it is necessary to map debates to those parts of the article content that have been shaped by the debate. Mapping turns (fine-grained segments from a discussion) to edits (fine-grained

changes to the article) enables a very detailed analysis of the flow of knowledge generated in a discussion among authors into the article content.

The detection of corresponding edit-turn-pairs also has practical implications for users of Wikipedia, as it might help to better understand the development of articles. Instead of having to read through all of the discussion page which can be an exhausting task for many of the larger articles in the English Wikipedia, users could focus on those discussions that actually had an impact on the article they are reading. Additionally, readers of an article could investigate the development of certain passages of an article if they were able to link the respective passage to a discussion thread. In the discussion, they might find additional material which better explains the content of the article and why the article has been written the way they found it. For example, the usage of a certain terminology in an article might be controversial and therefore negotiated on the discussion page. Once the dispute comes to an agreement and all relevant terminology in the article itself has been updated according to the result of the discussion, a link to the respective discussion page thread could be placed in a suitable location of the article.[87]

6.1.2 Corresponding and Non-Corresponding Edit-Turn-Pairs

Edits (as defined in section 4.1.2) are coherent modifications calculated from a pair of adjacent revisions from Wikipedia article pages. Edits are associated with metadata from the revision they belong to. This includes the comment (if present), the user name and the time stamp. In the above example, user `Sbharris` added the following comment to his edit: "Link the first use of borax".

Turns are segments from Wikipedia discussion pages (cf. section 3.2.5). To segment discussion pages into turns, we follow the procedure proposed by Ferschke et al. (2012a). With the help of the Java Wikipedia Library (Zesch et al., 2008), we access discussion pages from a database. Discussion pages are then segmented into topics based on the structure of the page (topics are separated by headlines). Individual turns are retrieved from topics by considering the revision history of the discussion page. Ferschke et al. (2012a) report that this procedure successfully segmented 94% of the turns in a corpus based on Simple English Wikipedia[88] turns. In this experiment, we are working with data from the English

[87] In Wikipedia, template messages (see section 3.2.1.1) are already used for such purposes, e.g. to inform users about a potential issue of neutrality in the article, which is or has been discussed on the discussion page. For example, the {{POV-check}} template (`http://en.wikipedia.org/wiki/Template:POV-check`, accessed May 25, 2015) is used to indicate that the article has problems with Wikipedia's neutral point of view policy, and that these are or should be addressed in a certain topic of the discussion page. The main drawback of the current usage of templates is that they are not used consistently. By far not all of the template messages are placed in the right location in the article and point to the right topic in the discussion page. Furthermore, since human interaction is required to add the templates, they are missing in many places. Our proposed approach to detect edit-turn-pairs could be used to automate or at least semi-automate this process by suggesting users where and when to place such templates.

[88] `http://simple.wikipedia.org`, accessed May 25, 2015

Wikipedia. Due to the larger average size of topic threads in discussion pages in the English Wikipedia, the number of segmentation errors might be slightly higher as compared to the Simple English Wikipedia. The segmentation of Wikipedia discussion pages into turns remains a challenging problem and we are not aware of any algorithm able to solve this task with perfect accuracy. Algorithms based on signatures and indentation reported in related work (Laniado et al., 2011) do not work perfect either, as not all turns are signed and indentation can be reset in longer discussions. Along with each turn, we store the name of the user who added the turn, the time stamp, and the name of the topic to which the turn belongs. In the above example, the turn is part of a topic with the name "Add a link?". This topic consists only of `JCM83`'s turn and a subsequent turn by `Sbharris`, who later on added the wikilink.

An edit-turn-pair is defined as a pair of an edit from a Wikipedia article's revision history and a turn from the discussion page which is bound to the same article. If an article has no discussion page, there are no edit-turn-pairs for this article. Ferschke et al. (2012a) suggest four types of explicit performatives in their annotation scheme for dialog acts of Wikipedia turns. Due to their performative nature, we assume that a turn labeled with one of these dialog acts makes a good candidate for a corresponding edit-turn-pair. We therefore define an *edit-turn-pair* as *corresponding*, if

1. the turn is an *explicit suggestion, recommendation or request* and the edit performs this suggestion, recommendation or request,

2. the turn is an *explicit reference or pointer* and the edit adds or modifies this reference or pointer,

3. the turn is a *commitment to an action in the future* and the edit performs this action, or

4. the turn is a *report of a performed action* and the edit performs this action.

We define all edit-turn-pairs which do not conform to the upper classification as *non-corresponding*.

Figure 6.1a shows examples for all of the four kinds of correspondence. The first example displays a turn by a user who suggests to add a special infobox. In the corresponding edit, an image is replaced with this infobox. In the second edit-turn-pair, the turn contains a URL to a reference that is missing in the article. The corresponding edit is performed by a user who inserts that reference to the respective sentence in the article. The third example contains a commitment to delete a text snippet that should have been deleted earlier and has gone unnoticed. In the corresponding edit, the respective text section is deleted from the article. In the last example from figure 6.1a, a user reports the addition of a footnote in the turn. This note has been added to the article before, as shown by the corresponding edit.

Explicit suggestion:

Now that I've taken an interest in the article one other item that might be useful in this very good article is the theatre infobox.	~~**[[Image:5th Ave Theater Marquee (Seattle) 2007-08.jpg]]{{Infobox Theatre}}**~~

Explicit reference:

The A330 comparison here is per the Flug Revue [LINK]. The comparison in the A340 article is currently unreferenced there.	The 767-400ER's closest competitor from Airbus is the [[Airbus A330]].<ref name=Flug767-400>[**LINK**]</ref>.

Commitment to action:

Yeeks, you are right! I've looked back through the page history, and it wasn't deleted by mistake any time recently. I'll correct the text. Thanks for finding that!	~~**The illustration on the right shows a thin section of one hemisphere of the brain of a Chlorocebus monkey, [...] or different types of brain tissue, in distinct ways; the Nissl stain shown here is probably the most widely used.**~~

Report of action:

I think I found the source of the confusion. Donald indicates 2 justices of the Illinois Supreme Ct. licensed AL to practice in Sept. of 1836, but the Supreme Ct. Clerk did not enroll him until March 1, 1837. I added a footnote to this effect.	Admitted to the bar in 1836, [...] **<ref> AL was added to roll of attorneys by the Clerk, Donald (1996), p.64.</ref>** he moved to Springfield, Illinois and began to practice law under John T. Stuart, Mary Todd's cousin.

(a) Corresponding edit-turn-pairs.

There are several compounds harder than cubic boron nitride, most of which are nanocomposites.	Unless otherwise stated, ~~**S**~~**s**tandard temperature and pressure conditions were used.

(b) A non-corresponding edit-turn-pair.

Figure 6.1: Corresponding and non-corresponding edit-turn-pairs, adapted from real-world examples. Turns are to the left; edits to the right. Formatting conventions: inserted text is boldfaced, deleted text is boldface and crossed out.

Figure 6.1b shows an example for a non-corresponding edit-turn-pair. In the turn, a user criticizes the factual accuracy of the article, whereas the edit is a spelling error correction. The latter is not related in any of the defined ways.

6.1.3 Previous Approaches

Despite the various applications given in section 6.1.1 which motivate bringing article history and discussion together, few previous works have tackled this task. Our intuition is that the lack of a suitable formalization of both the units of analysis and the task itself, as well as the inherent class imbalance problem explained subsequently in section 6.2.1, lead to the high complexity of this problem.

Besides the work by Ferschke et al. (2012a) which is the basis for our turn segmentation, there are several studies dedicated to discourse structure in Wikipedia. Viégas et al. (2007a) propose 11 dimensions to classify discussion page turns. The most frequent dimensions in their sample are requests for coordination and requests for information. Both of these may form part of a corresponding edit-turn-pair, according to our definition in section 6.1.2. A subsequent study (Schneider et al., 2010) adds more dimensions, among these an explicit category for references to article edits. This dimension accounts for roughly 5 to 10% of all

turns. Kittur and Kraut (2008) analyze correspondence between article quality and activity on the discussion page. Their study shows that both implicit coordination (on the article itself) and explicit coordination (on the discussion page of the article) play important roles for the improvement of article quality. In the present study, we have analyzed cases where explicit coordination leads to implicit coordination and vice versa.

Yasseri et al. (2012) find a positive correlation between the length of the discussion of a Wikipedia article and its controversiality, which they measure by the number of mutual reverts. Kaltenbrunner and Laniado (2012) analyze the development of discussion pages in Wikipedia with respect to time and compare dependencies between edit peaks in the revision history of the article itself and the respective discussion page. They find that the development of a discussion page is often bound to the topic of the article, i.e. articles on time-specific topics such as events grow much faster than discussions about timeless, encyclopedic content. Furthermore, they observe that the edit peaks in articles and their discussion pages are mostly independent. This partially explains the high number of non-corresponding edit-turn-pairs and the consequent class imbalance. Arazy et al. (2011) analyze the relationship between task conflict as reflected in Wikipedia's discussion pages, group composition (administrative-oriented vs. content-oriented) and information quality. They find that conflict on the discussion page negatively corresponds to article quality for homogenous group compositions.

While there are several studies which analyze the high-level relationship between discussion and edit activity in Wikipedia articles, very few have investigated the correspondence between edits and turns on the textual level. Among the latter, Ferron and Massa (2014) analyze 88 articles and their discussion pages related to traumatic events. In particular, they find a correlation between the article edits and their discussions around the anniversaries of the events.

6.2 Creating a Corpus of Annotated Edit-Turn-Pairs

To verify that our definition of corresponding and non-corresponding edit-turn-pairs is applicable in practice, we manually annotated a corpus of English Wikipedia edit-turn-pairs according to the definition given in section 6.1.2. In the following, we first explain the class imbalance problem we are facing, and how we solved it. The annotation of edit-turn-pairs was carried out with the help of crowdsourcing. Details about the annotation study and the resulting corpus are described in section 6.2.3.

6.2.1 The Class Imbalance Problem

We want to better understand the general connection between the activity on the discussion page of a Wikipedia article and the development of the article itself. Therefore, we

calculated the average overlap of users who had both edited an article (i.e. users who have created at least one revision in the article history, referred to as editors) and also participated in the discussion about this article (i.e. users who have created at least one revision in the history of the discussion page or one of its archives, referred to as discussants). The average proportion of editors to discussants in the English Wikipedia is about 8:1. By mid of 2014, 10% of the editors of an average English Wikipedia article have also contributed to its discussion page. To the opposite, 37% of the discussants of an average article have edited the article itself. This shows that there is a considerable amount of users (more than half of the discussants) who only contribute to the CW process via direct interaction on the discussion page.[89] However, the vast majority of users prefer to contribute solely via indirect interaction (90% of the editors).

Although editors might follow a discussion without actually contributing to it, we expect that a correspondence between article edits and turns requires a certain number of users active on both the article and the respective discussion page. If we use the number of editors who are involved in discussion as a proxy to estimate the upper bound of the share of corresponding edit-turn-pairs, we must assume that the number of corresponding pairs will be rather low as compared to non-corresponding pairs. This conclusion is reinforced by the fact that we are facing a pair classification problem (Jamison and Gurevych, 2014), i.e. we are combining each instance from one resource (edits from article revisions) with each instance from another resource (turns from discussion pages). For an article with x edits and y turns, each new edit creates y new edit-turn-pairs and each new turn creates x new edit-turn-pairs. If the new edit or turn does correspond to one or more edits or turns, it will automatically create a large number of non-corresponding edit-turn-pairs if x and y are big enough. The revision history of larger articles in the English Wikipedia often contains more than 1000 revisions, and the same applies to popular discussion pages. As a result of this consideration, we are to expect a highly imbalanced class distribution in a random sample of edit-turn-pairs extracted from Wikipedia (few corresponding and many non-corresponding edit-turn-pairs).

The class imbalance problem becomes important at two different stages of the task. First, for the annotation study, we need to present the annotators at least with a certain number of corresponding edit-turn-pairs to avoid them from labeling all instances as non-corresponding. Second, for the machine learning part, the classifier again needs to be trained with a minimum number of corresponding edit-turn-pairs. In section 6.2.2, we explain our approach to solve the first problem. With regard to machine learning, one way to address class imbalance is cost-sensitive classification (Elkan, 2001). Cost-sensitive

[89] Our analysis includes unregistered users who are identified via their IP address. That means that we do not distinguish users who contribute from different IPs to discussion and article history, or use a different Wikipedia account for their activity in a discussion page and in an article. We assume that this happens only in exceptional cases, and that the numbers reflect a realistic image.

classification helps to prevent the classifier from learning a trivial classification by labeling all instances with the majority class. A misclassification matrix is used to specify the cost for misclassified instances of the majority and minority classes. The effect of cost-sensitive classification can be very different depending on the algorithm underlying the classification (Elkan, 2001).

6.2.2 Creating a Corpus of Edit-Turn-Pairs

As argued in section 6.2.1, it might be necessary to search a large number of edit-turn-pairs to actually find a corresponding edit-turn-pair. It was important to find a reasonable number of corresponding pairs before the annotation study could take place, as we needed a certain number of both corresponding and non-corresponding edit-turn-pairs for quality control in the crowdsourcing annotation study. Therefore, we chose to take a stepwise approach to create the final corpus.

6.2.2.1 Limiting the Search Space for Edit-Turn-Pairs

We started with a basic sample of 26 random English Wikipedia articles. Within this sample, we calculated all edits in the revision history of the article pages based on the algorithm explained in section 4.1.1. For each article's discussion page and its potential archives, we calculated all turns following the procedure described in Ferschke et al. (2012a). To reduce the number of non-corresponding edit-turn-pairs (and hence, the class imbalance problem), we applied several filtering steps:

1. We automatically labeled the edits with the help of a model similar to the one described in section 4.3.3.2.[90] All edits labeled as Vandalism, Revert and Other were excluded from further processing.

2. We removed all edits which are part of a revision whose user is a bot, based on the Wikimedia user group scheme.[91]

3. We excluded trivial edits which only modify non-word characters or whitespace characters.

4. Based on the findings from manual analysis, we limited the time span between edits and turns to 86,000 seconds (roughly 24 hours).

The intuition behind steps 1 – 3 is that they will mostly filter out edits which cannot form part of a corresponding edit-turn-pair. Limiting the time span between edits and turns to 24

[90] We used a Random Forest classifier (Breiman, 2001) as base classifier; bipartition threshold 0.25; all other parameters as described in section 4.3.3.2

[91] `http://meta.wikimedia.org/wiki/User_classes`, accessed May 25, 2015

hours (step 4) certainly causes us to miss a number of corresponding edit-turn-pairs, but as we will show in section 6.2.3.3, it is very effective in reducing the class imbalance problem. We refer to the corpus resulting from combining all edits and turns after the filtering steps as ETP-ALL. Overall, ETP-ALL contains 13,331 edit-turn-pairs.

6.2.2.2 Preliminary Annotation Study

To detect corresponding edit-turn-pairs in ETP-ALL, we carried out a manual annotation study with the help of one graduate student. This step is intended to make sure that we have a critical number of corresponding edit-turn-pairs for the Mechanical Turk annotation study. Rather than annotating all 13,331 edit-turn-pairs in ETP-ALL, the annotator explicitly searched for corresponding pairs, based on the content of turns and changes in consecutive article revisions. As a result, a set of 262 edit-turn-pairs have been annotated as corresponding. However, we still expect a certain number of non-corresponding edit-turn-pairs in this data, as the correspondence was judged based on the entire revision and not the individual edit. For example, given a revision containing three edits, one of which is actually corresponding to a turn from the discussion page, all three possible edit-turn-pairs were annotated as corresponding. We refer to these 262 edit-turn-pairs as ETP-UNCONFIRMED.

6.2.3 Mechanical Turk Annotation Study

For the Mechanical Turk annotation study, we selected 500 random edit-turn-pairs from ETP-ALL excluding ETP-UNCONFIRMED. Based on the assumptions given in section 6.2.1, we expect to find mostly non-corresponding pairs among them. From ETP-UNCONFIRMED, we selected 250 random edit-turn-pairs. The resulting 750 pairs have each been annotated by five Mechanical Turk workers. The Mechanical Turk workers were presented the article title, the turn text, the turn topic name, the edit and its context, and the edit comment (if present). Any wiki markup used to specify formatting, links or templates has been removed from the text of the turns. The context of an edit is defined as one preceding and one following paragraph of the edited paragraph. Each edit-turn-pair could be labeled as "corresponding", "non-corresponding", or "can't tell". One human intelligence task (HIT) consisted of five edit-turn-pairs. Further details about the task description and the layout of the HITs can be found in appendix C.2.

6.2.3.1 Quality Assurance

The requirements which Mechanical Turk workers had to fulfill to participate in the annotation study included a minimum number of 2000 completed HITs, an acceptance rate of at least 97% and age over 18. We manually picked 20 corresponding edit-turn-pairs (labeled as corresponding by the author of this thesis) from the 750 pairs as seed examples. We equally distributed them across the HITs, so that we could rule out bots and vandals who labeled

the pairs randomly or as "non-corresponding" all the time. In addition to the general requirements, we only allowed Mechanical Turk workers to participate if they did a good job on either a pilot study which included the seed examples or a qualification test which also included the seed examples. Although these measures decreased the number of attracted Mechanical Turk workers and thereby extended the time necessary to collect all annotations, they helped to ensure the quality of the annotations as shown by the evaluation of the final gold standard in section 6.2.3.2.

6.2.3.2 Inter-rater Agreement and Gold Standard Creation

The average pairwise percentage agreement over all edit-turn-pairs is 0.66. This was calculated as

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{c=1}^{C} v_i^c}{C},$$

where $N = 750$ is the overall number of annotated edit-turn-pairs, $C = \frac{R^2-R}{2}$ is the number of pairwise comparisons, $R = 5$ is the number of raters per edit-turn-pair, and $v_i^c = 1$ if a pair of raters c labeled edit-turn-pair i equally, and 0 otherwise. The moderate pairwise agreement is an indicator for the complexity of this task for non-experts.

