# WikiConv

Building a rich conversation corpus
from Wikipedia Talk pages

Yiqing Hua, Cornell Tech

# Research interests on Wikipedia talk pages

http://thebluepaper.com/article/guest-commentary-grinder-pump-attacks-get-personal/

*Antisocial Behavior*

[Wulczyn et al. 2017]
[Zhang et al. 2018]

https://it.depositphotos.com/12630133/stock-photo-3d-talking-concept-over-white.html

*Disputes*

[Wang and Cardie 2014]

http://www.boerenopeenkruispunt.be/Tips/tabid/105/ArticleID/245/Default.aspx

*Conversational Behaviors*

[Danescu-Niculescu-Mizil 2012]

# Research interests on Wikipedia talk pages

http://thebluepaper.com/article/guest-commentary-grinder-pump-attacks-get-personal/

*Antisocial Behavior*

[Wulczyn et al. 2017]
[Zhang et al. 2018]

https://it.depositphotos.com/12630133/stock-photo-3d-talking-concept-over-white.html

*Disputes*

[Wang and Cardie 2014]

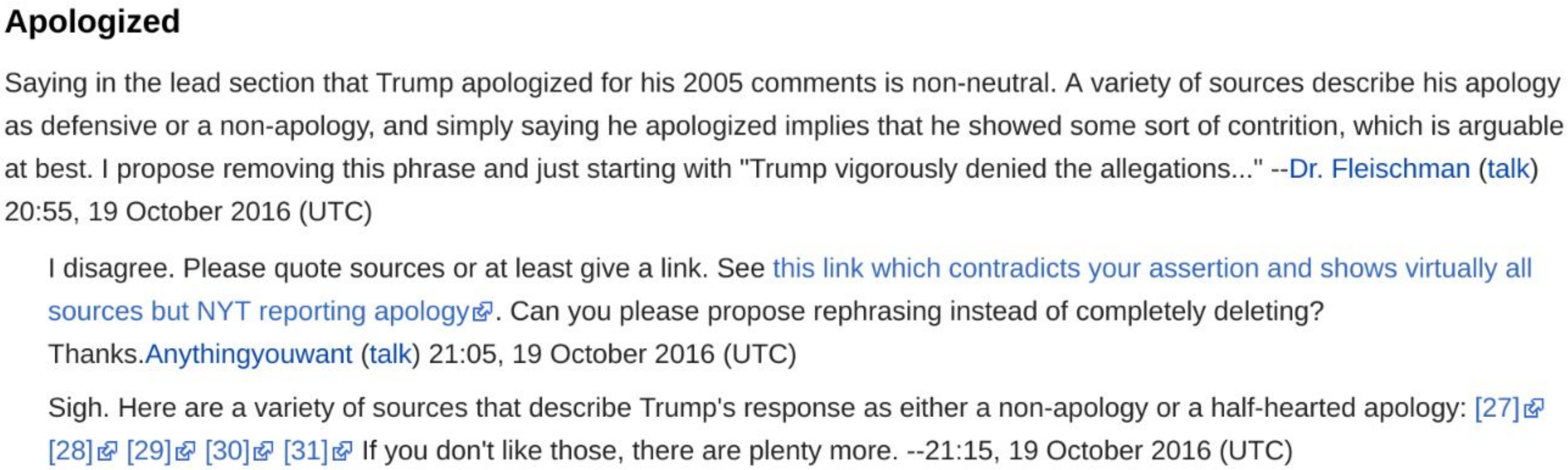
http://www.boerenopeenkruispunt.be/Tips/tabid/105/ArticleID/245/Default.aspx

*Conversational Behaviors*

[Danescu-Niculescu-Mizil 2012]

