Originally published in: Cotgrove, Louis/Herzberg, Laura/Lüngen, Harald/Pisetta, Ines (Eds.): Proceedings of the 10th International Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2023), 14–15 September 2023, University of Mannheim, Germany. - Mannheim: Leibniz-Institut für Deutsche Sprache, pp. 131-135
DOI: https://doi.org/10.14618/1z5k-pb25

Studying the distribution of reply relations in Wikipedia talk pages

Abstract

This paper presents an extended annotation and analysis of interpretative reply relations focusing on a comparison of reply relation types and targets between conflictual pages and neutral pages of German Wikipedia (WP) talk pages. We briefly present the different categories identified for interpretative reply relations to analyze the relationship between WP postings as well as linguistic cues for each category. We investigate referencing strategies of WP authors in discussion page postings, illustrated by means of reply relation types and targets taking into account the degree of disagreement displayed on a WP talk page. We provide richly annotated data that can be used for further analyses such as the identification of interactional relations on higher levels, or for training tasks in machine learning algorithms.

Keywords: Wikipedia talk pages, reply relations, referencing strategies

1. Introduction

This paper presents an extended annotation and analysis of reply relation types and targets in Wikipedia (WP) talk pages focusing on the investigation of reply relation (RR) types as well as target locations in Wikipedia talk pages containing different levels of disagreement.

Reply relations are a special kind of interactional relations that hold between postings, i.e. user contributions on a talk page. When Wikipedia authors communicate with each other on a talk page, a set of reply relations between postings on a talk page arises by the fact that the content of (one or more) previous posting(s) is directly addressed. Because computer-mediated communication (CMC) interactions vary based on genre and topic, reply relations are not limited to question-answer patterns. Reply relations involve any response or reaction that occurs when two authors communicate with one other.

We think that the annotation of reply relations is emerging as a promising method to reconstruct interaction structures on article discussion pages. By identifying reply relations, Wikipedia's usual convention of indentation can be substantiated or corrected if necessary, resulting in a more accurate representation of the underlying discussion structure.

Cleanup

> Does this complete the article cleanup? **Smack** (or others): you decide. -- hike395 02:16, 31 May 2005 (UTC)

It looks pretty good. […] --**Smack** (talk) 19:24, 31 May 2005 (UTC)

Example 1[1]: Interpretative reply relation, type: addressing, linguistic cue - username "Smack".

Indentations are a formal means to express reply relations in WP talk pages. The term *interpretative reply relations* refers to all reply action that is not realized formally but signaled by other structural or linguistic means.

In Example 1, the reply action is not realized formally by technical reply or indentation, but signaled by different linguistic means, e.g. via addressing, as in this case the use of *Smack* by author hike395. We refer to such indicators as *linguistic cues*. The term *reply target* denotes a previous posting which is being referred to by the current posting. In Example 1, author Smack refers to hike395' posting which means that hike395's posting is the reply target of Smack's answer. Besides different forms of addressing, such as in Example 1, Q-A structures, or quotes can be considered as cues for interpretative reply relation types. For the presentation, we will summarize our taxonomy of interpretative reply relation types which abstracts overgroups of linguistic cues in talk page postings.

2. Wikipedia talk pages

We aim to provide enriched CMC data by annotating reply relations between postings on Wikipedia talk pages. There are different strategies when trying to reconstruct reply relations, such as focusing on the microstructure (i.e. internal structure) of postings by annotating speech acts, as in Ferschke et al. 2012. We take a closer look at the mesostructure, i.e., the relation *between* postings, and build upon Lüngen/Herzberg (2019). By annotating reply relations between postings, we make explicit that a posting most of the time represents a reply to a previous posting and also to which posting exactly. In Wikipedia talk pages, the primary order structure is termed a *thread* (cf. Beißwenger et al. 2012). A thread contains a variable number of postings which the authors group thematically under different headings, such as "Cleanup" in Example 1. Wikipedia authors are requested to indent their contributions on the wiki page to build thread structures as known from other discussion forums. As a result, the amount of indentation is a property of the posting rather than something imposed by the server (cf. Beißwenger et al. 2012).