We created the final votes (gold standard) with the help of majority voting. An edit-turn-pair was counted as corresponding, if it was annotated as "corresponding" by least three out of five annotators. Likewise, an edit-turn-pair is non-corresponding in the gold standard, if it was annotated as "non-corresponding" by at least three out of five annotators. Furthermore, we deleted 21 pairs for which the turn segmentation algorithm clearly failed (e.g. when the turn text was empty). This resulted in 128 corresponding and 508 non-corresponsing pairs, 636 pairs in total. We refer to the resulting dataset as ETP-GOLD. To assess the reliability of these annotations, the author of this thesis manually annotated a random subset of 100 edit-turn-pairs contained in ETP-GOLD as corresponding or non-corresponding. The inter-rater agreement between ETP-GOLD (majority votes over Mechanical Turk annotations) and the expert annotations on this subset is Cohen's $\kappa = .72$. We consider this agreement high enough to draw further conclusions from the annotations (Krippendorff, 2004).

6.2.3.3 Properties of ETP-GOLD

As shown in figure 6.2, more than 50% of all corresponding edit-turn-pairs in ETP-GOLD occur within a time span of less than one hour. In our 24 hours search space, the probability to find a corresponding edit-turn-pair drops steeply for time spans of more than six hours. We therefore expect to cover the vast majority of corresponding edit-turn-pairs within a search space of 24 hours and keep this limitation for the remaining part of this study.

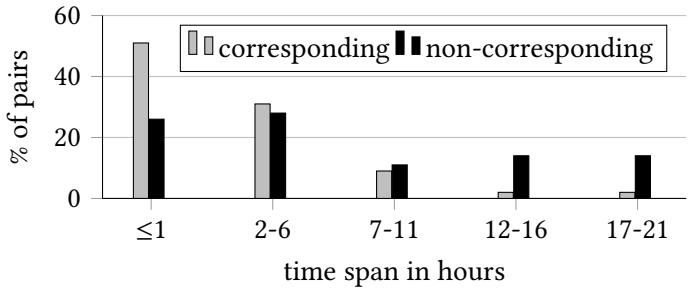


Figure 6.2: Percentage of (non-)corresponding edit-turn-pairs for various time intervals in ETP-gold.

	Number	Same author	Time Span (avg.)	Edit First
Corresponding	128	59%	204 min.	85%
Non-Corresponding	508	46%	482 min.	57%
All	636	49%	426 min.	63%

Table 6.1: Basic properties of corresponding and non-corresponding edit-turn-pairs in ETP-GOLD. The last column indicates the percentage of edit-turn-pairs in which the edit occurred before the turn.

Table 6.1 shows the number of corresponding and non-corresponding pairs in ETP-gold, along with the percentage of edits and turns in a pair which were created by the same user, the average time span between turn and edit (as shown in more detail in figure 6.2) and the percentage of edits made before the corresponding turn was added. The overall percentage of corresponding and non-corresponding pairs which have the same user is surprisingly high, particularely when considering that less than half of all discussants ever edit the article itself (cf. section 6.2.1). The higher overlap is likely to be caused by our restriction of the time span between edit and turns to 24 hours.

Obviously, this is a fairly small dataset which does not cover a representative sample of articles from the English Wikipedia. However, given the high price for a new corresponding edit-turn-pair (due to the high class imbalance in random data, cf. section 6.2.1), we consider it as a useful point of departure for research on edit-turn-pairs in Wikipedia.

6.3 Automatic Classification of Wikipedia Edit-Turn-Pairs

We used DKPro TC (cf. appendix B) to carry outin machine learning experiments on ETP-GOLD. We model edit-turn-pairs as document pairs, i.e. one edit and one turn are processed as pairs inside a pipeline. Each pair corresponds to a single training/test instance in the experiments. For each edit, we stored both the edited paragraph and its context from the old revision as well as the edited paragraph and context from the new revision. We used

Apache OpenNLP for the segmentation of edit and turn text.[92] Training and testing the classifiers has been carried out with the help of the Weka Data Mining Software (Hall et al., 2009). We used the Sweble parser (Dohrn and Riehle, 2011) to parse Wiki markup.

6.3.1 Proposed Feature Set

We divided our feature set in three categories determined by the type of information they reveal about edits, turns, or both. The *edit text* is composed of any inserted, deleted or relocated text from both the old and the new revision. The *edit context* includes the edited paragraph and one preceding and one following paragraph. The *turn text* includes the entire text from the turn.

Similarity between Turn and Edit Text We propose a number of features which are purely based on the similarity between the text of the turn and the edited text or context. We use several string-based similarity measures to compare the edit and turn texts. The measures are calculated between i) the plain edit text and the turn text, ii) the edit text after any wiki markup has been removed and the turn text, iii) the plain edit context and the turn text, and iv) the edit context without Wiki markup and the turn text. Cosine similarity was applied on binary weighted term vectors (L^2 norm). The word n-gram measure (Lyon et al., 2004) calculates a Jaccard similarity coefficient on trigrams. table 6.2a lists all edit and turn text based features.

Features Using Metadata of Edit and Turn An important piece of metadata for an edit is the comment of the revision it belongs to. For turns, we use the name of the topic in which the turn is located. These features extract information beyond the scope of individual edit-turn-pairs, as they refer to entities on a higher level than the edit (the revision) and the turn (the topic). This extension might introduce noise, e.g. when a revision consists of various edits and the comment talks about something different than what is actually performed in the edit of a particular edit-turn-pair. However, we expect the information extracted by these features to be more important than the potential noise they introduce. All features based on metadata are explained in table 6.2b.

Features Using Either Edit or Turn A number of our features are based on the edit or the turn alone and do not take into account the pair itself. Their purpose is to capture properties of the edit or turn, which qualify or disqualify it as a suitable part for a corresponding edit-turn-pair. The full list of those features can be found in table 6.2c. Turn n-grams are represented as binary features indicating the presence or absence of a certain n-gram.

[92] `http://opennlp.apache.org`, accessed May 25, 2015

Feature	Explanation
CosineSim-Edit-Turn	Cosine similarity between the edit text and the turn text
LCS-Edit-Turn	Longest common subsequence between the edit text and the turn text
N-Gram-Distance-Edit-Turn	Word n-gram distance between the edit text and the turn text

(a) Features based on edit and turn text.

Feature	Explanation
User-Is-Same	Whether the name of the edit user and the turn user are equal
Time-Distance	Absolute time difference between the turn and the edit
Pair-Sequence	Whether the edit or the turn occurred first
CosineSim-Edit-Comment-Turn	Cosine similarity between the turn text and the edit comment
LCS-Edit-Comment-Turn	Longest common subsequence between the turn text and the edit comment
N-Gram-Edit-Comment-Turn	Word n-gram distance between the turn text and the edit comment
CosineSim-Edit-Comment-Turn-Topic	Cosine similarity between the name of the topic and the edit comment
LCS-Edit-Comment-Turn-Topic	Longest common subsequence between the name of the topic and the edit comment
N-Gram-Edit-Comment-Turn-Topic	Word n-gram distance between the name of the topic and the edit comment

(b) Features based on metadata of the edit or the turn.

Feature	Explanation
Simple-Edit-Type	Whether the edit is an insertion, deletion, modification or relocation
Edit-Length	Length of the edit text
Turn-Length	Length of the turn text
Turn-N-Gram	N-grams extracted from the turn text

(c) Features based on specific properties of either the edit or the turn.

Table 6.2: Features for edit-turn-pair classification.

6.3.2 Experiments on ETP-gold

We treat the automatic classification of edit-turn-pairs as a binary classification problem. Given the small size of ETP-GOLD, we did not assign a fixed training/test split to the data.

	Baseline	R. Forest	SVM
Accuracy	.799 ±.031	**.866** ±.026†	.858 ±.027†
$F1_{mac.}$	NaN	**.789** ±.032	.763 ±.033
$Precision_{mac.}$	NaN	**.794** ±.031	.791 ±.032
$Recall_{mac.}$	.500 ±.039	**.785** ±.032†	.736 ±.034†
$F1_{non\text{-}corr.}$	.888 ±.025	**.917** ±.021	.914 ±.022
$F1_{corr.}$	NaN	**.661** ±.037	.602 ±.038

(a) Classification results from a 10-fold cross-validation experiment on ETP-GOLD with 95% confidence intervals compared to the majority class baseline. Best values are highlighted; non-overlapping intervals w.r.t. the majority baseline are marked by †.

		pred.	
		corr.	non-corr.
act.	corr.	**83**	40
	non-corr.	45	**468**

(b) Confusion matrix for the Random Forest classifier.

For the same reason, we did not further divide the data into training/test and development data. Rather, hyperparameters were optimized using grid-search over multiple cross-validation experiments, aiming to maximize accuracy. To deal with the class imbalance problem, we applied cost-sensitive classification. In correspondence with the distribution of class sizes in the training data, the cost for false negatives was set to 4, and for false positives to 1. A reduction of the feature set as judged by a χ^2 ranker improved the results for both Random Forest as well as the SVM, so we limited our feature set to the 100 best features.

In a 10-fold cross-validation experiment, we tested a Random Forest classifier (Breiman, 2001) and an SVM (Platt, 1998) with polynomial kernel. Previous work (Ferschke et al., 2012a; Bronner and Monz, 2012) has shown that these algorithms work well for edit and turn classification. As baseline, we defined a majority class classifier, which labels all edit-turn-pairs as non-corresponding.

Table 6.3a summarizes the results from the cross-validation experiment. Due to the high class imbalance in the data, the majority class baseline sets a challenging accuracy score of .80. With an overall macro-averaged F1 of .79, Random Forest yielded the best results, both with respect to precision as well as recall. The low F1 on corresponding pairs is likely due to the small number of training examples. As expected, cost-sensitive classification mainly improves F1 score on corresponding edit-turn-pairs, but it also improves the overall classification performance. More detailed results can be found in the confusion matrix in table 6.3b.

6.3.2.1 Error Analysis

To understand the mistakes of the classifier, we manually assessed error patterns within the model of the Random Forest classifier. Some of the false positives (i.e. non-corresponding pairs classified as corresponding) were caused by pairs where the revision (as judged by its comment or the edit context) is related to the turn text, however the specific edit in this pair is not. This might happen, when somebody corrects a spelling error in a paragraph that is heavily disputed on the discussion page. Another source of errors is a high textual overlap between edit and turn text with a small but important difference such as the orientation of an action: "I have removed X" (turn text) and the edit inserts X. Among the false negatives, we found errors caused by a missing direct textual overlap between edit and turn text. In these cases, the correspondence was indicated only (if at all) by some relationship between turn text and edit comment. For example, a turn might criticize the lack of information in the personal life section of an article about a movie actor. A corresponding edit could insert a sentence about a woman, where the comment explains that this woman is the movie actor's aunt. A really small textual overlap as compared to the entire turn or edit text length can also be problem.

6.3.2.2 Feature selection

We used the χ^2 measure to detect the most important features for our classifier. CosineSim-Turn-Edit-Comment appears to be most important, followed by Edit-Length, CosineSim-Edit-Turn, Turn-Length, LCS-Edit-Context-Turn-Topic and LCS-Edit-Comment-Turn. Among the most important turn unigrams we found *us*, *improved*, *further*, *needs* and *areas*. The fact that all three categories proposed in section 6.3.1 are covered by the most relevant features proves that our proposed set of features is suitable for the given task.

6.4 Edit-Turn-Pairs Across Wikipedia Articles

In this section, we want to demonstrate an application of the model described in section 6.3. As discussed in chapter 3.1, successful coordination via indirect user interaction is crucial in online mass collaboration. As shown in section 6.2.1, most authors in Wikipedia's CW process only participate via indirect interaction (i.e. editing the article itself), but do not contribute to coordination via direct interaction. From the users who are active in coordination on discussion pages, less than half choose to contribute to the CW process via indirect interaction. Their contribution to the CW process only becomes visible in the article itself if other users address these contributions via indirect interaction, effectively creating corresponding edit-turn-pairs. Both of these observations make us conclude that increasing the number of corresponding edit-turn-pairs is a promising way to strengthen coordination and its effects in the CW process and thus, improve the quality of the resulting

texts. To empirically analyze this assumption, we measure the influence of corresponding edit-turn-pairs on the development of articles. To this end, we will compare the percentage of corresponding edit-turn-pairs across different articles.

6.4.1 Edit-Turn-Pairs in Articles Suffering Quality Flaws

We chose to assess articles which are known to suffer from a certain quality flaw (Anderka et al., 2012), assuming that such flaws will be addressed in the discussion of the article. Quality flaws are marked by templates in the source of an article and are usually displayed in a tag box above the article text. We picked two very common quality flaws, the *Unreferenced* flaw and the *Refimprove* flaw.[93] *Unreferenced* is the most common flaw in the English Wikipedia (Anderka et al., 2012) and indicates the lack of any kind of reference or source in the article content. *Refimprove* is used to demonstrate that an article needs additional references. Both flaws have article scope, i.e. they are placed once per article and refer to the entire content (as opposed to inline flaws, which refer to text fragments only).

We selected all pages from the English Wikipedia Main namespace (i.e. encyclopedic articles) with an associated discussion page that had a minimum length of 10,000 characters. This limitation is intended to exclude articles with little or no discussion as they would not be suitable for our analysis. The articles are extracted from the dump of April 2011. As we are going to use the entire revision history of these articles, this sample will include a bit more than the first ten years of Wikipedia's history. From these pages, we picked a random subset of 100 articles, containing 50 articles marked with the *Unreferenced* flaw, and 50 articles marked with the *Refimprove* flaw.

We collected all edit-turn-pairs from the revision history of the 100 articles and their discussion pages, including archived discussion pages. Except for limiting the time span between edits and turns to 86,000 seconds (approx. 24 hours), as we did during the creation of ETP-GOLD, we applied no filtering. From the flawed articles, we included all revisions from their history and not just those revisions which actually contain the flaw marker (i.e., template) in their source. The reason for this is that we wanted to analyze the entire development of the articles. The pairs were classified using a Random Forest classifier trained on ETP-GOLD, as this classifier has shown the best performance in detecting corresponding edit-turn-pairs.

6.4.2 Implications from Classifying Edit-Turn-Pairs Across Articles

The results in table 6.4 show that the *Unreferenced* flaw generates a slightly higher average percentage of corresponding edit-turn-pairs as compared to the *Refimprove* flaw. This is an interesting finding, since the *Unreferenced* flaw is used mainly in young articles or stubs, as

[93] See `http://en.wikipedia.org/wiki/Template:Unreferenced`, accessed May 25, 2015 and `http://en.wikipedia.org/wiki/Template:Refimprove`, accessed May 25, 2015.

Flaw	Revisions	Pairs	Corresponding Pairs					
			Micro-Average	Macro-Average	Macro-Median	Max.	Min.	St.dev.
Unreferenced	17,504	18,001	4.54%	9.34%	3.31%	100%	0	0,17
Refimprove	42,491	96,703	2.28%	6.55%	4.38%	47%	0	0,08

Table 6.4: Overall number of revisions and edit-turn-pairs for our sample of *Refimprove* and *Unreferenced* flawed articles. Micro- and macro-averaged percentage of corresponding pairs, along with median, maximum, minimum, and standard deviation across all articles.

shown by the low average number of revisions. We conclude that certain quality problems (in this case, the *Unreferenced* flaw), are more likely to generate a knowledge flow from the discussion page to article content than others. However, as the experiment described in this section was carried out on a random subsample of flawed Wikipedia articles, the results could be distorted by outliers (this is reinforced by the higher percentage of edit-turn-pairs in *Refimprove* when measured as macro median). Therefore, far-ranging conclusions need to be considered carefully.

The macro-averaged numbers in table 6.4 weight each article equally. An average article in our sample has below 5% corresponding edit-turn-pairs, as shown by the median score. The standard deviation of this percentage across all articles is high, which indicates considerable variance between the articles. To better understand the reasons for the difference between individual articles, further investigation is necessary. One way to analyze the influence of corresponding pairs on the development of articles might be to take the temporal dimension into account, i.e. to analyze the distribution of corresponding edit-turn-pairs for different time periods in the development of an article.

Figure 6.3 shows two corresponding edit-turn-pairs discovered in our experiments, one for *Refimprove* and one for *Unreferenced*. In both pairs, a user adds one or more references to the article. The additions are both reported in the respective turn.

6.5 Implications beyond Wikipedia

In this chapter, we have presented a framework for the detection of correspondence between edits and turns in Wikipedia. We tested the framework on a range of flawed articles, showing how edit-turn-pairs can be used to analyze the development of articles. The detection of correspondence between edits and turns is also relevant to applications beyond Wikipedia. Commits (i.e., edits) to source code repositories might correspond to issues discussed on the respective mailing list. Automatically analyzing such correspondence can help the developers to prepare release summaries or track bugs. We also see applications in business. Many companies use wikis to store internal information and documentation

Hi and peace to everybody. Im a shia, and added some more text to the article. I expect i to be un-appreciated by sunnis, even though i only quote sunni sources. Mostly since i bring up events that sunnis rather would like to forget. It would be intresting to get some feedback from sunnis but also from non-muslims.	Sunni References: **[http://www.shianews.com/snObjects/documents/[...].txt | Musharriful Mahbubin by Hazrat Khuwaja Mehboob Qasim Chishti Mushrrafi Qadri (RA) Pages 216-218]**

(a) Corresponding edit-turn-pair from an article marked with *Refimprove*.

Wikipedia's Tigon and Liger articles contain info on Lady Kali, a ti-liger. However, the articles both initially merely repeated the claims made by her owners that she was part of a "research programme" and "behavioural" study. [...] What there IS, and I've provided a link to a copy of it, is a 2004 report from the local paper [...] Nonetheless, I think SOME reference to Lady Kali is worth leaving in the Tigon and Liger articles. I corrected the info and provided links to the Lady Kali webpage [...]	**~~A behavioural research programme in the USA has a female ti-liger called Lady Kali; at 2 years old she weighed 400 lb (180 kg).~~ A private exhibitor in the USA has a female ti-liger called Lady Kali, and displays a Class C USDA license [http://www.aphis.usda.gov/ac/publications/AWR/9cfr1.1.txt] valid through most of 2006 at their website. [http://www.domesticpanthera.com/legalstuff.htm] [...] Kali has reportedly been exhibited as a roadside zoo attraction in various North Carolina locations from at least 1999 through 2004, and is apparently not part of any active breeding program. [http://cwapc.org/news/NewsDescription.asp?FileName=news_-20041214.html]**

(b) Corresponding edit-turn-pair from an article marked with *Unreferenced*.

Figure 6.3: Corresponding edit-turn-pairs discovered with the help of our classification model. Turns are displayed left, and edits right. The examples have been slightly adapted to increase readability. Formatting conventions: inserted text is boldfaced, deleted text is boldface and crossed out.