***Rich** context of conversational interactions.*

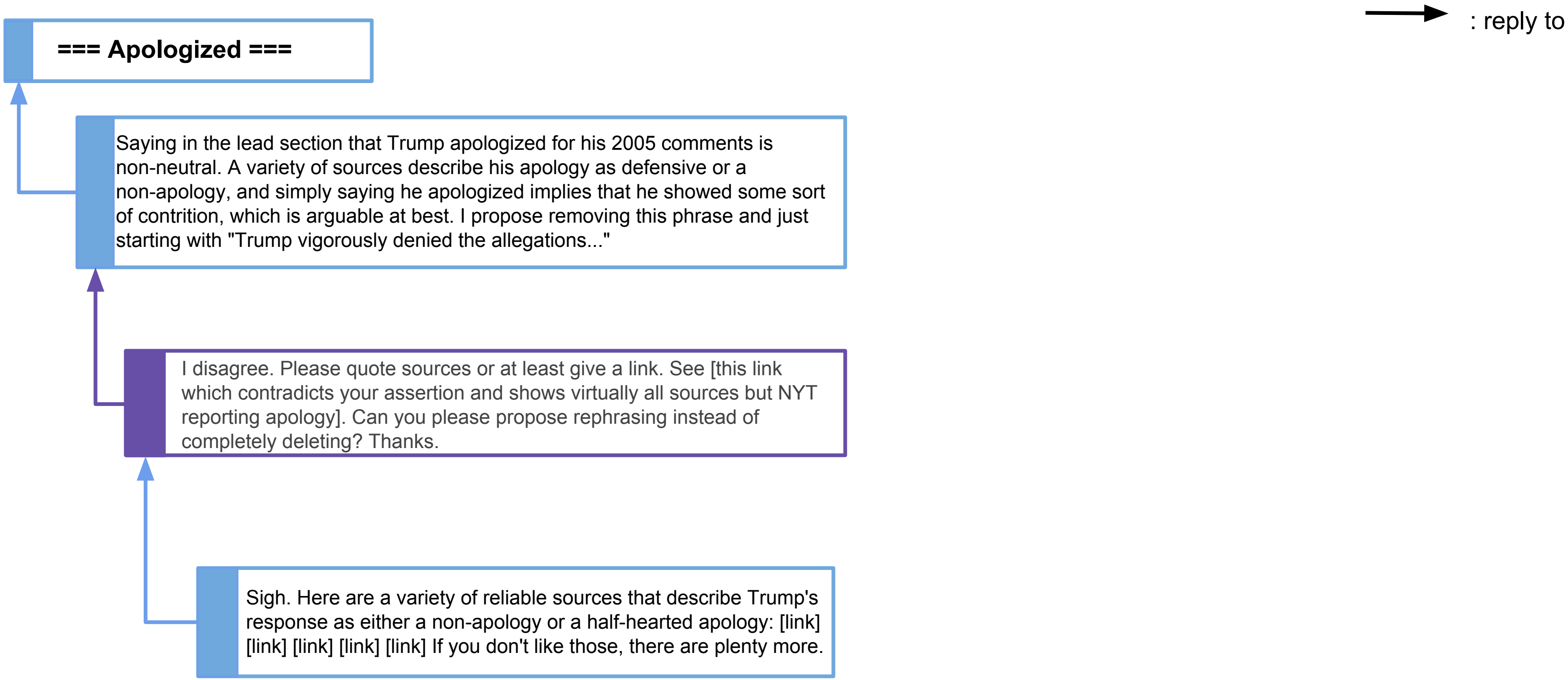
# Example: Extracting structured data from Wikipedia Talk Pages



**Apologized**

Saying in the lead section that Trump apologized for his 2005 comments is non-neutral. A variety of sources describe his apology as defensive or a non-apology, and simply saying he apologized implies that he showed some sort of contrition, which is arguable at best. I propose removing this phrase and just starting with "Trump vigorously denied the allegations..." --Dr. Fleischman (talk) 20:55, 19 October 2016 (UTC)

> I disagree. Please quote sources or at least give a link. See this link which contradicts your assertion and shows virtually all sources but NYT reporting apology. Can you please propose rephrasing instead of completely deleting? Thanks.Anythingyouwant (talk) 21:05, 19 October 2016 (UTC)
>
> Sigh. Here are a variety of sources that describe Trump's response as either a non-apology or a half-hearted apology: [27] [28] [29] [30] [31] If you don't like those, there are plenty more. --21:15, 19 October 2016 (UTC)