3. Linguistic Annotation

Identifying and annotating aspects such as the addressing cue in Example 1 is a first step to reconstructing the reply sequences in Wikipedia discussions. For a complete analysis, one would have to take into account the articles, the revision histories (of articles and talk pages), and the linked pages as well. The approach is a step towards the representation of interaction structures in CMC corpora, which will also allow for quantitative studies, similar to speech act annotations in speech corpora.

3.1 Research Questions

The annotation process addressed several goals. After demonstrating subtypes of reply relations and the reply strategies that occur within the extensive background of a Wikipedia talk page, we wanted to focus on the distribution of these reply types and strategies taking into account the degree of disagreement displayed on a WP talk page.

[1] https://en.wikipedia.org/wiki/Talk:Hiking/Archive_1.

DOI: https://doi.org/10.14618/1z5k-pb25

Therefore, the research questions are as follows:

RQ1: Do conflictual pages and neutral pages differ in the distribution of reply targets?
- RQ1a: Where in relation to the replying posting is the reply target posting located in conflictual vs. in neutral pages?
- RQ1b: How frequently does the reply target annotated actually match the one indicated by the indentation, i.e. the "parent" posting exactly one indentation level higher? And if the annotated reply target is not the parent posting, is it then the immediately preceding posting? Do the respective figures in the conflictual pages differ significantly from the neutral pages?
- RQ1c: How frequently is the reply target to be found in a different thread altogether in conflictual vs. in neutral pages?

RQ2: Do conflictual pages and neutral pages differ in the distribution of reply type categories?
- RQ2a: Which reply relation type can be identified for each subcorpus?
- RQ2b: Which reply relation type occurs most often in each subcorpus?

3.2 Data: Two Subcorpora of Wikipedia talk pages

Kittur et al. (2009) have shown in early Wikipedia research that articles of certain categories entail a high potential for disagreement and conflict (cf. Kittur et al. 2009). Categories in Wikipedia serve to group article pages according to certain characteristics. In addition to articles on religion and politics, article pages of the philosophy category and on personalities in particular contain an increased potential for conflict (cf. Kittur et al. 2009, p. 1512; Hara et al. 2010).

The Wikipedia Demo Corpus German Talk Pages Subcorpus[2] (WDC) provides the database of conflicting talk pages. Table 1 displays the WP pages (left column) of the WDC as well as the talk pages of the neutral corpus (right column). The neutral corpus consists of eight WP talk pages of less conflicting categories, such as technology, cities, animals, and represents the comparative, neutral data basis.

WDC; list of annotated conflict-prone WP pages	**Neutral corpus; list of annotated "neutral" WP pages**
Flüchtlingskrise in Europa ab 2015 *(Refugee crisis in Europe from 2015)* Chiropraktik *(Chiropractics)* Wladimir Wladimirowitsch Putin *(Vladimir Putin)* Terroranschläge am 11. September 2001 *(Terrorist attacks on 11 September 2001)* Psychoanalyse *(Psychoanalysis)* Gentechnisch veränderter Organismus *(Genetically modified organism)* Feminismus *(Feminism)* The Legend of Zelda *(The Legend of Zelda)*	Berlin *(Berlin)* Streifenhörnchen *(Chipmunk)* Großer Panda *(Giant panda)* Fernglas *(Binoculars)* Stadtbahn Bonn *(Bonn city rail)* Schwarzweißfotografie *(Black and white photography)* Grammatik *(Grammar)* Wandern *(Hiking)*

Table 1: Data basis: German Wikipedia talk pages, English translations added.

3.3 Annotation process and guidelines

The categories for annotating interpretative reply relations are based on suggested categories mentioned in Lüngen/Herzberg (2019). These suggestions were transformed into a set of nine categories and annotated in three annotation rounds by two encoders[3] for the WDC data set, and in one annotation round for the neutral corpus data set.