(Arazy et al., 2009). An alignment between edits in the company wiki and issues discussed in email conversations, on mailing lists, or other forums, can be helpful to track the flow or generation of knowledge within the company. This information can be useful to improve communication and knowledge sharing.

Edit-turn-pairs connect direct and indirect user interaction and demonstrate the effect of these modes of interaction with each other. It is known that coordination through direct interaction is crucial for successful CW projects (Allen et al., 1987; Erkens et al., 2005). However, direct interaction by itself does not improve the quality of a document, unless it leads to indirect interaction. Edit-turn-pairs enable a much more fine-grained analysis of the positive effect of direct interaction. Corresponding edit-turn-pairs can be detected individually for each document and thus potentially reveal successful and less successful coordination scenarios.

Based on the assumption that a higher number of corresponding edit-turn-pairs positively influences the CW process and potentially yields better document quality, authors should be encouraged to increase the percentage of corresponding edit-turn-pairs. One way to do this could be to identify those turns that are not part of any corresponding edit-turn-pair, as they have a higher potential to address issues that are still not implemented in the document. These turns could be shown to authors with a recommendation to take

action, if necessary. Furthermore, interested authors could also get recommendations to edit documents with a particularly low percentage of corresponding edit-turn-pairs.

6.6 Conclusion

In this chapter, we have presented a way to detect corresponding edit-turn-pairs in Wikipedia articles. We have addressed three questions. First, we wanted to understand the nature of correspondence between Wikipedia article edits and discussion page turns (first research question). To this end, we have shown that, in a corresponding edit-turn-pair, the turn contains an explicit performative and the edit corresponds to this performative. We have defined four types of correspondence, based on the performative expressed in the turn. Second, we have asked for the distinctive properties of corresponding edit-turn-pairs and how these can be detected automatically (second research question). To find out about this, we annotated a corpus of corresponding and non-corresponding edit-turn-pairs with the help of crowdsourcing. We have identified a number of distinguished properties of corresponding edit-turn-pairs as compared to non-corresponding pairs. For example, an edit and a turn are much more likely to correspond, if they occur within a window of less than six hours. Using the annotated corpus, we train a model based on textual similarity and metadata of edits and turns, such as the edit comment and the turn topic. We show that these features can effectively distinguish between corresponding and non-corresponding edit-turn-pairs. In a cross-validation experiment, we achieve an accuracy of .87 on this model. Third and finally, with respect to the overall research question of this thesis, we wanted to know what edit-turn-pairs tell us about the relationship between direct and indirect user interaction in online mass collaboration (third research question). In a corpus of flawed English Wikipedia articles, we have shown that the percentage of corresponding edit-turn-pairs is below 5% on average and varies considerably across different articles. Based on our observations regarding the participation of users in the CW process in Wikipedia, we suggested that increasing the percentage of corresponding edit-turn-pairs might be a promising way to improve article quality. This could be achieved by pointing users to articles with a particularly low percentage of corresponding edit-turn-pairs. A more sophisticated approach would highlight those turns on a discussion page which contain explicit performatives other than reports of actions (cf. section 6.1.2) and which have not been addressed in the article as part of a corresponding edit-turn-pair.

We see three directions for future research. First, to improve the predictive power of the model presented in section 6.3, a larger number of edit-turn-pairs, in particular, corresponding pairs, should be annotated. The existing data can be used to bootstrap the annotation of further data. In particular, the corresponding edit-turn-pairs can be used to detect pairs in unseen data which are highly likely to be corresponding, so that with less effort a balanced corpus can be created. Second, although we have taken a step in this direction, the influence

of edit-turn-pair correspondence on the success of a CW project is not fully understood. We have analyzed two sets of flawed Wikipedia articles, with more than 100,000 edit-turn-pairs. However, as we have only experimented with flawed articles, it remains unclear to which degree this set of articles can be considered representative with respect to its distribution of edit-turn-pairs. Further research should be conducted, both on a representative set of Wikipedia articles (cf. section 5.2), and on a set of high-quality articles. Third, rather than analyzing the mere distribution of corresponding and non-corresponding edit-turn-pairs, the kind of correspondence should be evaluated. We based our definition of edit-turn-pair correspondence on four types of performatives, but so far, we did not make use of these types when detecting corresponding edit-turn-pairs. Knowledge about the distribution of kinds of edit-turn-pair correspondence in a large corpus would help to get further insights into the nature of edit-turn-pairs.

Chapter 7

Conclusion

The development of documents on the web, where potentially thousands of authors participate in the writing process, is a complex area of study. In this work, we have presented several NLP-based approaches to analyze CW in online mass collaboration by the example of Wikipedia. Within this analysis, our main focus was on indirect user interaction, which can be studied with the help of document revision histories. Revision histories record all past versions of a document including meta data such as their author or time stamp. We also studied direct user interaction, which takes place when authors are engaged in direct communication. In particular, we analyzed how direct interaction influences indirect interaction, and vice versa. This chapter summarizes the main conclusions and findings presented in the course of this work. Furthermore, we show the impact of our research for the related theoretical frameworks and give broader practical recommendations with respect to online mass collaboration. Finally, we discuss open issues.

7.1 Summary of Main Contributions and Findings

While we presented previous research and theoretical considerations for the writing process and online mass collaboration in chapters 2 and 3, chapters 4 through 6 contain our main contributions.

We introduced **writing as a process** and **revision as a part of the writing process** in chapter 2. Furthermore, we outlined the history and recent developments in **CW** with a particular focus on computer-supported CW. Related to these concepts, we discussed two major shifts in writing research and composition teaching. In the 1970s, teachers and scholars shifted their focus from the product of writing to the process of writing, resulting in a substantial body of research on revision. A couple of years later, technological advance gave raise to tools which support CW. When multiple authors work on the same document, the challenges of the writing process become quickly visible. Tools designed to support writers

in collaborative environments started to be used in the classroom as well as in research and industry. The success of CW tools and CW itself was disputed in the beginning. In the following years, a growing number of studies highlighted the positive influence of CW on the quality of documents, provided that the writing process is well-coordinated.

In chapter 3, we suggested to divide the study of the **CW process in open online collaboration** into direct and indirect user interaction. Coordination is an important factor for successful CW projects. Under the concept of user interaction we gathered all CW activities in online mass collaboration. Direct interaction happens when authors communicate (e.g. on a discussion platform) among each other during the CW process. Indirect interaction does not involve oral or written communication between authors, but is produced when more than one author edits a document or part of a document previously written by another author. Based on wikis, a popular online CW tool, we illustrated and discussed these concepts in more depth. In particular, we based our analysis on the online encyclopedia Wikipedia, one of the most successful online CW projects. We outlined the structure of the Wikipedia community and introduced the technologies which enable direct and indirect interaction in Wikipedia and other wikis. Finally, we outlined a few aspects of quality in Wikipedia, in particular those based on the Wikipedia-internal quality assessment project, which labels articles according to certain quality criteria.

Chapter 4 contains our first major contribution and lays the foundation for our analysis of **indirect user interaction in Wikipedia**. Chapters 5 and 6 are based on the results of chapter 4. To better understand the particularities of the writing process in Wikipedia, we presented an algorithm for the segmentation of consecutive revision pairs into finer-grained edits. Based on the concept of edits, we created a novel taxonomy for the classification of Wikipedia edits into categories such as spelling correction or vandalism. We demonstrated the suitability of the novel taxonomy in two annotation studies on corpora of English and German Wikipedia edits. Furthermore, we present a novel machine learning model which we trained and tested on corpora of (a) English Wikipedia edits, (b) German Wikipedia edits, and (c) a third-party corpus of English Wikipedia revisions. Our models achieved a micro-averaged F1 score of .62 on English edits, .55 on German edits and .78 on English revisions. The corpora deviate considerably in size and granularity. We concluded chapter 4 with an analysis of the relationship between indirect user interaction and article quality in Wikipedia. We showed that the information content in high-quality articles tends to become more stable once they have been promoted. Furthermore, we found that high-quality articles are more homogeneous with respect to their collaboration patterns as compared to random articles.

In chapter 5, we switched to a higher-level perspective, away from the fine-granular analysis of Wikipedia edits, and turned to **activity-based roles of users**. We introduced the concept of emergent roles. These roles are assigned to users based on their edit behavior. To this end, we created a sample of representative articles from the English Wikipedia and

classified the entire revision history of these articles with the approach presented in chapter 4. We then created activity profiles over all revision categories performed by each user. The profiles were clustered into seven roles with the help of an unsupervised machine learning approach. These clusters mark our emergent roles and include clusters that have not been detected before, e.g. Content-Shapers or Vandals. By calculating the stability of our clustering solution, we showed that our proposed emergent roles describe the data well. Furthermore, we found that the nature of the roles is quite similar across different models of the input space and also across time. Finally, we tested whether users change emergent roles over time, finding that more than half of the users in our sample played different roles in the same article before and after the year 2007. We concluded that this behavior is partly related to changes in Wikipedia's community structure, which around this time substantially increased its focus on content quality and consistency.

The **relationship between direct and indirect user interaction** has been our focus in chapter 6. We introduced the concept of **edit-turn-pairs**, which connect edits from a Wikipedia article revision history and turns from the discussion pages of the same article. Based on whether the turn expresses a performative and the type of this performative, we defined four kinds of correspondence between edits and turns. We labeled edit-turn-pairs as corresponding if the edit corresponds to a performative expressed by the turn, and as non-corresponding otherwise. Given that the vast majority of edit-turn-pairs in Wikipedia is non-corresponding, we presented a step-wise approach to create a small corpus of corresponding and non-corresponding edit-turn-pairs, showing how to overcome the class imbalance problem. In a crowdsourcing annotation study, we labeled a corpus of 750 edit-turn-pairs according to our definition. The resulting gold standard data has reasonable agreement with expert annotations. Based on textual similarity and metadata of edits and turns, e.g. the edit comment, we presented a machine learning model which is able to automatically detect corresponding edit-turn-pairs. The model distinguished corresponding and non-corresponding edit-turn-pairs with .87 accuracy in a cross-validation experiment. We concluded the chapter by showing that the percentage of corresponding edit-turn-pairs is on average below 5% for a corpus of Wikipedia articles suffering certain quality flaws. Increasing the number of corresponding edit-turn-pairs, e.g. by directing users to articles with a high number of unaddressed issues on the discussion page, could be a promising approach to improve article quality.

7.2 Theoretical Impact and Implications

Based on the main contributions of this thesis as outlined in the previous section, we will now summarize the resulting implications for the related theoretical frameworks and for further research. Although the basic principles of traditional research on the writing process also apply to CW in online mass collaboration, the open and often large-scale setting in

online scenarios implies several particularities. With respect to the edit categories in online CW, it should be considered that not all authors contribute with good intentions, but that vandalism can become a serious problem. We accounted for this fact by creating an edit taxonomy which captures the entire range of editing activity in the online encyclopedia Wikipedia. Based on the full range of edit categories, we analyzed different Wikipedia articles with respect to edit patterns and found that articles typically enter a phase of higher stability after they reached a certain quality level. This phase is characterized by an increased number of surface edits (e.g. copy-editing) and a lower number of text-base edits (e.g. new information). Although these phases cannot be properly mapped onto traditional CW activities proposed in previous work (Galegher and Kraut, 1994; Lowry et al., 2004), they suggest that there are phases emphasizing different activities in online mass collaboration as well.

We identified seven activity-based roles which were determined based on the individual contributions of each user. Again here, the traditional model of activity-based roles in CW does not map onto the range of possibilities in online mass collaboration. For example, we identified roles such as vandals, who destroy value during the CW process and watchdogs, who try to restore the value. In contrast to previous work, we also identified roles which are associated with shaping tasks, i.e. formatting and layout editing.

As a unique feature to CW projects in online settings, we can analyze the coordination processes "behind the scenes" via direct interaction. This has not been possible, at least not to the same degree of detail, in many traditional CW projects as coordination tasks were typically not recorded by any means. While most CW theories are aware of the importance of coordination via direct user interaction, none offers an explanation on how to integrate direct user interaction with the actual editing process process. We therefore created the concept of edit-turn-pairs which connect coordination, conflict resolution, and ultimately, knowledge building, to individual edit actions.

Online Mass Collaboration and Interdisciplinary Research While we could not cover all aspects of the CW process in online mass collaboration with the same level of detail (e.g. we analyze direct user interaction only in the context of edit-turn-pairs) and while there are certainly other ways to discover the writing process in online mass collaboration (e.g. visual exploration, cf. Viégas et al. (2004); Flöck et al. (2015); Borra et al. (2015)), we are confident that our framework will be useful to subsequent research. Given that both the writing process as well as the resource Wikipedia are subject of study in several disciplines including education, sociology, information science, business and computer science, our findings have the particular potential to foster interdisciplinary research.

We presented a detailed framework for the analysis of indirect interaction in online mass collaboration, based on the notion of edits. Whereas edits in Wikipedia have been studied before (Liu and Ram, 2011; Bronner and Monz, 2012), our research goes beyond

existing work by applying a detailed edit category taxonomy, based on the fundamental distinction between text-base and surface changes (Faigley and Witte, 1981), to Wikipedia revisions. In an analysis of the quality of Wikipedia articles, we made advantage of this distinction and found that the writing process in Wikipedia articles changed after they had reached a certain quality stage. While stages of the writing process have been studied with frequency in educational settings (Hayes, 2004; Myhill and Jones, 2007), we are not aware of any studies that have shown this effect in the context of online mass collaboration. Furthermore, we applied a corpus of German and English Wikipedia edits, annotated with the categories from our taxonomy, to train a model for automatic edit classification. The resulting model can be used to annotate large numbers of Wikipedia revisions fully automatically. Due to the breadth of categories in our taxonomy, a wide range of applications benefits from this outcome. Beyond the basic distinction between meaning-altering changes and meaning-preserving changes, a fine-granular analysis of certain edit categories such as spelling corrections, paraphrases, or additions of information is possible. Such analysis should be particularly interesting for applications in the domains of education or psycholinguistics. For example, research on the relationship between personality and revision (Jensen and DiTiberio, 1984) might discover new insights from the large-scale analysis that is enabled by our tools.

It is widely accepted that coordination is a crucial factor in online mass collaboration systems. We thus connected our edit framework, which quantitatively and qualitatively judges indirect user interaction, to the main coordination space of Wikipedia, namely its discussion pages. Using the concept of edit-turn-pairs, we made implicit links between direct and indirect user interaction in online CW visible. For the first time, edit-turn-pairs enable to measure the degree of correspondence between article edits and coordination efforts on the discussion page. Our experiments yield initial insight about the impact of direct correspondence between edits and discussion turns, showing that the percentage of discussion page issues which are addressed in the respective article is typically below 5% in a corpus of flawed articles in the English Wikipedia. While this quantification by itself might be interesting to some, our framework lays the foundation to address further issues both in the organizational and sociological domains, as well as for writing research. Understanding the knowledge flow in large online collaboration projects can help managers to channel resources into the right direction. Knowing which issues have or have not been addressed in a document should also result in a better understanding of more and less successful coordination scenarios and thus in a better understanding of the online CW process.

The concept of emergent roles presented in chapter 5 of this thesis opens new directions for research on social networks and user roles (Forestier et al., 2012; McCallum et al., 2007). We have extended the notion of activity-based roles in online CW projects proposed by previous work (Welser et al., 2011; Liu and Ram, 2011) with a more detailed conceptualization in which each kind of edit performed by an author is considered and evaluated. As an

extension to previous work, we have analyzed the stability of roles over time, showing that although the roles themselves are quite stable, a substantial number of authors change their role over time. In addition to the findings about the nature and impact of emergent roles, our research paves the way for further research in the organizational, educational, and sociological domains. For all of these, it seems particularly promising to analyze the impact of emergent roles on document quality, e.g. by analyzing Wikipedia co-author networks based on emergent roles for articles with known quality ratings.

7.3 Practical Recommendations

Beyond the implications for online collaboration systems other than Wikipedia which we gave for each of the chapters 4 through 6, here, we will give some more generic recommendations for applications related to online mass CW.

On Community Decline in Open Online Collaboration Systems We have addressed the issue of community decline due to undesired effects of measures to improve quality and consistency in chapter 5. Halfaker et al. (2012) argue that the reinforced implementation of new policies, starting in the English Wikipedia in the year 2007, substantially lowered newcomer retention. This effect is well-known for Wikipedia, but it has also been observed in other online communities, e.g. Question-and-Answer platforms such as Stack Exchange. While over-regulation and newcomer deterrence may have a strong influence on community decline in Wikipedia, there are certainly other reasons which have contributed likewise.[94] One of the further reasons for the slowing community growth might be the high coverage of topics in larger Wikipedias, such as the English Wikipedia. Compared to the early phase of the online encyclopedia, it has certainly become more difficult to find topics of general interest (cf. the notability criteria for new articles in Wikipedia[95]) which are not or incompletely covered. In chapter 4, we found preliminary evidence that mature articles attract more edits to the surface (meaning-preserving) as compared to young articles. Assuming that new users are more likely to continue editing Wikipedia, when they are also able to perform text-base edits (with the intent to add knowledge to the encyclopedia), rather than "polishing" existing content, we might conclude that the growing number of mature articles is, to a certain degree, responsible for a lower newcomer retention. Our findings from chapter 5 about the activity-based role changes of authors over time tend to confirm that the need for polishing articles has grown, as many users have moved from generic roles such as All-round Contributor to specialized roles such as Layout-Shapers. Certainly, based

[94] See also this Slate article from late 2014: http://www.slate.com/articles/technology/bitwise/2014/12/wikipedia_editing_disputes_the_crowdsourced_encyclopedia_has_become_a_rancorous.single.html, accessed May 25, 2015

[95] http://en.wikipedia.org/wiki/Wikipedia:Notability, accessed May 25, 2015

on the existing evidence, these conclusions are mere assumptions, but they clearly reflect the benefits of a better understanding of the CW process in online mass collaboration.