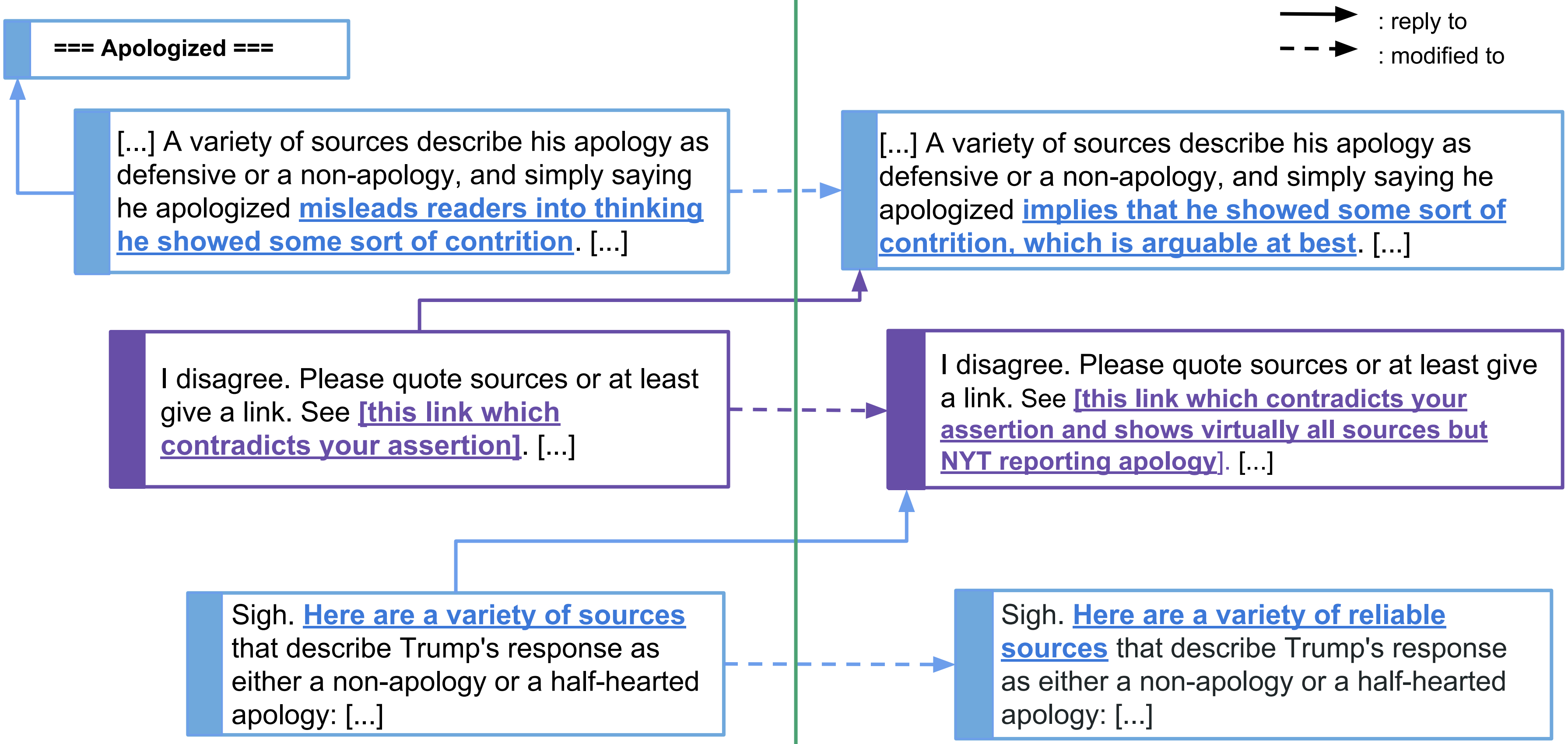
: reply to

=== Apologized ===

Saying in the lead section that Trump apologized for his 2005 comments is non-neutral. A variety of sources describe his apology as defensive or a non-apology, and simply saying he apologized implies that he showed some sort of contrition, which is arguable at best. I propose removing this phrase and just starting with "Trump vigorously denied the allegations..."

I disagree. Please quote sources or at least give a link. See [this link which contradicts your assertion and shows virtually all sources but NYT reporting apology]. Can you please propose rephrasing instead of completely deleting? Thanks.

Sigh. Here are a variety of reliable sources that describe Trump's response as either a non-apology or a half-hearted apology: [link] [link] [link] [link] [link] If you don't like those, there are plenty more.



→ : reply to

- - → : modified to

=== Apologized ===

[...] A variety of sources describe his apology as defensive or a non-apology, and simply saying he apologized **misleads readers into thinking he showed some sort of contrition**. [...]