For the annotations, simplified I5 versions of the pages devoid of all inline annotations (e.g. *italics*) were prepared, and the annotators used the simplified I5 XML files in the *Oxygen XML Editor* as well as annotation guidelines which explained the attributes and elements accordingly. Three attributes were annotated during the process: @relationTarget, @relationType, and @cueTarget, the latter in combination with the element <cue>.

To finish the annotation process, we adjudicated the annotations of both subcorpora (relation targets, types and cues) to create *master annotations* which constitute a gold standard dataset[4]. We took over the roles of adjudicators, as it is essential "to have adjudicators who were involved in creating the annotation guidelines, as they will have the best understanding of the purpose of the annotation" (Pustejovsky/Stubbs 2013, 134).

4. Results

The results section provides answers to the research questions using the WDC and neutral corpus master files. We extended the results of the WDC annotation process by comparisons to neutral WP sites in order to find out whether the observations made by Kittur et al. (2009) apply to referencing strategies as well.

[2] The *Wikipedia Demo Corpus* was developed within the framework of a multilingual research project and is currently available via the Corpus plattform *KorAP*: https://korap.ids-mannheim.de/instance/wikidemo.

[3] There were different encoders involved in the annotation processes: two encoders annotated the WDC data over a total of three consecutive annotation rounds. As the categories were solidified in the initial annotation process and the annotation guidelines existed in final form, two different encoders got by with fewer rounds of annotation when annotating the neutral corpus. Overall, the process also took less time, which can also be attributed to the different sizes of the subcorpora, with the WDC containing 572,968 tokens and the neutral corpus 21,131 tokens, as well as posting and thread sizes, cf. Table 2.

[4] Implementing this additional step is beneficial when planning to train and test machine learning (ML) algorithms.

We assumed that the more disagreement is displayed in the author's exchanges on a WP talk page, the more complex the referencing between the postings will get, creating long and branched discussion threads.
Before reporting on the results of the research questions, Table 2 presents some descriptive statistics about the sizes of pages, threads and posts in the two subcorpora.

	Conflictual pages	**Neutral pages**
avg #threads by page*	27.24	16.62
avg #posts by page*	278.27	47.50
avg #posts by thread*	10.20	2.86
avg #tokens by page	17,362.67	2,641.38
avg #tokens by post	67.80	54.53

Table 2: Sizes of pages, threads and posts in the two WP subcorpora. The asterisk * symbolizes a significant size difference between the subcorpora[5].

The two subcorpora differ significantly in the size of threads per page, posts per page as well as posts per thread. On the conflictual WDC talk pages there are more threads that contain a larger amount of postings which are longer as well in comparison to the talk pages of the neutral corpus. This confirms our assumption that the greater the amount of displayed disagreement in the author's exchanges on a WP talk page, the longer and more complex discussion threads are emerging.

Reply relation target	**% in conflictual pages**	**% in neutral pages**
Target is in the same thread	99.66	99.12
Target is in a different thread	0.34	0.88
Target is parent posting	66.30	76.79
Target is parent and parent is preceding posting	93.79	96.51
Posting has more than one target	4.83	5.59

Table 3: Distribution of reply type targets in the two WP subcorpora.