Educational Applications The use of automatized tools to support students in their writings is highly debated.[96] In North America, automated essay scoring based on models created with the help of machine learning is a widely used approach to reduce costs and increase comparability of gradings (Shermis and Burstein, 2003). The fairness and accuracy of this technique is highly disputed.[97] However, the amount of data created in educational environments is constantly growing. Massive open online courses (MOOCs) are drawing really large numbers of students from all over the world. When tens of thousands of students participate, the load of works to be corrected by the instructors of MOOCs often exceeds any reasonable limit. Therefore, the need to make use of NLP-based technology in teaching is a highly relevant issue. However, given the doubts about automated (essay) grading, NLP-supported applications in teaching should be used to improve, rather than grade, the writing skills of students.[98] While the amount of data suitable for analyzing the writing process produced by single authors is rather small, open CW systems like Wikipedia produce a huge amount of data which can be freely accessed. In this work, we have presented various ways to analyze the output of open CW systems. Detailed knowledge about the writing process (e.g. categorizing revision, correlation between document quality and revision) helps to develop applications which support writers in their writing, both for single-author writing, as well as for CW. When used for this purpose, our contributions can be very useful to educational applications.

Applications in Industry Techniques to support the CW process can also be of benefit to companies. Large companies operating globally, but also more and more small- and medium-size companies, store and extend their knowledge databases in wikis or wiki-like platforms (Tapscott and Williams, 2008), where employees exchange ideas, persist frequent problem solutions or describe workflows and best practices in a collaborative manner. The revising process of technical writers has been investigated to some extent (Rosner, 1992), however, the technologies presented in this work enable new ways of analyzing revision in industry. In this regard, we see two promising applications. First, the awareness of preferred editing behavior and activity-based roles in collaborative systems will help administrators to channel and guide time and effort put into these resources. For example, authors

[96] See e.g. `https://www.edsurge.com/n/2014-09-22-where-does-automated-essay-scoring-belong-in-k-12-education`, accessed May 25, 2015.

[97] See e.g. `http://www.nytimes.com/2012/04/23/education/robo-readers-used-to-grade-test-essays.html`, accessed May 25, 2015.

[98] This has been proposed before. See e.g. this blog entry by Elijah Mayfield, the founder of LightSIDE Labs, from 2013: `http://mfeldstein.com/si-ways-the-edx-announcement-gets-automated-essay-grading-wrong`, accessed May 25, 2015.

who constantly add new content, but rarely structure existing content, might be advised to put more emphasis onto the organization of knowledge, to support their colleagues in actually finding and using the newly added content. Second, the techniques described in chapter 6 which bundle editing and discussion activity in related resources, can reveal otherwise hidden knowledge in companies. Many companies use wikis, trouble ticket systems, and mailing lists in parallel, so that documentation and knowledge are likely to be distributed across various places. Whenever a CW system such as a wiki and a discussion platform such as a mailing list exist together, the detection of (non-)corresponding edit-turn-pairs seems to be a promising way to collect existing information from different sources. Like this, it becomes possible to analyze the flow of knowledge within the company, which in turn can be used to improve communication and knowledge sharing.

7.4 Open Issues and Limitations

Beyond the specific open issues addressed in each of the chapters 4 through 6, here, we will give a summary of open issues along with some higher-level future work.

Edit and Revision Classification We have extensively discussed and analyzed approaches to segment and classify Wikipedia edits and revisions. We have also applied the former to analyze the relationship between indirect user interaction and article quality in Wikipedia. However, we have not extended our experiments beyond featured and non-featured articles (chapter 4) and articles with quality flaws (chapter 6). There are several reasons why this might not yield a complete picture. First and foremost, the ratings from the Wikipedia quality assessment project are based on criteria defined and judged upon by members of the Wikipedia community. Such measures do not necessarily correspond to the perception of quality that the readers of an article might have. The social processes during the promotion process of high-quality articles could be investigated in more detail to understand the influence of the authors of an article during its nomination and promotion as featured article. Wikipedia's Article Feedback Tool (Halfaker et al., 2013; Flekova et al., 2014) has been designed to overcome the limitations of the internal quality rating systems by collecting feedback about article quality from readers along several dimensions. The v.5 pilot of the Feedback Tool collected more than 1.5 million feedback messages in three language versions of Wikipedia (English, German, French). A further concern with the Wikipedia-internal quality ratings is their up-to-dateness (Ferschke, 2014). Although it is possible to demote featured articles, the CW process cannot assure that once an article was featured, it will remain in such a good shape and that, if the quality drops over time, it will be demoted.

Activity-based Roles Our analysis of emergent roles in Wikipedia, based on the edit behavior of users, yields several promising options for future work. First, the analysis of emergent roles enables new possibilities to explore Wikipedia article co-author networks (cf. section 3.1.1.3). Wikipedia co-author networks have already been explored by several studies (Brandes et al., 2009; Laniado and Tasso, 2011), however, in these studies, the network nodes typically represent individual authors, and not roles. We already carried out a preliminary exploration of this task and expect novel insights from establishing a network over emergent roles rather than individual authors. Subsequently, an analysis of the relationship between frequent motifs in the co-author network of an article and other commensurable properties of the article, e.g. the article's text quality, could be explored. Second, we suggest to connect emergent roles and edit-turn-pair correspondence, so that the profile of an author would take into account both his or her direct and indirect interaction. Like this, it would be possible to study whether users in certain emergent roles are more likely to address their editing activity in direct interaction than others.

Corresponding Edit-Turn-Pairs With respect to edit-turn-pair correspondence, we have already started to explore the potential of a large-scale analysis across Wikipedia articles. However, we believe that there is room for further investigation. On a high level, the distribution of edit-turn-pairs across different topical categories of articles could be compared. On a lower level, i.e. for individual articles, it would be interesting to compare the distribution of edit-turn-pairs across time. For example, an analysis of the percentage of corresponding edit-turn-pairs in flawed Wikipedia articles before and after a flaw is reported would help to extend our analysis in section 6.4. Finally, increasing the amount of annotated data in ETP-GOLD would probably help to improve the predictive power of the classifier which detects corresponding edit-turn-pairs.

Generic Beyond article and discussion, Wikipedia contains several namespaces (see table 3.2) with user-specific, multimedia, technical, and organizational content. Among these, user pages (see section 3.2.2.1) are a very interesting part of Wikipedia data, which we did not consider in this work. As these pages reveal further details about users, they might well serve as an extension to the data we gathered about editing behavior. In addition to the User namespace, policy pages (mostly located in the Wikipedia namespace) could be used as a source of further information on the CW process in Wikipedia, in particular about those users, that are involved deeper in the organizational and technical aspects of editing.

Another source of data not considered in this work but potentially relevant to CW in Wikipedia is Wikidata (see section 3.2.2.1), Wikimedia's central knowledge platform. The introduction of Wikidata as a centralized edit interface to maintain basic knowledge such as interlanguage links, info box entries and lists, is likely to affect the editing behavior, as certain types of trivial edits (often carried out by bots) to articles might not be neces-

sary anymore in the future. While this feature certainly helps to increase consistency and up-to-dateness of articles, it decouples a part of the CW process from the article at hand. For example, interlanguage links are not managed within the source of a Wikipedia article anymore, so that changes to interlanguage links of an article are not directly visible in the article history, but only in the history of the respective Wikidata object.

For the analysis in chapters 4 and 6, we looked at CW processes in individual Wikipedia articles or compared them across articles. Ransbotham et al. (2012) found that, in the context of an English WikiProject, articles develop not fully independently from each other, but that there are certain dependencies which they measured in a bipartite network of articles and authors. We have addressed this issue in two ways. First, in chapter 5, we established the concept of emergent roles, which are determined over all edits users performed in a sample of representative Wikipedia articles. Second, in chapter 6 we analyzed the flow of knowledge between Wikipedia articles and discussion pages. We have, however, so far not measured the relationship of CW processes between individual articles, based on e.g. the author network. This could be addressed by analyzing bipartite networks of authors and articles in combination with collaboration patterns (cf. section 4.6.2) applied in individual articles.

CW in Wikipedia: Limiting Factors We should mention that, although we have carefully selected the data sources for our explorations, there are certain limits to the conclusions drawn from this study. Due to Wikipedia's encyclopedic nature, the text types under collaboration are quite restricted. The extent to which these factors have an influence on the CW process has not yet been investigated in depth (Fitzgerald, 1987). As explained in section 3.2.2, Wikipedia suffers a systematic bias, in that its authors reflect an unbalanced picture with respect to e.g. gender and age. The particular setting of Wikipedia, both in terms of its community and its implementation, has consequences on the inferences about user interaction in CW which we can draw from it.

The design of the edit and revision category taxonomies, which are the foundation for many of our findings in chapters 4 and 5, sets the range of activities our analysis can account for. Since the design of the taxonomies is based on CW activities in Wikipedia, certain categories which we had to include might not be relevant to the same degree in other online CW projects, and vice versa. As a consequence, the edit classification models trained and tested in chapter 4 cannot be applied to non-wiki-based CW projects without further effort. Despite this limitation, we expect the edit category taxonomies presented in chapter 4 to account for all important editing activities in online mass CW. Another, more important, limiting factor is the distribution of edit activities across CW projects. For example, in collaborative online word processors such as Google Docs, the importance of editing markup or using templates might be different as authors are editing in a WYSIWYG editor rather than the wiki markup language. Furthermore, vandalism only becomes a substantial prob-

lem in large open online projects with high visibility. It is thus to be expected that the importance of individual edit categories such as vandalism or markup modifications varies across online CW projects, implying that some of our findings based on edit distributions, e.g. the composition of emergent roles explained in chapter 5, might not hold across all online CW projects.

We analyzed direct user interaction in Wikipedia mainly through the activity on discussion pages. Here again, the fact that in Wikipedia CW coordination is mainly supported through asynchronous discussion in a dedicated forum has some limiting consequences for our findings in chapter 6. For example, given the distance between a turn on the discussion page and an edit on the article page, detecting corresponding edits and turns becomes a tricky problem. This would be much easier in CW platforms which support local comments on the edited document itself, e.g. Google Docs (cf. figure 3.3). The same is true in CW platforms which offer synchronous discussion through a chat system in parallel to the edited document. Detecting correspondences between edits and turns is likely to be simplified when authors are discussing about what they are *currently* editing. We assume, that this happens less frequently in the static, asynchronous setting of Wikipedia's discussion pages.[99] In chapter 6 we found that less than 5% of all edit-turn-pairs in corpus of flawed Wikipedia articles were corresponding. In the light of the above considerations, it is not clear whether this finding also holds for online CW projects with different support for direct user interaction. For systems with support for real-time chatting or local comments, the percentage of corresponding pairs might be much higher. On the other hand, given that Wikipedia discussions often develop their own dynamics (Viégas et al., 2007a; Ferschke et al., 2012a), much of the knowledge created in a separate discussion space would probably be lost in such systems.

The first step to understand different editing patterns and edit category distributions across online CW projects, including findings based on the latter such as emergent roles, could be to analyze more wiki-based CW communities, e.g. wikis hosted by Wikia.[100] To better understand the influence of the kind of support for direct user interaction, revision data extracted from non-wiki-based CW platforms such as Etherpad or Google Docs should be analyzed as a next step.

7.5 Concluding Remarks

This work is a step towards a composite model of collaborative revision in online communities. Such a composite model of collaborative revision needs to take into account not only

[99] It should be noted that discussion pages are not exclusively intended for the purpose of discussing matters related to the current editing process. However, the recommended usage of discussion pages is to "discuss improvements" in the associated subject page (cf. `http://en.wikipedia.org/wiki/Help:Using_talk_pages`, accessed May 25, 2015), which suggests a usage to that effect.

[100] `http://www.wikia.com`, accessed May 25, 2015.

the revising process "on paper" (i.e. the edit itself) but all the events happening around the revision, in CW especially the interaction between authors. This interaction takes place before, while and after writing (Fitzgerald, 1987). We tried to contribute to the picture by establishing, discussing, and connecting two aspects of collaborative revision in Wikipedia, namely direct and indirect interaction.

Appendix

The appendix contains additional material which was not included in the main part of this thesis due to its technical nature.

A Supervised Machine Learning on Textual Data: Foundations

Many of the experiments carried out as part of this work use machine learning technology. All of them are carried out on textual data, and most are supervised, i.e. we use labeled data to train a classifier.[101] This work has also contributed to the development of an open-source text classification framework with a focus on NLP applications, DKPro TC. In the light of these conditions, we decided to add a very brief introduction into the basic concepts of supervised text classification. Machine learning is a highly complex research area, so that by no means we can give a complete overview of its methods and applications. For a more detailed introduction to machine learning, see e.g. Bishop (2006).

Supervised text classification has become a popular solution to many problems brought forward by the massive growth of user-generated content in the web. The growing text analytics market heavily relies on text classification to offer services such as sentiment analysis, document categorization, or scientific discovery. In a nutshell, supervised text classification extracts relevant information from manually classified documents and learns a model from the extracted information. Machine learning classifiers *learn* to take decisions autonomously, so that there is no need to programmatically implement rules which are later used to automatically take decisions. Approaches based on the latter paradigm are often referred to as *rule-based*. The drawback of rule-based approaches is that they tend to fall short on generalizing (resulting in high precision but low recall). Furthermore, the rules need to be defined manually and are usually not updated once written. Machine learning classifiers overcome these drawbacks as they are not bound to explicit human-

[101] In chapter 5, we also use unsupervised machine learning, to cluster activity profiles of users to CW roles.

created rules, but learn from real-world data, which can be continuously updated. They automatically learn to take decisions (which might be based on rules but not necessarily are). The drawback of supervised learning is that classifiers require annotated data from which to learn (this requirement is omitted in unsupervised machine learning). Both supervised and unsupervised machine learning approaches are prone to suffer from a domain bias when the data on which they are trained is not appropriately selected, i.e. they might produce good results when tested on data from the same domain as the training data, but not otherwise.

Supervised machine learning is used to solve NLP applications including language identification, part-of-speech (POS) tagging, word sense disambiguation, and sentiment detection. One of the most important tasks of a machine learning classifier for textual data is to abstract over the actual content of the document it learns from. It should only extract relevant information, e.g. the nouns of a document. We refer to this information as *features*. We extensively discuss different types of features in the course of this thesis, from simple n-grams to more complex features based on document metadata or document similarity.

A.1 Supervised Text Classification Tasks

In supervised classification, given an instance $d \in D$ (e.g. a document) and a set of labels $C = \{c1, c2, ..., c_n\}$ (e.g. topics), we want to label each instance d with $L \subset C$, where L is the set of relevant or true labels. In single-label classification, each instance d is labeled with exactly one label, i.e. $|L| = 1$, whereas in multi-label classification, a set of labels is assigned to each instance, i.e. $|L| \geq 1$.[102] Single-label classification can further be divided into binary classification ($|C| = 2$) and multi-class classification ($|C| > 2$).

A supervised machine learning task on textual data is often divided into a training phase (on labeled training data), and testing (on labeled test data) or prediction (on unlabeled data). It typically requires the following steps to be performed in the given order (Bethard et al., 2014):

1. **reading labeled input data**: read raw data from any source, segment data into instances (if necessary), label each instance according to the given gold standard (true labels)

2. **preprocessing**: add further linguistic information to each of the instances (and their context if any), e.g. lemmatization, POS tagging

3. **feature extraction**: based on the text of the instances and the additional information added during preprocessing, extract information to be employed by the machine learning algorithm

[102] The case of empty predictions, where $L = \emptyset$, is discussed in Liu and Chen (2015).

4. **machine learning**: use a machine learning algorithm to train a classifier model based on the extracted features (training); or use a trained classifier model to predict the label of a given instance (prediction/testing)[103]

5. **evaluation** (only for testing): on test data, compare the predictions of a classifier trained with train data to the gold standard labels and calculate evaluation measures

In appendix B, we will explain how these steps have been implemented in the text classification framework DKPro TC. We refer to these steps as supervised text classification *tasks*.

A.2 Approaches to Supervised Machine Learning

Supervised machine learning approaches can be divided into classification (see appendix A.1) and regression approaches, where the latter assigns real numbers instead of labels to documents. In the following, we give a short list of the major algorithm families used in this thesis:

- **Tree-based**: Decision tree-based classifiers partition the input space in such a way that many simple decisions on feature values are combined into a common model, represented as a tree. Due to their tree-like structure, which follows simple choices, decision trees tend to produce intuitive models. Example: C4.5 (Quinlan, 1993).

- **Kernel-based**: Kernel-based algorithms extend otherwise linear classifiers to solve non-linear problems by mapping the input space into a higher-dimensional space. Example: Support Vector Machines with polynominal kernel (Boser et al., 1992).

- **Ensemble Methods**: Ensemble models combine multiple models into a single predictor. Combining several diverse models (even random ones) often produces better classification results as compared to a single model. Example: A Random Forest classifier (Breiman, 2001) combines several decision tree models with a random selection of features.

- **Neural Networks**: Artificial neural networks are able to process a large number of input signals from "raw data" which activate layers of interconnected neurons. Neural networks, in particular deep ones (with many layers) have been shown to produce state-of-the-art result for many NLP problems (Goodfellow et al., 2015). Example: Perceptron.

[103] Some text classification setups do not explicitly distinguish between feature extraction and machine learning. Particularly, in neural networks, features are learned automatically during the training process.

Further algorithm families for supervised text classification which are frequently used in the NLP community include Linear Regression models and Sequence Classification models. Fernández-Delgado et al. (2014) carried out an extensive comparison of almost 200 single-label classifiers on more than 100 datasets, finding that Random Forest and SVM are most likely to yield the best result. A comprehensive overview of multi-label classification algorithms and evaluation measures can be found in Madjarov et al. (2012).

The evaluation scores to measure the performance of a machine learning model highly depend on the task and the community in which it is reported. Although there is no agreed upon set of measures for the evaluation of a classifier, there is a number of measures which are frequently used, especially within the NLP community. They are based on the number of true positives *tp* (correctly classified instances with label A), true negatives *tn* (correctly classified instances with label $\bar{A}$), false positives *fp* (misclassified instances with label A), and false negatives *fn* (misclassified instances with label $\bar{A}$). The calculation of these measures ist straightforward for binary classification:

$$\text{Precision: } \frac{tp}{tp + fp}$$

$$\text{Recall: } \frac{tp}{tp + fn}$$

$$\text{Accuracy: } \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{F1 score: } 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

In multi-class classification, overall scores need to be micro- or macro-averaged over all labels. More complex evaluation scores used in this thesis are defined when mentioned for the first time.

B The DKPro Text Classification Framework

The open-source text classification framework DKPro TC has been substantially extended in the course of the work carried out in this thesis.[104] DKPro TC has evolved from a prototype of the FlawFinder System (Ferschke et al., 2012b). In the following Sections, we give a short overview of the involved technologies, and how the requirements outlined in appendix A.1 are implemented in DKPro TC. Furthermore, we explain where and how DKPro TC has been used to support the experiments carried out as part of this thesis.

[104] DKPro TC is joint work under the main guidance of Johannes Daxenberger, Oliver Ferschke, and Torsten Zesch. The source code is freely available at `https://dkpro.github.io/dkpro-tc`, accessed May 25, 2015.

B.1 Underlying technologies

DKPro TC is implemented in Java and based on the Apache UIMA and uimaFIT frameworks for unstructured information management (Ferrucci and Lally, 2004; Ogren and Bethard, 2009). We use the Common Analysis Structure (CAS), provided by UIMA, to represent instances (input documents) and annotations (added during preprocessing) in a standardized way. DKPro TC is part of the DKPro software family (Eckart de Castilho and Gurevych, 2014). It relies on the DKPro Lab (Eckart de Castilho and Gurevych, 2011) workflow engine, which allows fine-grained control over the dependencies between single tasks, e.g. the preprocessing of a document needs to happen before the feature extraction. As part of the reading input data and preprocessing tasks, DKPro TC mostly reuses DKPro Core (Eckart de Castilho and Gurevych, 2014), e.g. to apply segmentation, POS tagging or parsing on the input documents. For the actual machine learning part, DKPro TC integrates several frameworks, among them Weka (Hall et al., 2009), Meka[105], and CRFsuite[106]. As part of the evaluation routine, DKPro TC offers integration with the STATSREP-ML framework for statistical evaluation (Guckelsberger and Schulz, 2014).

B.2 Standardized document processing and feature development

DKPro TC makes use of the sequential nature of the tasks involved in supervised text classification (see appendix A.1) and assembles them in a comprehensible way. Figure B.1 visualizes the DKPro TC task sequence in a train/test setup. It reflects the tasks explained in appendix A.1 and additionally includes a task which collects information over the whole training data. The latter is necessary for feature extractors which need to relate information from a particular instance to the same information for the whole (training) data, e.g. TF-IDF scores for n-gram feature extractors. The input data (split into training and test data) is read and (if necessary) split into instances, one instance per CAS. Training and test data are processed separately, the global information used during feature extraction on the test data stems from the training data.

Beyond simplifying the setup of a text classification pipeline, DKPro TC's main focus is to support users in determining an appropriate set of features for a given task. To keep the framework flexible, we support several kinds of instances and features which are processed in the pipeline:

- **document**: a full document, e.g. a newspaper text, which should be classified according to its topic

[105] `http://meka.sourceforge.net`, accessed May 25, 2015
[106] `http://www.chokkan.org/software/crfsuite`, accessed May 25, 2015

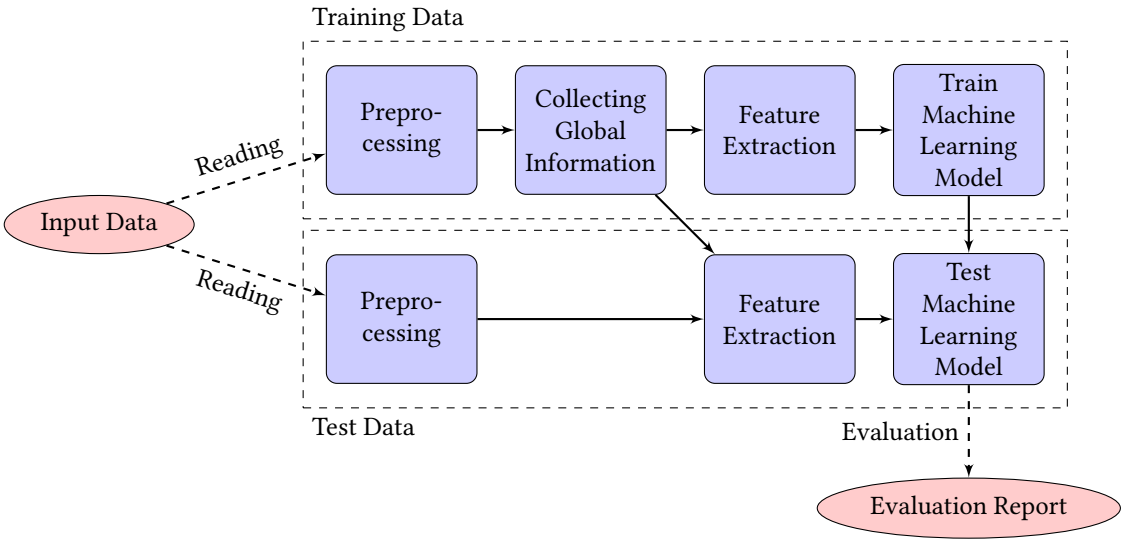


Figure B.1: A prototypical train/test pipeline with tasks in DKPro TC.

- **pair**: two instances which should be classified according to a given property, e.g. two sentences which should be classified according to whether they form a paraphrase or non-paraphrase
- **unit**: instances as part of a text, e.g. the nouns of a document which should be classified according to whether they form a Named Entity and if so, which one
- **sequence**: instances as part of a sequence, e.g. the words of a sentence (word: instance, sentence: sequence) which should be labeled with POS tags

We call the different approaches *feature modes*, as they influence the way in which features are extracted. In **document mode**, features are extracted based on the entire document. In **pair mode**, features from both documents, or features describing differences/similarities between the documents are extracted. In **unit mode**, features based on the particular instance and its context are extracted. In **sequence mode**, features based on the particular instance and its subsequent or previous instances are extracted. Unit and sequence modes are intended to be used for the classification of small units within a text (e.g. words or sentences), and which can only be reliably classified within their context, as determined by the surrounding text. Apart from different document modes, DKPro TC also supports several (machine) *learning modes*, namely **single-label** and **multi-label** classification, **sequence classification**, and **regression**. The learning mode influences how instances are labeled when they are first read, and how machine learning and evaluation are carried out.

Fokkens et al. (2013) noted that NLP experiments are not replicable in most cases. DKPro TC addresses this issue because it (a) encourages users to reuse existing components which they can refer to in research papers rather than writing their own components (due

to its modular and flexible architecture), (b) documents all performed steps through extensive logging, and (c) makes it possible to re-run experiments with minimal effort (by sharing the basic configuration and novel feature extractors). Apart from helping the replicability of experiments, DKPro TC encourages reusing existing components and therefore allows the user to concentrate on the new functionality that is specific to the planned experiment instead of having to reinvent the wheel. Many parts of a text classification system are not specific to a particular corpus or experiment setup and can thus be reused. In DKPro TC, this includes readers and preprocessing components from DKPro Core, generic feature extractors (e.g. n-gram extractors), machine learning algorithms from various third-party frameworks, and evaluation functionality.

B.3 DKPro TC in this work

DKPro TC or one of its prototypes have been used to carry out the following experiments, as described in the main part of this thesis:

- classification of English edits: train/test and prediction setup; document mode, multi-label learning (section 4.3.3)
- classification of German edits: train/test setup; pair mode, multi-label learning (section 4.4.2)
- classification of English revisions: train/test and prediction setup; pair mode, multi-label learning (section 4.5.2)
- classification of edit-turn-pairs: cross-validation and prediction setup; pair mode[107], single-label learning (section 6.3)

During the course of this work, the core functionality of DKPro TC has been extended to make the framework more generic. The experiments described in section 4.4.2 and section 4.5.2 (WPEC-GER and WPRC) take advantage of the pair mode. Rather than calculating features based on a previously calculated diff of consecutive revisions as we did in section 4.3.3 (WPQAC), we calculated the features on a pair of instances, namely for each edit the edited text span from the old revision and the edited text span from the new revision. The current state of DKPro TC can therefore be attributed partly to the experiments carried out in this work. To demonstrate how we applied DKPro TC, we list a couple of code snippets.[108]

Listing 1 is a very simple feature extractor which retrieves metadata from the Wikipedia revisions. The `WikipediaRevision` annotation, which persists revision metadata, needs to be created during the instantiation of the CAS, i.e. in the reader, so that it can be accessed

[107] To be precise, we extended the pair mode to account for three instances: the edited text span from the old revision, the edited text span from the new revision, and the turn text.

[108] The code refers to the 0.6.0. version of DKPro TC.

```
public class RevisionIsMinor
extends FeatureExtractorResource_ImplBase implements PairFeatureExtractor
{
    @Override
    public List<Feature> extract(JCas oldEdit, JCas newEdit){
        // retrieving previously annotated metadata of the newer revision
        WikipediaRevision revision = JCasUtil.selectSingle(newEdit, WikipediaRevision.class);
        boolean isMinor = revision.isMinor();
        return new Feature("RevisionIsMinor", isMinor);
    }
}
```

Listing 1: Java code of a simple DKPro TC pair mode feature extractor which we used in our experiments to determine whether a revision is marked as minor change or not (imports omitted).

```
public class DiffNumberInternalLinks
extends FeatureExtractorResource_ImplBase implements PairFeatureExtractor
{
    @Override
    public List<Feature> extract(JCas oldEdit, JCas newEdit){
        // retrieving previously annotated internal links, based on wiki markup
        Collection <WikiInternalLink> oldLinks = JCasUtil.select(oldEdit, WikiInternalLink.class);
        Collection <WikiInternalLink> newLinks = JCasUtil.select(newEdit, WikiInternalLink.class);

        // only record differences
        double diffLinks = Math.abs(oldLinks.size() - newLinks.size());
        return new Feature("DiffNumberInternalLinks", diffLinks).asList();
    }
}
```

Listing 2: Java code of a DKPro TC pair mode feature extractor which we used in our experiments to extract the number of added/deleted internal links (imports omitted).

during feature extraction, as demonstrated in listing 1. Listing 2 shows a slightly more complex feature extractor which calculates the number of changed (inserted or deleted) internal Wikipedia links. It retrieves `WikiInternalLink` annotations, which have been added by a wiki markup parser during preprocessing.

C Annotation Guidelines

In this part of the Appendix, we include a condensed version of the annotation guidelines given to our annotators for the annotation of the corpora that were created in the course of this work: WPEC, WPEC-GER, and ETP-GOLD. WPEC and WPEC-GER were annotated

with the help of a small number of annotators on site, whereas ETP-GOLD was annotated via crowdsourcing on Amazon Mechanical Turk.[109]

C.1 Annotation Guidelines for WPEC and WPEC-GER

These guidelines have been designed to annotate *edits*, as calculated from a pair of consecutive Wikipedia revisions r_{v-1}, r_v, extracted from MediaWiki dumps. The revision text has not been parsed, but includes the source text as-is, including wiki markup. The annotation editor (cf. section 4.3.1) gives the annotator access to the following data (referring to r_v):

1. the title of the article
2. the time stamp of the revision
3. the user name or IP address
4. the comment of the user (if present)
5. a link to the corresponding diff page of Wikipedia's online API
6. the source text of r_{v-1}
7. the source text of r_v
8. a diff view where edits are highlighted and listed with the above properties

C.1.1 Detailed Guidelines

Our edit category taxonomy is hierarchical, cf. figure 4.4. Except for the categories REVERT, VANDALISM and OTHER, all categories are divided into edits altering the surface of the text, i.e. edits, which do not affect the meaning, and edits that actually alter the text-base of the article's content, i.e. its meaning. Edits which alter the meaning are further divided according to whether they insert, delete or modify text. These actions correspond partly with the basic edit types that are calculated by the line-based edit segmentation algorithm (Insertion, Deletion, Modification, Relocation), but are not to be confused. The algorithm that calculates the edits is designed to generate human readable edits. Though, it might not always find the nearby way to transform one revision into another. For edits altering the form, only MARKUP edits are further divided. We allow for multi-labeling, i.e. one edit may be labeled with more than one category. In what follows, we list detailed explanation for each of the categories.

[109] The Human Intelligence Task (HIT) Layout and Setup were carried out by Emily Jamison.

Template Templates have wide usage in Wikipedia. Their basic task is to display some kind of information in the same way in many pages. That includes infoboxes, navigational boxes, warnings and others. The wiki syntax to create templates is double curly brackets (`{{template name}}`). All edits related to this syntax have to be annotated as Template. Many times, editing templates (especially more complex ones) also brings a gain or loss of information which has to be annotated accordingly.

Reference References include any kind of link, no matter if they are internal or external. Wiki links are usually generated using single or double square brackets, while references can be added with *</ref>* tags. Both of them are annotated as reference. If attributes inside a reference are changed or corrected (e.g. the ISBN inside a bibliographical reference), we consider this as Information-Modify or accordingly. Reference-Modify only applies, where the reference itself (i.e. the internal or external link/anchor or in the case of bibliographic references the book, article etc.) is changed. Many bibliographical references are generated via templates, as in `<ref>cite book | ... </ref>`, such an edit has to be annotated as Reference-Insert *and* Template-Insert.

File File references behave quite similar to other references, but have to be annotated differently, as we consider embedding images and other files inside the article substantially different than linking to other resources. For that reason, only sources that are actually displayed on the article page itself have to be annotated as File. References like `[[Media:Name_of_file|Link]]`, which might link to an image, but do not display the image inside the article, have to be annotated as Reference. In contrast to other references, descriptions of files which are displayed together with the file on the article page (i.e. captions) are to be labeled separately as they usually contain textual content. If, for example, a new image and its description are inserted into an article, this has to be marked as both File-Insert and Information-Insert. Changes in the markup information of files like their alignment or size have to be annotated as Markup-Modify.

Information Information includes all basic textual changes which are not used for formatting or reference handling and change the meaning of the text. Textual changes which do not affect the meaning (e.g. of a sentence) are to be annotated as Paraphrase.

Relocation The Relocation category is for annotating all sorts of (often copy-paste) edits, which move parts of the article to other places without changing them. This should be annotated only if the content of the fragment that has been moved did not change at all. Text including markup as well as files, references etc., can be affected by relocations. A fragment that has been relocated cannot be multi-labeled, no matter if it is a link, an image, or text.

Paraphrase Paraphrase is intended for edits processing changes inside a sentence, which do not change the meaning of the sentence, i.e. paraphrases. Edits, that change the meaning are to be annotated as Information-Modify. In very rare cases this might happen to a portion of text spanning across more than one single sentence. A typical example of a Paraphrase edit would be to replace a single word with a synonym to avoid repetition. Local modifications that only affect grammatical or spelling corrections are to be annotated as Spelling/Grammar.

Spelling/Grammar Any kind of grammar and/or spelling error correction is to be labeled with Spelling/Grammar.

Markup Markup essentially refers to wiki syntax. Only changes made to wiki markup characters (as long as they are not covered by other categories like Template or Reference) are to be annotated as Markup including any HTML characters that can be used inside the wiki markup. To give an example: `===Some Heading===` is changed to `==Some other Heading==`. That would affect both the markup of the text as well as the textual content, since the name of the headline has been changed. Consequently, the edit must be labeled with both Markup-Modify and Information-Modify.

Vandalism Generally, edits marked with Vandalism, must not be multi-labeled. In virtually all cases of vandalism, the entire revision will be affected, so if more than one edit is present, all of them should be labeled as Vandalism. An edit has to be labeled as vandalism, if

- (almost) the entire article is deleted (page blanking)
- paragraphs or sentences are removed, changed or inserted without any reasons (which would usually be stated in the comment)
- letters or words are removed which are needed to guarantee the readability and comprehensibility of a sentence or word
- files, references or templates are removed or changed without reasons
- (almost) the entire article is rephrased
- paragraphs, sentences, words or letters are replaced by nonsense text or new content that impedes the readability of the fragment or the entire article
- facts like numbers or names are changed to wrong values
- nonsense/sexist/offense sentences or words are inserted

- other changes are present that are obviously vandalism.

Most vandalistic edits do not have a comment.

Revert A revert undoes the effects of one or more edits, usually restoring a previous revision of a page. Most of the time, this refers to reverting the last change and happens due to vandalism. The latest version of an article can be reverted to the previous revision or to any older revision. The most common and most obvious way of denoting revert-actions by users is to signal them in the revision comment (typically something like "Reverted [one or more] edit[s] by [username_1] to last revision by [username_2]"). The best way to detect reverts is the user's comment. Keywords for Reverts in comments include morphological derivations of the words "revert" (abr. "rv" or "rvv") and "undo". Reverts cannot be multi-labeled.

Other Edits that are hard or impossible to classify because of their segmentation, should be marked as other. Edit labeled as Other cannot be multi-labeled.

C.1.2 Insertions, Deletions and Modifications

This Section explains how to distinguish Insertions, Deletions and Modifications.

Insert Insert actions are only present, if new content or markup is added to the article, i.e. if the edit's content or markup has not been present in r_{v-1} but is present in r_v. Insertions that include text and markup (e.g. inserting a new headline), have to be multi-labeled according to the general guidelines with both Information-Insert, Markup-Insert.

Delete An edit is a deletion, if the edit's content or markup has been present in r_{v-1} but is not present in r_v. Edits deleting various basic elements, e.g. text and markup, have to be multi-labeled correspondingly.

Modification Modifications only apply for content and markup belonging to the *same* segment (template, word, markup element) in both r_{v-1} and r_v. Template-Modify applies, when the type of a template (i.e. its name) is changed but not if textual contents inside the template are edited. The latter has to be labeled as Information-Modify or Paraphrase depending on whether the change affects the meaning of the text or not. The same is true for formatting issues: only changes in the actual wiki syntax are considered as Markup-Modify, while edits in between wiki markup characters or HTML tags which only affect the textual content are to be labeled as Information-Modify or Paraphrase.