Table 3 presents the distribution difference in percent of reply type targets in the two WP subcorpora. When investigating RR targets, we wanted to identify where in relation to the replying posting the reply target posting is located in conflictual in comparison to neutral pages. For both subcorpora, the clear majority of targets is located within the same thread: 99.66 % of annotated targets in the WDC and 99.12 % in the neutral corpus respectively (RQ1a). In 66.30 % of the annotated WDC data, the reply target annotated actually matched the one indicated by the indentation, i.e. the "parent" posting exactly one indentation level higher (RQ1b). In the neutral corpus, the amount is slightly higher with a total of 76.79 % that matched the "parent" posting exactly one indentation level higher. That means that the indentation had been used correctly (i.e. according to the Wikipedia guidelines). We then asked how frequently the reply target is, if not the parent posting, the immediately preceding posting. In 93.79 % of annotated targets in the WDC, the targets corresponded to the immediately preceding posting and were the parent posting at the same time. This result is almost identical in the neutral corpus, with 96.51 % annotated targets that corresponded to the immediately preceding posting and were the parent posting at the same time. Lastly, we analyzed whether the frequency of the reply target to be found in a different thread altogether differs between the subcorpora (RQ1c). Again, both subcorpora show a similar distribution. In around 5 % of postings, reply relations are identified to refer to more than just one other posting, i.e. where several interpretative reply relations that were identified within one posting show replies to more than one previous posting. The reply relations from postings like this can currently not be correctly identified by relying on the indentation only.
The respective figures in the conflictual pages do not differ significantly from the neutral pages. To conclude, the level of disagreement does not lead to a more branched and expanded discussion thread as the distribution of the annotated reply relation target locations does not differ between the subcorpora.
By contrast, the analyses of the RR types distribution revealed differences between the conflictual pages and neutral pages (RQ2). The annotators distinguished between eight reply relation types[6] (RQ2a), cf. the column "Reply relation type" in Table 4, while it was

[5] We used Pearson's chi-square statistic $\chi 2$ to calculate differences in reply relation type distribution between the subcorpora, cf. https://www.socscistatistics.com/tests/chisquare2/default2.aspx. The p-values range between < .00001 (avg #threads by page and avg #posts by thread), and .047415 (avg #posts by page). The results are significant at $p < .05$.

[6] Additionally to the presented eight RR types in Table 4, the category "title-relation" was annotated as well. We do not include it here and in other presented results in this paper because it had been annotated largely automatically.

possible to identify more than just one type per posting.[7] The relation types arise from abstracting over the nature and forms of their linguistic indicators in the postings, cf. Example 1 in which the linguistic cue *Smack* used by author hike395 allows for interpreting the reply relation type "addressing".

Reply relation type	**% in conflictual pages**	**% in neutral pages**
2ndPerson*	28.18	6.09
implied*	23.97	8.12
anaphor*	12.61	21.83
response-token*	12.10	25.89
quoting	9.02	7.11
addressing	8.90	6.60
QA-relation*	5.16	24.37
no relation annotated	0.07	0.00
Sum	100	100

Table 4[8]: Distribution of reply type categories in the two WP subcorpora, sorted by frequencies highest to low of the WDC. The asterisk * symbolizes a significant difference in the RR type distribution between the subcorpora[9].

As results of annotating RR types in the WDC conflict-prone WP talk pages, the two relation types occurring most often are "2ndPerson" and "implied" for both encoders[10] (RQ2b), cf. Table 4. These two types count for over 50% of all reply types assigned by them. Putting this into perspective in terms of the relation type "implied", we can see that for almost a quarter of all relations between postings in the WDC corpus, no specific textual cues could be identified, such as a greeting, "Hi Anna", or a direct request to action, for example in questions like "Can you change...?".

In the neutral subcorpus the two relation types occurring most often are "response-token" and "QA-relation" for both encoders (RQ2b). Comparable to the WDC RR category type results, also two RR types count for over 50% of all reply types assigned. However, in the neutral corpus, the RR type "anaphor" was identified with almost similar frequency, turning the dual lead into a trio.

Interestingly, the aforementioned three RR types "response-token", "QA-relation" and "anaphor" can be identified significantly less often in the WDC. The relations between postings in the WDC arise rather implicitly by interpreting the contents of all participants' postings involved and understanding their connection whereas on the neutral pages, referencing strategies between postings are signaled explicitly.

To conclude, the level of disagreement on a Wikipedia talk page impacts the distribution of RR types, but not RR targets. We could identify that on shorter and content-wise, more neutral talk pages, the distribution of RR types, namely the following five out of eight RR categories: "response-token", "QA-relation", "anaphor", "2ndPerson" and "implied" differs significantly with regard to the conflictuality degree of a talk page. The category "implied" that contains relations established implicitly by the general content of the postings and the readers ability to infer and understand that a reply relation holds without reading a specific linguistic cue, finds proportionately large usage on the conflict-prone WDC talk pages.

Identifying interpretative reply relations helps account for postings whose references cannot be reconstructed via the indentation itself. By annotating @relationType and @relationTarget to identify the posting which another posting refers to, we know the extensiveness of discussion threads and the multipurposeness of one singular posting in which the author can address numerous issues simultaneously. Moreover, the developed reply relation type categories can be applied to a variety of talk pages, regardless of their potential for disagreement.

5. References

Beißwenger, M./Ermakova, M./Geyken, A./Lemnitzer, L./Storrer, A. (2012): A TEI Schema for the Representation of Computer-mediated Communication. In: Journal of the Text Encoding Initiative (Issue 3). https://doi.org/10.4000/jtei.476.

Ferschke, O./Gurevych, I./Chebotar, Y. (2012): Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. Präsentiert auf: EACL 2012,

[7] e.g., relationTarget="p1 p2" relationType="2ndPerson QA-relation" cueTarget="c2 c3" would encode that the posting includes two different types of reply relations, a "2ndPerson" as well as a "QA-relation".

[8] We calculated the frequencies relatively in % to take into account the different subcorpora sizes.

[9] We used Pearson's chi-square statistic $\chi 2$ to calculate differences in reply relation type distribution between the subcorpora, cf. https://www.socscistatistics.com/tests/chisquare2/default2.aspx. The p-values range between < .00001 (2ndPerson, response-token , QA-relation), .000015 (implied) and .001303 (anaphor). The results are significant at $p < .05$.

[10] We calculated an inter-rater agreement between the encoders for eight categories using Cohen's κ. We counted all pairings of relation types at postings with one identical relation target, according to the following rules: label an empty relation type as the type 'NO_REL', if there are one or more identical pairs of relation types, count the first one, and if there is no identical pair of relation types, count the first non-identical pair.
κ for the conflictual pages, i.e, over the sum of all WDC postings, was 0.63; κ for the neutral pages was 0.63 as well. An agreement level between 0.61–0.80 classifies as a substantial agreement (Landis/Koch 1977, 165). This result shows that RR annotations in both subcorpora can be covered substantially well with the developed RR type categories, regardless of the level of disagreement on a talk page.

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: Association for Computational Linguistics. S. 777–786. https://www.aclweb.org/anthology/E12-1079.

Hara, Noriko/Shachaf, Pnina/Hew, Khe Foon (2010): Cross-Cultural Analysis of the Wikipedia Community. In: Journal of the American Society for Infor- mation Science and Technology 61(10), S. 2097–2108. https://doi.org/10.1002/asi.21373.

Kittur, Aniket/Chi, Ed H./Suh, Bongwon (2009): What's in Wikipedia?: Mapping Topics and Conflict Using Socially Annotated Category Structure. Präsentiert auf: CHI '09: CHI Conference on Human Factors in Computing Systems, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Boston MA USA: ACM. S. 1509–1512. https://doi.org/10.1145/1518701.1518930.

Landis, J. R./Koch, G. G. (1977): The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310.

Lüngen, H./Herzberg, L. (2019): Types and annotation of reply relations in computer-mediated communication. In: European Journal of Applied Linguistics 7(2), S. 305–332. https://doi.org/10.1515/eujal-2019-0006.

Pustejovsky, J./Stubbs, Amber (2013): Natural language annotation for machine learning. Sebastopol, CA: O'Reilly Media.

WikiDemo Corpus in *KorAP* (*Corpus analysis platform*). <https://korap.ids-mannheim.de/instance/wikidemo>

Wikipedia. <https://www.wikipedia.de/>