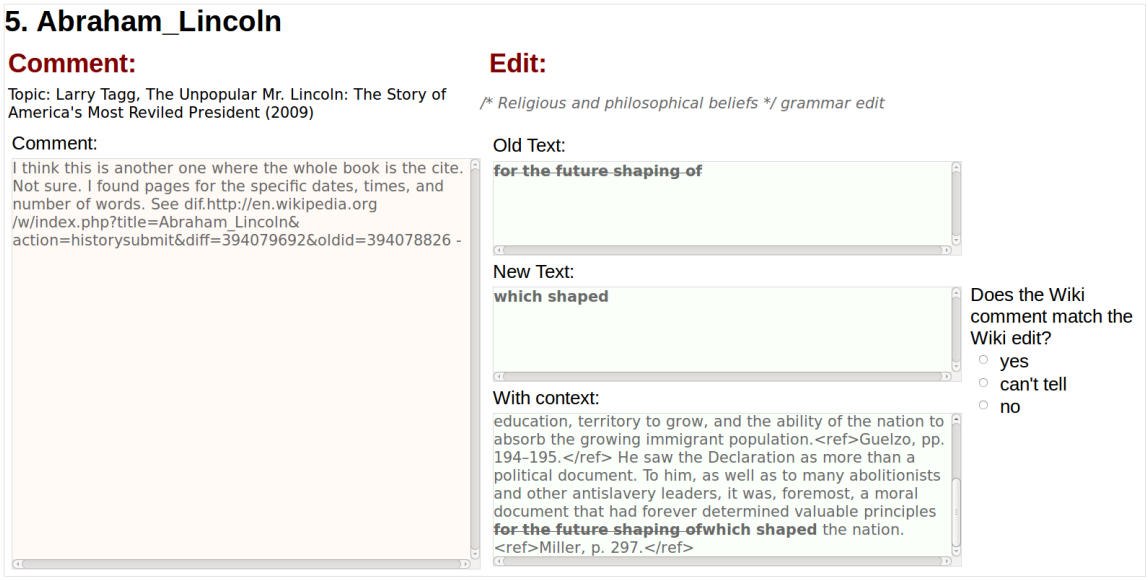


Figure C.2: Layout of a HIT to annotate an edit-turn-pair; the four upper edits are omitted.

C.1.3 General Remarks

- Lists and Tables: When a list or a table is started or deleted, both Information and Markup are affected. If a new line is added to a list, this is considered as an addition of Information, not Markup. If a former list was transformed into plain text or vice versa without changing the actual text, this had to be labeled as Markup.
- Templates: If text inside templates is changed, but not the template itself, the edit is to be labeled as Information, File, Reference and the like. Generally, templates are not to be considered as markup as they have their own category. Templates that generate references on the surface are also considered as templates.
- Bot edits: Edits by bots are annotated as if they were made by humans.
- Major edits: Edits affecting entire paragraphs or large portions of text sometimes change, insert or delete text, references, files and markup at the same time, but are marked as single edits. In those cases, multi-labeling applies and all basic elements like templates, markup, references and text which have been inserted, deleted or changed have to be considered and labeled with their corresponding categories.

C.2 Annotation Guidelines for ETP-GOLD

ETP-GOLD contains pairs of edits and turns (segments from Wikipedia discussion pages), which should be labeled according to whether they correspond to each other or not. The

Determine whether or not this comment describes this edit.

- Comments are user contributions from discussion pages in the English Wikipedia.
- Edits are modifications to existing articles in the English Wikipedia.

Comments and edits are taken from Wikipedia pages, so they may contain Wiki markup.
Edits:
Boldface text has been inserted.
~~**Boldface and crossed out text**~~ has been deleted.
Boldface and italics text was moved.

Here are the ways a comment can describe an edit:

- The comment is an explicit suggestion, recommendation or request and the edit performs this suggestion, recommendation or request
- The comment is an explicit reference and the edit adds or modifies this reference
- The comment is a commitment to an action in the future and the edit performs this action
- The comment is a report of a performed action (self-performed or performed by another user) and the edit performs this action

Each comment has a *Topic* (headline of the thread it is part of)
Each edit may have a *Description* (short text describing the nature of the edit, manually inserted by its creator)

Objective: This data will be used for natural language processing research.

Figure C.3: The instructions displayed to Mechanical Turk workers in each HIT (examples omitted).

Mechanical Turk workers had access to Human Intelligence Tasks (HITs) which each contained five edit-turn-pairs. We parsed the text of the turn and displayed it without wiki markup; whereas the text affected by the edit is reproduced with markup. In addition to the edited text, we also displayed the context in which the edit happened. Figure C.2 shows the web interface we used. For each of the pairs, the Mechanical Turk workers needed to answer the question: "Does the Wiki comment match the Wiki edit?" Possible answer options were: "Yes", "Can't tell", and "No". The Mechanical Turk workers were also given instructions at the head of each HIT, reproduced in figure C.3. We labeled turns as "comments" as this term seemed to be less confusing to Mechanical Turk workers. In addition to the instructions, each of the four ways in which edits and turns may correspond, were illustrated by an example. The examples are reproduced in figure C.4a. Furthermore, we also added an example for a non-corresponding edit-turn-pair, cf. figure C.4b.

Explicit suggestion:

Now that I've taken an interest in the article one other item that might be useful in this very good article is the theatre infobox.	~~**[[Image:5th Ave Theater Marquee (Seattle) 2007-08.jpg|thumb|right|The off-season [[marquee]] of the 5th Avenue Theatre.]]**~~**{{Infobox Theatre |name = 5th Avenue Theatre |website = www.5thAvenue.org}}**

Explicit reference:

The A330 comparison here is per the Flug Revue [http://web.archive.org/web/20080513172529/LINK source]. The comparison in the A340 article is currently unreferenced there.	The 767-400ER's closest competitor from Airbus is the [[Airbus A330#A330-200|A330-200]].<ref name=Flug767-400>[**http://web.archive.org/web/20080513172529/**LINK "Boeing 767-400ER"].

Commitment to action:

Yeeks, you are right! I've looked back through the page history, and it wasn't deleted by mistake any time recently. I'll correct the text. Thanks for finding that!	~~**The illustration on the right shows a thin section of one hemisphere of the brain of a Chlorocebus monkey, [...] or different types of brain tissue, in distinct ways; the Nissl stain shown here is probably the most widely used.**~~

Report of action:

I think I found the source of the confusion. Donald indicates 2 justices of the Illinois Supreme Ct. licensed AL to practice in Sept. of 1836, but the Supreme Ct. Clerk did not enroll him until March 1, 1837. I added a footnote to this effect.	[[Admission to the bar in the United States|Admitted to the bar]] in 1836, [...] **<ref> AL was added to roll of attorneys by the Clerk, Illinois Supreme Court on March 1, 1837[[#Donald|Donald (1996)]], p.64.</ref>** he moved to Springfield, Illinois and began to practice law under [[John T. Stuart]], Mary Todd's cousin.

(a) Corresponding edit-turn-pairs.

There are several compounds harder than cubic boron nitride, most of which are nanocomposites.	[[de:Aminosäure]] ~~**[[he:חומצתאמינ ו]]**~~ [[es:Aminoácido]]

(b) A non-corresponding edit-turn-pair.

Figure C.4: Example edit-turn-pairs given to the Mechanical Turk workers. Turns are to the left; edits to the right. Formatting conventions: inserted text is boldfaced, deleted text is boldface and crossed out.

List of Tables

List of Figures

Bibliography

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. (Cited on pages 52, 68, 77, 79 and 84)

Aggarwal, C. C. (2011). *Social Network Data Analytics*. Springer US, Boston, MA. (Cited on page 36)

Aji, A., Wang, Y., and Agichtein, E. (2010). Using the Past To Score the Present: Extending Term Weighting Models Through Revision History Analysis. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 629–638, Toronto, Canada. (Cited on page 52)

Allen, N., Atkinson, D., Morgan, M., Moore, T., and Snow, C. (1987). What Experienced Collaborators Say About Collaborative Writing. *Journal of Business and Technical Communication*, 1(2):70–90. (Cited on pages 2, 22, 23, 32 and 142)

Anderka, M., Stein, B., and Lipka, N. (2012). Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. In *35th International ACM Conference on Research and Development in Information Retrieval*, pages 981–990, Portland, OR, USA. (Cited on pages 57 and 140)

Antin, J., Cheshire, C., and Nov, O. (2012). Technology-Mediated Contributions: Editing Behaviors Among New Wikipedians. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 373–382, Seatttle, WA, USA. (Cited on pages 53, 54, 68 and 91)

Arazy, O., Gellatly, I., Jang, S., and Patterson, R. (2009). Wiki deployment in corporate settings. *IEEE Technology and Society Magazine*, 28(2):57–64. (Cited on pages 29, 59 and 142)

Arazy, O., Nov, O., and Oded, N. (2010). Determinants of Wikipedia Quality: the Roles of Global and Local Contribution Inequality. In *Proceedings of the 2010 ACM Conference on*

Computer Supported Cooperative Work, pages 233–236, Savannah, GA, USA. (Cited on page 57)

Arazy, O., Nov, O., and Ortega, F. (2014). The [Wikipedia] World is not flat: on the Organizational Structure of Online production Communities. In *European Conference on Information Systems*, Tel Aviv, Israel. (Cited on pages 44, 47, 48, 105, 117 and 118)

Arazy, O., Nov, O., Patterson, R., and Yeo, L. (2011). Information Quality in Wikipedia: The Effects of Group Composition and Task Conflict. *Journal of Management Information Systems*, 27(4):71–98. (Cited on pages 90, 109, 110 and 130)

Arazy, O., Ortega, F., Nov, O., Yeo, L., and Balila, A. (2015). Functional Roles and Career Paths in Wikipedia. In *Proceedings of the 18th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, pages 1092–1105, Vancouver, BC, Canada. (Cited on pages 44, 105 and 117)

Arazy, O., Yeo, L., and Nov, O. (2013). Stay on the Wikipedia task: When task-related disagreements slip into personal and procedural conflicts. *Journal of the American Society for Information Science and Technology*, 64(8):1634–1648. (Cited on pages 90 and 109)

Arthur, D. and Vassilvitskii, S. (2007). K-means++: The Advantages of Careful Seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, Philadelphia, PA, USA. (Cited on page 112)

Bayá, A. E. and Granitto, P. M. (2013). How Many Clusters: A Validation Index for Arbitrary-Shaped Clusters. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(2):401–414. (Cited on page 115)

Behles, J. (2013). The Use of Online Collaborative Writing Tools by Technical Communication Practitioners and Students. *Technical Communication*, 60(1):28–44. (Cited on pages 29 and 59)

Benkler, Y. (2002). Coase's Penguin, or, Linux and the Nature of the Firm. *Yale Law Journal*, 112(3):367–445. (Cited on page 31)

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press. (Cited on page 31)

Bethard, S., Ogren, P., and Becker, L. (2014). ClearTK 2.0: Design Patterns for Machine Learning in UIMA. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3289–3293, Reykjavik, Iceland. (Cited on page 158)

Birnholtz, J. and Ibara, S. (2012). Tracking Changes in Collaborative Writing: Edits, Visibility and Group Maintenance. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 809–818, Seattle, WA, USA. (Cited on page 59)

Bisaillon, J. (2007). Professional Editing Strategies Used by Six Editors. *Written Communication*, 24(4):295–322. (Cited on page 24)

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer New York. (Cited on page 157)

Blumenstock, J. E. (2008). Size Matters: Word Count as a Measure of Quality on Wikipedia. In *Proceedings of the 17th International Conference on World Wide Web*, pages 1095–1096, Beijing, China. (Cited on page 57)

Borra, E., Laniado, D., Weltevrede, E., Mauri, M., Magni, G., Venturini, T., Ciuccarelli, P., Rogers, R., and Kaltenbrunner, A. (2015). A Platform for Visually Exploring the Development of Wikipedia Articles. In *Proceedings of the 9th International AAAI Conference of Web and Social Media*, page (to appear), Oxford, UK. (Cited on page 148)

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, USA. (Cited on page 159)

Brandes, U., Kenis, P., Lerner, J., and van Raaij, D. (2009). Network analysis of collaboration structure in Wikipedia. In *Proceedings of the 18th International Conference on World Wide Web*, pages 731–740, Madrid, Spain. (Cited on pages 56 and 153)

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32. (Cited on pages 52, 84, 93, 132, 138 and 159)

Bridwell, L. S. (1980). Revising Strategies in Twelfth Grade Students' Transactional Writing. *Research in the Teaching of English*, 14(3):197–222. (Cited on page 18)

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. (Cited on pages 5, 52, 54, 77, 79, 84, 138 and 148)

Bryant, S. L., Forte, A., and Bruckman, A. (2005). Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*, pages 1–10, Sanibel Island, FL, USA. (Cited on page 44)

Buriol, L. S., Castillo, C., Donato, D., Leonardi, S., and Millozzi, S. (2006). Temporal Analysis of the Wikigraph. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 45–51, Washington DC, USA. (Cited on page 52)

Burke, M. and Kraut, R. (2008). Mopping Up: Modeling Wikipedia Promotion Decisions. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, pages 27–36, San Diego, California, USA. (Cited on page 44)

Butler, B., Joyce, E., and Pike, J. (2008). Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in Wikipedia. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1101–1110, Florence, Italy. (Cited on page 39)

Butler, B. and Sproull, L. (2007). Community Effort in Online Groups: Who Does the Work and Why? In Weisband, S. P., editor, *Leadership at a Distance: Research in Technologically-Supported Work*, chapter 9, pages 171–194. Psychology Press. (Cited on pages 44 and 105)

Cabrio, E., Magnini, B., and Ivanova, A. (2012). Extracting Context-Rich Entailment Rules from Wikipedia Revision History. In *Proceedings of the 3rd Workshop on The People's Web meets NLP*, pages 34–43, Jeju Island, Republic of Korea. (Cited on page 51)

Cahill, A., Madnani, N., Tetreault, J., and Napolitano, D. (2013). Robust Systems for Preposition Error Correction Using Wikipedia Revisions. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 507–517, Atlanta, GA, USA. (Cited on page 52)

Callahan, E. S. and Herring, S. C. (2011). Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*, 62(10):1899–1915. (Cited on page 44)

Callero, P. L. (1994). From Role-Playing to Role-Using: Understanding Role as Resource. *Social Psychology Quarterly*, 57(3):228–243. (Cited on page 44)

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, Hyderabad, India. (Cited on page 54)

Cosley, D., Frankowski, D., Terveen, L., and Riedl, J. (2006). Using Intelligent Task Routing and Contribution Review to Help Communities Build Artifacts of Lasting Value. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1037–1046, Montreal, Canada. (Cited on page 122)

Cress, U. and Kimmerle, J. (2008). A systemic and cognitive view on collaborative knowledge building with wikis. *International Journal of Computer-Supported Collaborative Learning*, 3(2):105–122. (Cited on pages 24 and 126)

Daiute, C. (1986). Physical and Cognitive Factors in Revising: Insights from Studies with Computers. *Research in the Teaching of English*, 20(2):141–159. (Cited on pages 13 and 28)

Daiute, C. A. (1982). Psycholinguistic perspectives on revising. *Revising: New essays for teachers of writing*, pages 109–120. (Cited on page 21)

Damerau, F. J. (1964). A Technique for Computer Detection and Correction of Spelling Errors. *Communications of the ACM*, 7(3):171–176. (Cited on page 79)

Daniels, P. T. and Bright, W. (1996). *The world's writing systems*. Oxford University Press. (Cited on page 13)

Daxenberger, J., Ferschke, O., Gurevych, I., and Zesch, T. (2014). DKPro TC: A Java-based Framework for Supervised Learning Experiments on Textual Data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pages 61–66, Baltimore, MD, USA. (Cited on page 9)

Daxenberger, J. and Gurevych, I. (2012). A Corpus-Based Study of Edit Categories in Featured and Non-Featured Wikipedia Articles. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 711–726, Mumbai, India. (Cited on page 8)

Daxenberger, J. and Gurevych, I. (2013). Automatically Classifying Edit Categories in Wikipedia Revisions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 578–589, Seattle, WA, USA. (Cited on page 8)

Daxenberger, J. and Gurevych, I. (2014). Automatically Detecting Corresponding Edit-Turn-Pairs in Wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Short Papers*, pages 187–192, Baltimore, MD, USA. (Cited on page 9)

Dishaw, M., Eierman, M. A., Iversen, J. H., and Philip, G. C. (2011). Wiki or Word? Evaluating Tools for Collaborative Writing and Editing. *Journal of Information Systems Education*, 22(1):43–54. (Cited on page 59)

Dohrn, H. and Riehle, D. (2011). Design and implementation of the Sweble Wikitext parser. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 72–81, Mountain View, CA, USA. (Cited on pages 65, 77 and 136)

Dutrey, C. C., Bouamor, H., Bernhard, D. D., and Max, A. (2011). Local Modifications and Paraphrases in Wikipedia's Revision History. *Procesamiento del lenguaje natural*, 46:51–58. (Cited on page 52)

Eckart de Castilho, R. and Gurevych, I. (2011). A Lightweight Framework for Reproducible Parameter Sweeping in Information Retrieval. In *Proceedings of the Workshop on Data Infrastructures for Supporting Information Retrieval Evaluation*, pages 7–10, Glasgow, UK. (Cited on page 161)

Eckart de Castilho, R. and Gurevych, I. (2014). A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In Ide, N. and Grivolla, J., editors, *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, Dublin, Ireland. (Cited on page 161)

Ede, L. and Lunsford, A. (1990). *Singular Text/Plural Authors: Perspectives on Collaborative Writing*. Southern Illinois University Press. (Cited on pages 22 and 24)

Einsohn, A. (2011). *The Copyeditor's Handbook A Guide for Book Publishing and Corporate Communications*. University of California Press. (Cited on page 15)

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pages 973–978, Seattle, WA, USA. (Cited on pages 131 and 132)

Emig, J. A. (1971). *The composing processes of twelfth graders*. National Council of Teachers of English. (Cited on pages 15 and 21)

Erkens, G., Jaspers, J., Prangsma, M., and Kanselaar, G. (2005). Coordination processes in computer supported collaborative writing. *Computers in Human Behavior*, 21(3):463–486. (Cited on pages 20, 23, 29, 32 and 142)

Faigley, L. and Miller, T. (1982). What we learn from writing on the job. *College English*, 44(6):557–569. (Cited on page 22)

Faigley, L. and Witte, S. (1981). Analyzing revision. *College Composition and Communication*, 32(4):400–414. (Cited on pages 15, 17, 18, 19, 53, 67, 68 and 149)

Fan, S., Li, X., and Zhao, J. L. (2012). Collaboration process patterns and efficiency of issue resolution in software development. In *2012 International Conference on Collaboration Technologies and Systems*, pages 559–565, Denver, CO, USA. (Cited on page 21)

Faraj, S. and Johnson, S. L. (2011). Network Exchange Patterns in Online Communities. *Organization Science*, 22(6):1464–1480. (Cited on page 106)

Feldstein, A. (2011). Deconstructing Wikipedia: Collaborative Content Creation in an Open Process Platform. *Procedia - Social and Behavioral Sciences*, 26:76–84. (Cited on page 57)

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, 15:3133–3181. (Cited on page 160)

Ferron, M. and Massa, P. (2014). Beyond the encyclopedia: Collective memories in Wikipedia. *Memory Studies*, 7(1):22–45. (Cited on page 130)

Ferrucci, D. and Lally, A. (2004). UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10(3-4):327–348. (Cited on pages 70 and 161)

Ferschke, O. (2014). *The Quality of Content in Open Online Collaboration Platforms: Approaches to NLP-supported Information Quality Management in Wikipedia*. PhD thesis, Darmstadt. (Cited on pages 25, 57, 103 and 152)

Ferschke, O., Daxenberger, J., and Gurevych, I. (2013). 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, pages 121–160. Springer. (Cited on pages 9 and 57)

Ferschke, O., Gurevych, I., and Chebotar, Y. (2012a). Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786, Avignon, France. (Cited on pages 41, 54, 56, 127, 128, 129, 132, 138 and 155)

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. (Cited on page 160)

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. (Cited on page 62)

Fitzgerald, J. (1987). Research on Revision in Writing. *Review of Educational Research*, 57(4):481–506. (Cited on pages 1, 15, 17, 18, 41, 154 and 156)

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–381. (Cited on page 73)

Flekova, L., Ferschke, O., and Gurevych, I. (2014). What Makes a Good Biography? Multidimensional Quality Analysis Based on Wikipedia Article Feedback Data. In *Proceedings of the 23rd International World Wide Web Conference*, pages 855–866, Seoul, Korea. (Cited on page 152)

Flöck, F., Laniado, D., Stadthaus, F., and Acosta, M. (2015). Towards Better Visual Tools for Exploring Wikipedia Article Development — The Use Case of "Gamergate Controversy". In *Proceedings of the Workshop on Wikipedia, a Social Pedia: Research Challenges and Opportunities*, page (to appear), Oxford, UK. (Cited on page 148)

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. (Cited on pages 52, 54 and 68)

Flower, L. and Hayes, J. R. (1981). A Cognitive Process Theory of Writing. *College Composition and Communication*, 32(4):365. (Cited on pages 14, 15, 16, 18, 23 and 26)

Fokkens, A., van Erp, M., Postma, M., Pedersen, T., Vossen, P., and Freire, N. (2013). Offspring from Reproduction Problems: What Replication Failure Teaches Us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*,

pages 1691–1701, Sofia, Bulgaria. (Cited on page 162)

Forestier, M., Stavrianou, A., Velcin, J., and Zighed, D. A. (2012). Roles in social networks: Methodologies and research issues. *Web Intelligence and Agent Systems*, 10(1):117–133. (Cited on page 149)

Forte, A., Larco, V., and Bruckman, A. (2009). Decentralization in Wikipedia Governance. *Journal of Management Information Systems*, 26(1):49–72. (Cited on page 32)

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 1606–1611, Hyderabad, India. (Cited on page 79)

Galegher, J. and Kraut, R. E. (1994). Computer-mediated communication for intellectual teamwork: a field experiment in group writing. *Information Systems Research*, 5(2):110–138. (Cited on pages 23, 26, 28 and 148)

Geiger, R. S. and Halfaker, A. (2013). When the levee breaks: without bots, what happens to Wikipedia's quality control processes? In *Proceedings of the 9th International Symposium on Open Collaboration*, pages 1–6, Hong Kong, China. (Cited on page 49)

Georgescu, M., Kanhabua, N., Krause, D., Nejdl, W., and Siersdorfer, S. (2013). Extracting Event-Related Information from Article Updates in Wikipedia. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*, pages 281–284, Moscow, Russia. (Cited on page 52)

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901. (Cited on pages 41 and 57)

Goldberg, A., Russell, M., and Cook, A. (2003). The Effect of Computers on Student Writing: A Meta-analysis of Studies from 1992 to 2002. *The Journal of Technology, Learning and Assessment*, 2(1). (Cited on page 28)

Goodfellow, I., Courville, A., and Bengio, Y. (2015). *Deep Learning*. Book in preparation for MIT Press. (Cited on page 159)

Guckelsberger, C. and Schulz, A. (2014). STATSREP-ML: Statistical Evaluation & Reporting Framework for Machine Learning Results. (Cited on page 161)

Halfaker, A., Geiger, R. S., Morgan, J. T., and Riedl, J. (2012). The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline. *American Behavioral Scientist*, 57(5):664–688. (Cited on pages 39, 47, 118, 120, 121 and 150)

Halfaker, A., Keyes, O., and Taraborelli, D. (2013). Making peripheral participation legitimate: reader engagement experiments in Wikipedia. In *Proceedings of the 16th Conference on Computer Supported Cooperative Work*, pages 849–860, San Antonio,

Texas, USA. (Cited on page 152)

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10–18. (Cited on pages 78, 112, 136 and 161)

Han, J., Wang, C., and Jiang, D. (2011). Probabilistic Quality Assessment Based on Article's Revision History. In *Proceedings of the 22nd International Conference on Database and Expert Systems Applications*, pages 574–588, Toulouse, France. (Cited on page 57)

Hanjani, A. M. and Li, L. (2014). Exploring L2 writers' collaborative revision interactions and their writing performance. *System*, 44:101–114. (Cited on page 24)

Harris, J. (2003). Revision as a critical practice. *College English*, 65(6):577–592. (Cited on page 20)

Hasan Dalip, D., André Gonchaptercalves, M., Cristo, M., Calado, P., Dalip, D. H., and Gonchaptercalves, M. A. (2009). Automatic Quality Assessment of Content Created Collaboratively by Web Communities. In *Proceedings of the Joint International Conference on Digital Libraries*, pages 295–304, Austin, TX, USA. (Cited on pages 57 and 58)

Hayes, J. R. (2004). What Triggers Revision? In Allal, L., Chanquoy, L., and Largy, P., editors, *Revision Cognitive and Instructional Processes*, volume 13 of *Studies in Writing*, pages 9–20. Springer Netherlands. (Cited on page 149)

He, J., Tan, A.-H., Tan, C.-L., and Sung, S.-Y. (2004). On Quantitative Evaluation of Clustering Systems. In Wu, W., Xiong, H., and Shekhar, S., editors, *Clustering and Information Retrieval*, pages 105–133. Springer US. (Cited on page 112)

Heckel, P. (1978). A technique for isolating differences between files. *Communications of the ACM*, 21(4):264–268. (Cited on page 63)

Heidorn, G. (2000). Intelligent Writing Assistance. In Dale, R., Moisl, H., and Somers, H., editors, *Handbook of Natural Language Processing*, pages 181–207. Marcel Dekker, Inc. (Cited on page 21)

Herrington, A. J. and Cadman, D. (1991). Peer Review and Revising in an Anthropology Course: Lessons for Learning. *College Composition and Communication*, 42(2):184–199. (Cited on page 22)

Hildick, W. (1965). *Word for word: the rewriting of fiction.* Norton, New York. (Cited on page 18)

Hirate, Y. and Yamana, H. (2006). Generalized Sequential Pattern Mining with Item Intervals. *Journal of Computers*, 1(3):51–60. (Cited on page 100)

Horning, A. and Becker, A. (2006). *Revision: History, Theory, and Practice.* Parlor Press LLC. (Cited on page 20)

Hyland, K. and Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, 39(02):83. (Cited on page 20)

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall, Inc. (Cited on page 115)

Jamison, E. and Gurevych, I. (2014). Needle in a Haystack: Reducing the Costs of Annotating Rare-Class Instances in Imbalanced Datasets. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*, pages 244–253, Phuket,Thailand. (Cited on page 131)

Javanmardi, S., McDonald, D. W., and Lopes, C. V. (2011). Vandalism Detection in Wikipedia: A High-Performing, Feature-Rich Model and its Reduction Through Lasso. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 82–90, Mountain View, CA, USA. (Cited on pages 52, 77 and 79)

Jensen, G. H. and DiTiberio, J. K. (1984). Personality and Individual Writing Processes. *College Composition and Communication*, 35(3):285–300. (Cited on page 149)

Johansen, R. (1988). *Groupware: Computer support for business teams*. The Free Press. (Cited on page 27)

Johnson, D. W., Johnson, R. T., and Holubec, E. J. (1994). *Cooperative Learning in the Classroom*. Association for Supervision and Curriculum Development. (Cited on page 23)

Jones, J. (2008). Patterns of Revision in Online Writing: A Study of Wikipedia's Featured Articles. *Written Communication*, 25(2):262–289. (Cited on pages 53, 58, 68, 97 and 98)

Jones, S. L. (2005). From Writers to Information Coordinators: Technology and the Changing Face of Collaboration. *Journal of Business and Technical Communication*, 19(4):449–467. (Cited on pages 22, 28 and 29)

Kallass, K. (2012). *Schreibprozesse in der Wikipedia: Eine linguistische Analyse*. PhD thesis, Universität Koblenz-Landau. (Cited on page 4)

Kaltenbrunner, A. and Laniado, D. (2012). There is No Deadline - Time Evolution of Wikipedia Discussions. In *Proceedings of the Annual International Symposium on Wikis and Open Collaboration*, Linz, Austria. (Cited on page 130)

Kane, G., Johnson, J., and Majchrzak, A. (2014). Emergent Life Cycle: The Tension Between Knowledge Change and Knowledge Retention in Open Online Coproduction Communities. *Management Science*, 60(12):3026 – 3048. (Cited on pages 106 and 108)

Kim, H.-C. and Eklundh, K. S. (2001). Reviewing Practices in Collaborative Writing. *Computer Supported Cooperative Work*, 10(2):247–259. (Cited on page 38)

Kirby, A. and Rodden, T. (1995). Contact: Support for Distributed Cooperative Writing. In *Proceedings of the Fourth European Conference on Computer-Supported Cooperative Work*,

pages 101–116, Stockholm, Sweden. (Cited on page 38)

Kittur, A., Chi, E., and Suh, B. (2009a). What's in Wikipedia? Mapping Topics and Conflict Using Socially Annotated Category Structure. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1509–1512, Boston, MA, USA. (Cited on page 109)

Kittur, A., Chi, E. H., Pendleton, B., Suh, B., and Mytkowicz, T. (2007a). Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Proceedings of the 25th Annual ACM Conference on Human Factors in Computing Systems*, San Jose, CA, USA. (Cited on pages 57 and 106)

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*, pages 37–46, San Diego, CA, USA. (Cited on pages 32, 49, 57, 58 and 130)

Kittur, A. and Kraut, R. E. (2010). Beyond Wikipedia: coordination and conflict in online production groups. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, pages 215–224, Savannah, GA, USA. ACM. (Cited on pages 54 and 55)

Kittur, A., Lee, B., and Kraut, R. E. (2009b). Coordination in collective intelligence. In *Proceedings of the 27th international Conference on Human Factors in Computing Systems*, pages 1495–1504, Boston, MA, USA. (Cited on page 90)

Kittur, A., Suh, B., Pendleton, B. A., Chi, H., and Chi, E. H. (2007b). He Says, She Says: Conflict and Coordination in Wikipedia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 453–462, San Jose, CA, USA. (Cited on pages 46, 52 and 54)

Kriplean, T., Beschastnikh, I., and McDonald, D. W. (2008). Articulations of wikiwork: uncovering valued work in wikipedia through barnstars. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, pages 47–56, San Diego, CA, US. (Cited on pages 49, 54 and 91)

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage Publications, 2nd edition. (Cited on pages 74, 86, 91 and 134)

Krumov, L., Fretter, C., Müller-Hannemann, M., Weihe, K., and Hütt, M. T. (2011). Motifs in co-authorship networks and their relation to the impact of scientific publications. *The European Physical Journal B*, 84(4):535–540. (Cited on page 36)

Lange, T., Roth, V., Braun, M. L., and Buhmann, J. M. (2004). Stability-Based Validation of Clustering Solutions. *Neural computation*, 16(6):1299–1323. (Cited on pages 115 and 116)

Laniado, D. and Tasso, R. (2011). Co-authorship 2.0: Patterns of collaboration in Wikipedia. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, pages 201–210, Eindhoven, Netherlands. (Cited on pages 56 and 153)

Laniado, D., Tasso, R., Kaltenbrunner, A., Milano, P., and Volkovich, Y. (2011). When the Wikipedians Talk : Network and Tree Structure of Wikipedia Discussion Pages. In *Proceedings of the 5th International Conference on Weblogs and Social Media*, pages 177–184, Barcelona, Spain. (Cited on pages 44, 56 and 128)

Leland, M. D. P., Fish, R. S., and Kraut, R. E. (1988). Collaborative document production using quilt. In *Proceedings of the 1988 ACM Conference on Computer Supported Cooperative Work*, pages 206–215, Portland, OR, USA. (Cited on pages 26 and 28)

Leuf, B. and Cunningham, W. (2001). *The Wiki Way: Quick Collaboration on the Web*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. (Cited on page 40)

Lipka, N. and Stein, B. (2010). Identifying featured articles in Wikipedia. In *Proceedings of the 19th International Conference on World Wide Web*, pages 1147–1148, Raleigh, NC, USA. (Cited on page 58)

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):11. (Cited on pages 5, 53, 54, 56, 68, 98, 107, 108, 109, 110, 111, 112, 113, 115, 120, 122, 148 and 149)

Liu, S. M. and Chen, J.-H. (2015). An empirical study of empty prediction of multi-label classification. *Expert Systems with Applications*, 42(13):5567–5579. (Cited on pages 84 and 158)

Lowry, P. B., Curtis, A., and Lowry, M. R. (2004). Building a Taxonomy and Nomenclature of Collaborative Writing to Improve Interdisciplinary Research and Practice. *Journal of Business Communication*, 41(1):66–99. (Cited on pages 3, 22, 23, 24, 25, 26, 27, 28, 50 and 148)

Lowry, P. B. and Nunamaker, J. F. (2003). Using Internet-Based, Distributed Collaborative Writing Tools to Improve Coordination and Group Awareness in Writing Teams. *IEEE Transactions on Professional Communication*, 46(4):277–297. (Cited on pages 28, 29 and 59)

Lyon, C., Barrett, R., and Malcolm, J. (2004). A theoretical basis to the automated detection of copying between texts, and its practical implementation in the Ferret plagiarism and collusion detector. In *Plagiarism: Prevention, Practice and Policy Conference*, Newcastle, UK. (Cited on page 136)

Mabroukeh, N. R. and Ezeife, C. I. (2010). A taxonomy of sequential pattern mining algorithms. *ACM Computing Surveys*, 43(1):1–41. (Cited on page 35)

Madjarov, G., Kocev, D., Gjorgjevikj, D., and Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104. (Cited on pages 81 and 160)

Mahlow, C. and Piotrowski, M. (2008). Linguistic Support for Revising and Editing. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, pages 631–642. Springer Berlin Heidelberg, Berlin, Heidelberg. (Cited on page 21)

Majchrzak, A., Wagner, C., and Yates, D. (2006). Corporate Wiki Users: Results of a Survey. In *Proceedings of the 2006 International Symposium on Wikis*, pages 99–104, Odense, Denmark. (Cited on page 59)

Majchrzak, A., Wagner, C., and Yates, D. (2013). The Impact of Shaping on Knowledge Reuse for Organizational Improvement with Wikis. *Mis Quarterly*, 37(2):455–469. (Cited on pages 106 and 108)

Marttunen, M. and Laurinen, L. (2012). Participant profiles during collaborative writing. *Journal of Writing Research*, 4(1):53–79. (Cited on pages 25, 26 and 126)

Max, A. and Wisniewski, G. (2010). Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History. In *Proceedings of the 7th Conference on International Language Resources and Evaluation*, Valletta, Malta. (Cited on pages 52 and 101)

McCallum, A., Wang, X., and Corrada-Emmanuel, A. (2007). Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. *Journal of Artificial Intelligence Research*, 30(1):249–272. (Cited on page 149)

McNally, K., O'Mahony, M. P., and Smyth, B. (2013). A comparative study of collaboration-based reputation models for social recommender systems. *User Modeling and User-Adapted Interaction*, 24(3):219–260. (Cited on page 39)

Merton, R. (1968). *Social Theory and Social Structure*. Free Press. (Cited on page 44)

Mesgari, M., Okoli, C., Mehdi, M., Nielsen, F. Å., and Lanamäki, A. (2015). "The sum of all human knowledge": A systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology*, 66(2):219–245. (Cited on page 41)

Mihalcea, R. and Nastase, V. (2002). Letter level learning for language independent diacritics restoration. In *Proceedings of the 6th Conference on Natural Language Learning*, volume 20, Morristown, NJ, USA. (Cited on page 85)

Mizumoto, T., Komachi, M., Nagata, M., and Matsumoto, Y. (2011). Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155, Chiang Mai, Thailand. (Cited on page 20)

Moran, A., Favela, J., Martinez, A., and Decouchant, D. (2001). Document presence notification services for collaborative writing. In *Proceedings Seventh International Workshop on Groupware*, pages 125–133, Darmstadt, Germany. (Cited on page 39)

Morgan, J. T., Gilbert, M., McDonald, D. W., and Zachry, M. (2014). Editing Beyond Articles: Diversity and Dynamics of Teamwork in Open Collaborations. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social*

Computing, pages 550–563, Baltimore, MD, USA. (Cited on page 46)

Moskaliuk, J., Kimmerle, J., and Cress, U. (2009). Wiki-supported learning and knowledge building: effects of incongruity between knowledge and information. *Journal of Computer Assisted Learning*, 25(6):549–561. (Cited on page 24)

Murray, D. M. (1978a). Internal revision: A process of discovery. In Cooper, C. R. and Odell, L., editors, *Research on composing: Points of departure*, pages 85–104. (Cited on pages 15 and 18)

Murray, D. M. (1978b). Teach the Motivating Force of Revision. *The English Journal*, 67(7):56–60. (Cited on page 20)

Myers, E. W. (1986). An O(ND) difference algorithm and its variations. *Algorithmica*, 1(1-4):251–266. (Cited on page 63)

Myhill, D. and Jones, S. (2007). More Than Just Error Correction: Students' Perspectives on Their Revision Processes During Writing. *Written Communication*, 24(4):323–343. (Cited on page 149)

Nelken, R. and Yamangil, E. (2008). Mining Wikipedia's Article Revision History for Training Computational Linguistics Algorithms. In *Proceedings of the 1st AAAI Workshop on Wikipedia and Artificial Intelligence*, pages 31–36, Chicago, IL, USA. (Cited on pages 51 and 101)

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409. (Cited on page 36)

Noël, 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. (Cited on pages 23, 25 and 29)

Nov, O. (2007). What motivates Wikipedians? *Communications of the ACM*, 50(11):60–64. (Cited on page 44)

Nunes, S., Ribeiro, C., and David, G. (2011). Term weighting based on document revision history. *Journal of the American Society for Information Science and Technology*, 62(12):2471–2478. (Cited on page 52)

Ogren, P. and Bethard, S. (2009). Building Test Suites for UIMA Components. In *Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 1–4, Boulder, CO, USA. (Cited on page 161)

O'Mahony, S. and Ferraro, F. (2007). The emergence of governance in an open source community. *Academy of Management Journal*, 50(5):1079–1106. (Cited on page 39)

Onrubia, J. and Engel, A. (2009). Strategies for collaborative writing and phases of knowledge construction in CSCL environments. *Computers & Education*,

53(4):1256–1265. (Cited on pages 4, 23, 24, 25, 26 and 57)

Ortega, F., Gonzalez-Barahona, J. M., and Robles, G. (2008). On the Inequality of Contributions to Wikipedia. In *Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, Waikoloa, HI, USA. (Cited on page 109)

Passig, D. and Schwartz, G. (2007). Collaborative Writing: Online Versus Frontal. *International Journal on E-Learning*, 6(3):395–412. (Cited on page 29)

Passonneau, R. J. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy. (Cited on pages 74, 86 and 91)

Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., Dayal, U., and Hsu, M.-C. (2004). Mining sequential patterns by pattern-growth: the PrefixSpan approach. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1424–1440. (Cited on page 100)

Perl, S. (1979). The composing processes of unskilled college writers. *Research in the Teaching of English*, 13(4):317–336. (Cited on page 19)

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. (Cited on pages 53 and 68)

Philippe Fournier-Viger, Nkambou, R., Fournier-Viger, P., and Nguifo, E. M. (2008). A Knowledge Discovery Framework for Learning Task Models from User Interactions in Intelligent Tutoring Systems. In *Proceedings of the 7th Mexican International Conference on Artificial Intelligence*, pages 765–778, Atizapán de Zaragoza, Mexico. (Cited on page 100)

Platt, J. C. (1998). Fast training of support vector machines using sequential minimal optimization. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in Kernel Methods: Support Vector Learning*, pages 185–208. (Cited on pages 116 and 138)

Posner, I. and Baecker, R. (1992). How people write together. In *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences*, pages 127–138, Kauai, HI, USA. (Cited on pages 23, 24, 26, 27, 28 and 29)

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. (Cited on page 52)

Potthast, M. and Holfeld, T. (2011). Overview of the 2nd International Competition on Wikipedia Vandalism Detection. In *Notebook Papers of CLEF 2011 Labs and Workshops*, Amsterdam, Netherlands. (Cited on page 52)

Priedhorsky, R., Chen, J., Lam, S. K., Panciera, K., Terveen, L., and Riedl, J. (2007). Creating, destroying, and restoring value in Wikipedia. In *Proceedings of the 2007*

International ACM Conference on Supporting Group Work, pages 259–268, Sanibel Island, FL, USA. (Cited on pages 57 and 106)

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers. (Cited on pages 81, 93, 116 and 159)

Ransbotham, S., Kane, G. C., and Lurie, N. H. (2012). Network Characteristics and the Value of Collaborative User-Generated Content. *Marketing Science*, 31(3):387–405. (Cited on page 154)

Reagle, J. and Rhue, L. (2011). Gender Bias in Wikipedia and Britannica. *International Journal of Communication*, 5:1138–1158. (Cited on page 44)

Reagle, J. M. J. (2010). *Good Faith Collaboration: The Culture of Wikipedia*. The MIT Press. (Cited on page 46)

Recasens, M., Danescu-Niculescu-Mizil, C., and Jurafsky, D. (2013). Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics*, pages 1650–1659, Sofia, Bulgaria. (Cited on pages 52 and 101)

Reynolds, S., Woolley, B., and Woolley, T. (1911). *Seems So! A Working-Class View of Politics*. MacMillan London. (Cited on page 30)

Rice, R. P. and Huguley J.T., J. (1994). Describing collaborative forms: a profile of the team-writing process. *IEEE Transactions on Professional Communication*, 37(3):163–170. (Cited on page 23)

Rimmershaw, R. (1992). Collaborative writing practices and writing support technologies. *Instructional Science*, 21(1-3):15–28. (Cited on pages 22 and 28)

Rohman, D. G. (1965). Pre-Writing the Stage of Discovery in the Writing Process. *College Composition and Communication*, 16(2):106–112. (Cited on page 15)

Rosner, M. (1992). Engineered Revisions in Industry. In Charney, D. and Ebbitt, W. R., editors, *Constructing Rhetorical Education*, pages 318–329. SIU Press. (Cited on page 151)

Rzeszotarski, J. and Kittur, A. (2012). Learning from history: Predicting reverted work at the word level in Wikipedia. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 437–440, Seattle, WA, USA. (Cited on pages 52 and 54)

Sanger, L. (2005). The Early History of Nupedia and Wikipedia: A memoir. In DiBona, C., Stone, M., and Cooper, D., editors, *Open Sources 2.0: The Continuing Evolution*, pages 307–338. (Cited on page 40)

Scardamalia, M. and Bereiter, C. (1985). Development of Dialectical Processes in Composition. In Olson, D. R., Torrance, N., and Hildyard, A., editors, *Language, Literacy, and Learning: the nature and consequences of reading and writing*, pages

307–329. (Cited on page 21)

Scardamalia, M. and Bereiter, C. (1994). Computer Support for Knowledge-Building Communities. *Journal of the Learning Sciences*, 3(3):265–283. (Cited on pages 21 and 23)

Scardamalia, M. and Bereiter, C. (2003). Knowledge Building. In Guthrie, J., editor, *Encyclopedia of Education*, pages 1370–1373. (Cited on pages 23 and 24)

Schneider, J., Passant, A., and Breslin, J. G. (2010). A Content Analysis: How Wikipedia Talk Pages Are Used. In *Proceedings of the 2nd International Conference of Web Science*, pages 1–7, Raleigh, NC, USA. (Cited on pages 55 and 129)

Segall, J. and Greenstadt, R. (2013). The Illiterate Editor: Metadata-driven Revert Detection in Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration*, Hong Kong, China. (Cited on page 52)

Sepehri Rad, H., Makazhanov, A., Rafiei, D., and Barbosa, D. (2012). Leveraging editor collaboration patterns in Wikipedia. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, pages 13–22, Milwaukee, WI, USA. (Cited on page 56)

Sharples, M. (1993). *Computer Supported Collaborative Writing*. Springer. (Cited on page 28)

Sharples, M., Goodlet, J. S., Beck, E. E., Wood, C. C., Easterbrook, S. M., and Plowman, L. (1993). Research Issues in the Study of Computer Supported Collaborative Writing. In Sharples, M., editor, *Computer Supported Collaborative Writing*, pages 9–28. (Cited on pages 21, 22, 26, 28, 29, 30, 46 and 59)

Shermis, M. D. and Burstein, J. C. (2003). *Automated Essay Scoring: A Cross-disciplinary Perspective*. Lawrence Erlbaum Associates, Inc. (Cited on page 151)

Simon, P. (2013). *Too Big to Ignore: The Business Case for Big Data*. John Wiley & Sons Inc. (Cited on page 21)

Sommers, N. (1980). Revision strategies of student writers and experienced adult writers. *College composition and communication*, 31(4):378–388. (Cited on pages 14, 17, 18, 19, 20 and 98)

Stallard, C. K. (1974). An Analysis of the Writing Behavior of Good Student Writers. *Research in the Teaching of English*, 8(2):206–218. (Cited on page 18)

Steiner, T. (2014). Bots vs. Wikipedians, Anons vs. Logged-Ins (Redux). In *Proceedings of The International Symposium on Open Collaboration*, Berlin, Germany. ACM Press. (Cited on page 49)

Storch, N. (2005). Collaborative writing: Product, process, and students' reflections. *Journal of Second Language Writing*, 14(3):153–173. (Cited on pages 20 and 24)

Stratton, C. (1989). Collaborative writing in the workplace. *IEEE Transactions on Professional Communication*, 32(3):178–182. (Cited on page 27)

Stvilia, B., Gasser, L., Twidale, M. B., and Smith, L. C. (2007). A Framework for Information Quality Assessment. *Journal of the American Society for Information Science*, 58(12):1720–1733. (Cited on page 57)

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. (Cited on pages 44, 57, 58 and 105)

Suh, B., Convertino, G., Chi, E. H., and Pirolli, P. (2009). The singularity is not near: slowing growth of Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, Orlando, FL, USA. (Cited on page 46)

Sumi, R., Yasseri, T., Rung, A., Kornai, A., and Kertesz, J. (2011). Edit Wars in Wikipedia. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 724–727, Boston, MA, USA. (Cited on page 52)

Tam, J. and Greenberg, S. (2006). A framework for asynchronous change awareness in collaborative documents and workspaces. *International Journal of Human-Computer Studies*, 64(7):583–598. (Cited on page 39)

Tapscott, D. and Williams, A. (2008). *Wikinomics: How Mass Collaboration Changes Everything*. Portfolio Trade. (Cited on pages 29, 48 and 151)

Trohidis, K., Tsoumakas, G., Kalliris, G., and Vlahavas, I. (2008). Multi-label classification of music into emotions. In *9th International Conference on Music Information Retrieval*, pages 325–330, Philadelphia, PA, USA. (Cited on page 81)

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2008). Effective and Efficient Multilabel Classification in Domains with Large Number of Labels. In *Proceedings of the ECML/PKDD 2008 Workshop on Mining Multidimensional Data*, Antwerp, Belgium. (Cited on pages 80 and 88)

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. (Cited on pages 74, 75 and 78)

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2011). Random k-Labelsets for Multi-Label Classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1079–1089. (Cited on pages 81 and 88)

Tuzi, F. (2004). The impact of e-feedback on the revisions of L2 writers in an academic writing course. *Computers and Composition*, 21(2):217–235. (Cited on page 22)

Vacc, N. N. (1986). Word Processor versus Handwriting: A Comparative Study of Writing Samples Produced by Mildly Mentally Handicapped Students. *Exceptional Children*, 54(2):156–65. (Cited on page 13)

Viégas, F. B., Wattenberg, M., and Dave, K. (2004). Studying Cooperation and Conflict Between Authors with History Flow Visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 575–582, Vienna, Austria. (Cited on pages 52 and 148)

Viégas, F. B., Wattenberg, M., Kriss, J., and Ham, F. (2007a). Talk Before You Type: Coordination in Wikipedia. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, pages 78–78, Big Island, HI, USA. (Cited on pages 41, 54, 129 and 155)

Viégas, F. B., Wattenberg, M., and McKeon, M. M. (2007b). The Hidden Order of Wikipedia. In Schuler, D., editor, *Online Communities and Social Computing*, pages 445–454. Springer Berlin Heidelberg. (Cited on pages 31 and 32)

Wang, H., Tudorache, T., Dou, D., Noy, N. F., and Musen, M. A. (2014). Analysis and Prediction of User Editing Patterns in Ontology Development Projects. *Journal on Data Semantics*, 4(2):117–132. (Cited on page 99)

Wang, L. and Cardie, C. (2014). A Piece of My Mind: A Sentiment Analysis Approach for Online Dispute Detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 693–699. (Cited on page 54)

Warncke-Wang, M., Ayukaev, V. R., Hecht, B., and Terveen, L. (2015). The Success and Failure of Quality Improvement Projects in Peer Production Communities. In *Proceedings of the 18th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, pages 743–756, Vancouver, BC, Canada. (Cited on page 57)

Warncke-Wang, M., Cosley, D., and Riedl, J. (2013). Tell Me More: An Actionable Quality Model for Wikipedia. In *Proceedings of the International Symposium on Wikis and Open Collaboration*, Hong Kong, China. (Cited on page 58)

Welser, H., Cosley, D., and Kossinets, G. (2011). Finding Social Roles in Wikipedia. In *Proceedings of the 2011 iConference*, pages 122–129, Seattle, WA, USA. (Cited on pages 44, 106, 107, 123 and 149)

Wichmann, A. and Rummel, N. (2013). Improving revision in wiki-based writing: Coordination pays off. *Computers & Education*, 62:262–270. (Cited on pages 20 and 23)

Wilkinson, D. M. and Huberman, B. A. (2007). Cooperation and Quality in the Wikipedia. In *Proceedings of the International Symposium on Wikis and Open Collaboration*, pages 157–164, Montreal, Canada. (Cited on pages 57, 58 and 106)

Wöhner, T., Köhler, S., and Peters, R. (2011). Automatic Reputation Assessment in Wikipedia. In *Proceedings of the International Conference on Information Systems*, Shanghai, China. (Cited on page 49)

Wöhner, T. and Peters, R. (2009). Assessing the quality of Wikipedia articles with lifecycle based metrics. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, Orlando, FL, USA. (Cited on page 57)

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. (Cited on page 51)

Wu, G., Harrigan, M., and Cunningham, P. (2011). Characterizing Wikipedia Pages Using Edit Network Motif Profiles. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management*, pages 45–51, Glasgow, Scotland, UK. (Cited on page 56)

Xue, H. and Hwa, R. (2010). Syntax-driven machine translation as a model of ESL revision. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1373–1381, Beijing, China. (Cited on page 20)

Yagelski, R. P. (1995). The Role of Classroom Context in the Revision Strategies of Student Writers. *Research in the Teaching of English*, 29(2):216–238. (Cited on page 5)

Yamangil, E. and Nelken, R. (2008). Mining Wikipedia Revision Histories for Improving Sentence Compression. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Short Papers*, pages 137–140, Columbus, OH, USA. (Cited on pages 51 and 101)

Yang, H.-L. and Lai, C.-Y. (2010). Motivations of Wikipedia content contributors. *Computers in Human Behavior*, 26(6):1377–1383. (Cited on page 44)

Yasseri, T., Sumi, R., Rung, A., Kornai, A., and Kertész, J. (2012). Dynamics of conflicts in Wikipedia. *PloS one*, 7(6). (Cited on pages 52 and 130)

Yates, D., Wagner, C., and Majchrzak, A. (2010). Factors Affecting Shapers of Organizational Wikis. *Journal of the American Society for Information Science and Technology*, 61(3):543–554. (Cited on pages 106 and 108)

Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., and Lee, L. (2010). For the Sake of Simplicity: Unsupervised Extraction of Lexical Simplifications from Wikipedia. In *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 365–368, Los Angeles, CA, USA. (Cited on page 51)

Zamel, V. (1982). Writing: The Process of Discovering Meaning. *TESOL Quarterly*, 16(2):195. (Cited on pages 14, 15 and 20)

Zamel, V. (1983). The Composing Processes of Advanced ESL Students: Six Case Studies. *TESOL Quarterly*, 17(2):165. (Cited on page 20)

Zanzotto, F. M. and Pennacchiotti, M. (2010). Expanding textual entailment corpora from Wikipedia using co-training. In *Proceedings of the COLING-Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 28–36, Beijing, China. (Cited on page 51)

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. (Cited on pages 52 and 101)

Zesch, T., Müller, C., and Gurevych, I. (2008). Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 861–866, Chicago, IL, USA. (Cited on pages 62, 78 and 127)

Zhang, H., Zhang, S., Wu, Z., Huang, L., and Ma, Y. (2014). Predicting Wikipedia Editor's Editing Interest Based on Factor Graph Model. In *2014 IEEE International Congress on Big Data*, pages 382–389, Anchorage, AK, USA. (Cited on page 122)

Index

Ehrenwörtliche Erklärung[†]

Hiermit erkläre ich, die vorgelegte Arbeit zur Erlangung des akademischen Grades "Dr.-Ing." mit dem Titel "The Writing Process in Online Mass Collaboration: NLP-Supported Approaches to Analyzing Collaborative Revision and User Interaction" selbständig und ausschließlich unter Verwendung der angegebenen Hilfsmittel erstellt zu haben. Ich habe bisher noch keinen Promotionsversuch unternommen.

Darmstadt, den 28. Mai 2015

Johannes Daxenberger, M.A.

[†] Gemäß § 9 Abs. 1 der Promotionsordnung der TU Darmstadt

Wissenschaftlicher Werdegang des Verfassers[‡]

2006–2011	Magisterstudium Hauptfach: Sprachliche Informationsverarbeitung Nebenfächer: Romanistik (Spanisch), Allgemeine Sprachwissenschaft Universität zu Köln
Juni 2011	Abschluss als Magister Artium (M.A.) Magisterarbeit: „Automatische Klassifikation von Sprechakten unter besonderer Berücksichtigung syntaktischer Eigenschaften. Implementation und Analyse in einem komponentenbasierten System.“ Gutachter: Prof. Dr. Jürgen Rolshoven
2011–2014	Promotionsstipendium im Rahmen der hessischen Exzellenzinitiative „Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz“ am Ubiquitous Knowledge Processing Lab Technische Universität Darmstadt
seit Ende 2014	Wissenschaftlicher Mitarbeiter am Ubiquitous Knowledge Processing Lab Technische Universität Darmstadt

[‡] Gemäß § 20 Abs. 3 der Promotionsordnung der TU Darmstadt

Publikationsverzeichnis des Verfassers

Habernal, I., **Daxenberger, J.**, and Gurevych, I. (2016). Mass Collaboration on the Web: Textual Content Analysis by Means of Natural Language Processing. In Cress, U., Moskaliuk, J., and Jeong, H., editors, *Mass Collaboration and Education*, in press, Springer International Publishing.

Daxenberger, J., Ferschke, O., Gurevych, I., and Zesch, T. (2014). A Java-based Framework for Supervised Learning Experiments on Textual Data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pages 61–66, Baltimore, MD, USA.

Daxenberger, J. and Gurevych, I. (2014). Automatically Detecting Corresponding Edit-Turn-Pairs in Wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Short Papers*, pages 187–192, Baltimore, MD, USA.

Ferschke, O., **Daxenberger, J.**, and Gurevych, I. (2013). 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*, Chapter 5, pages 121–160, Springer.

Daxenberger, J., and Gurevych, I. (2013). Automatically Classifying Edit Categories in Wikipedia Revisions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 578–589, Seattle, WA, USA.

Daxenberger, J., and Gurevych, I. (2012). A Corpus-Based Study of Edit Categories in Featured and Non-Featured Wikipedia Articles. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 711–726, Mumbai, India.

Daxenberger, J. (2011). Automatische Klassifikation von Sprechakten unter besonderer Berücksichtigung syntaktischer Eigenschaften. Implementation und Analyse in einem komponentenbasierten System. M.A. thesis, Universität zu Köln.