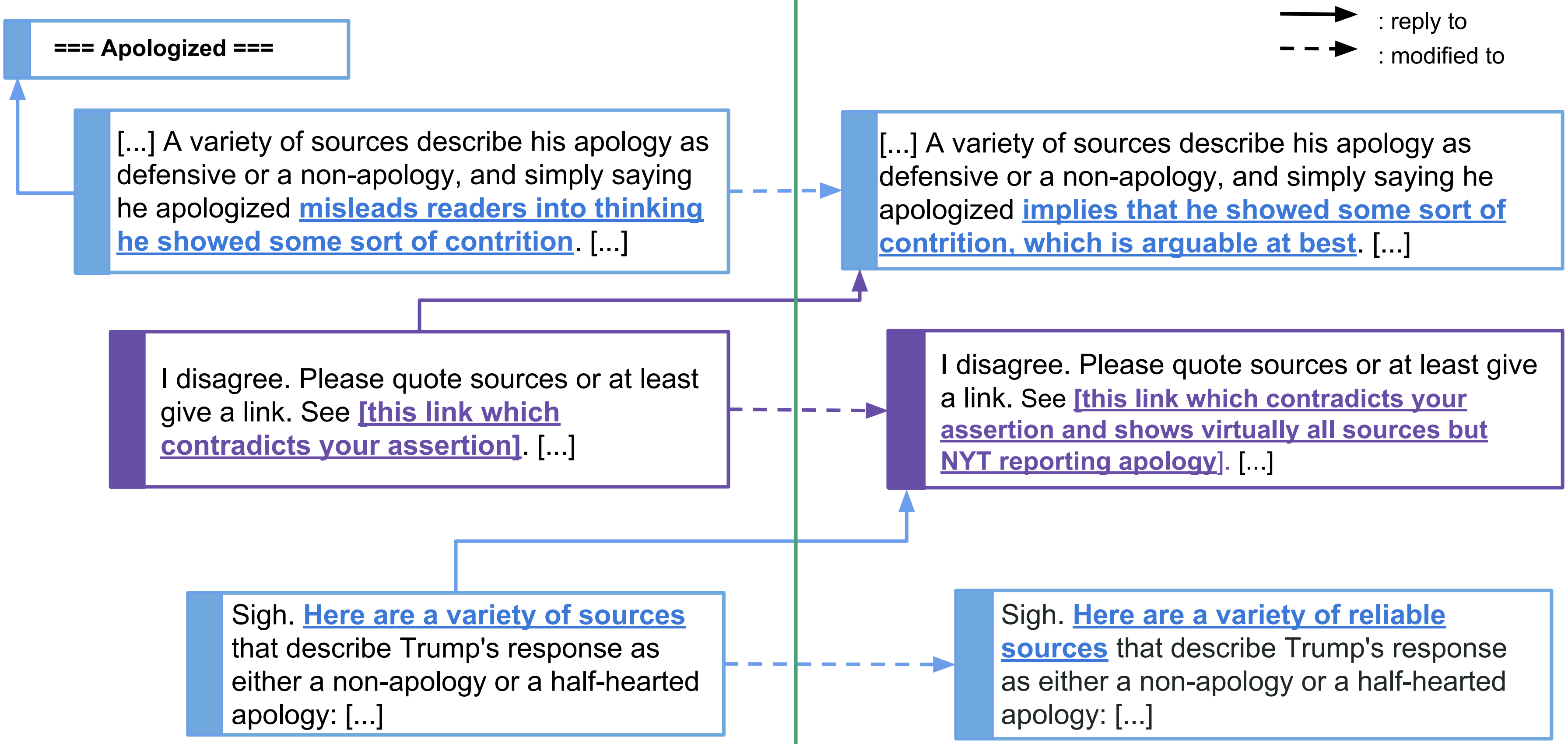
I disagree. Please quote sources or at least give a link. See **[this link which contradicts your assertion]**. [...]

Sigh. **Here are a variety of sources** that describe Trump's response as either a non-apology or a half-hearted apology: [...]

[...] A variety of sources describe his apology as defensive or a non-apology, and simply saying he apologized **implies that he showed some sort of contrition, which is arguable at best**. [...]

I disagree. Please quote sources or at least give a link. See **[this link which contradicts your assertion and shows virtually all sources but NYT reporting apology]**. [...]

Sigh. **Here are a variety of reliable sources** that describe Trump's response as either a non-apology or a half-hearted apology: [...]



=== Apologized ===

[...] A variety of sources describe his apology as defensive or a non-apology, and simply saying he apologized misleads readers into thinking he showed some sort of contrition. [...]

I disagree. Please quote sources or at least give a link. See [this link which contradicts your assertion]. [...]

Sigh. Here are a variety of sources that describe Trump's response as either a non-apology or a half-hearted apology: [...]

[...] A variety of sources describe his apology as defensive or a non-apology, and simply saying he apologized implies that he showed some sort of contrition, which is arguable at best. [...]

I disagree. Please quote sources or at least give a link. See [this link which contradicts your assertion and shows virtually all sources but NYT reporting apology]. [...]

Sigh. Here are a variety of reliable sources that describe Trump's response as either a non-apology or a half-hearted apology: [...]

***Rich** detail of conversational interactions.*

# Roadmap

- Reconstruction Pipeline
- Dataset Statistics
- Result Evaluation
- Case Studies
- Next Step

# The Data Structure

**Revision R1**
**Author: U1**

**== Improving the Article ==**

Let's discuss how to write this article!

**Revision R2**
**Author: U2**

**== Improving the Article ==**

You are an IDIOT!

**Revision R3**
**Author: U1**

**== Improving the Article ==**

Let's discuss how to write this article!

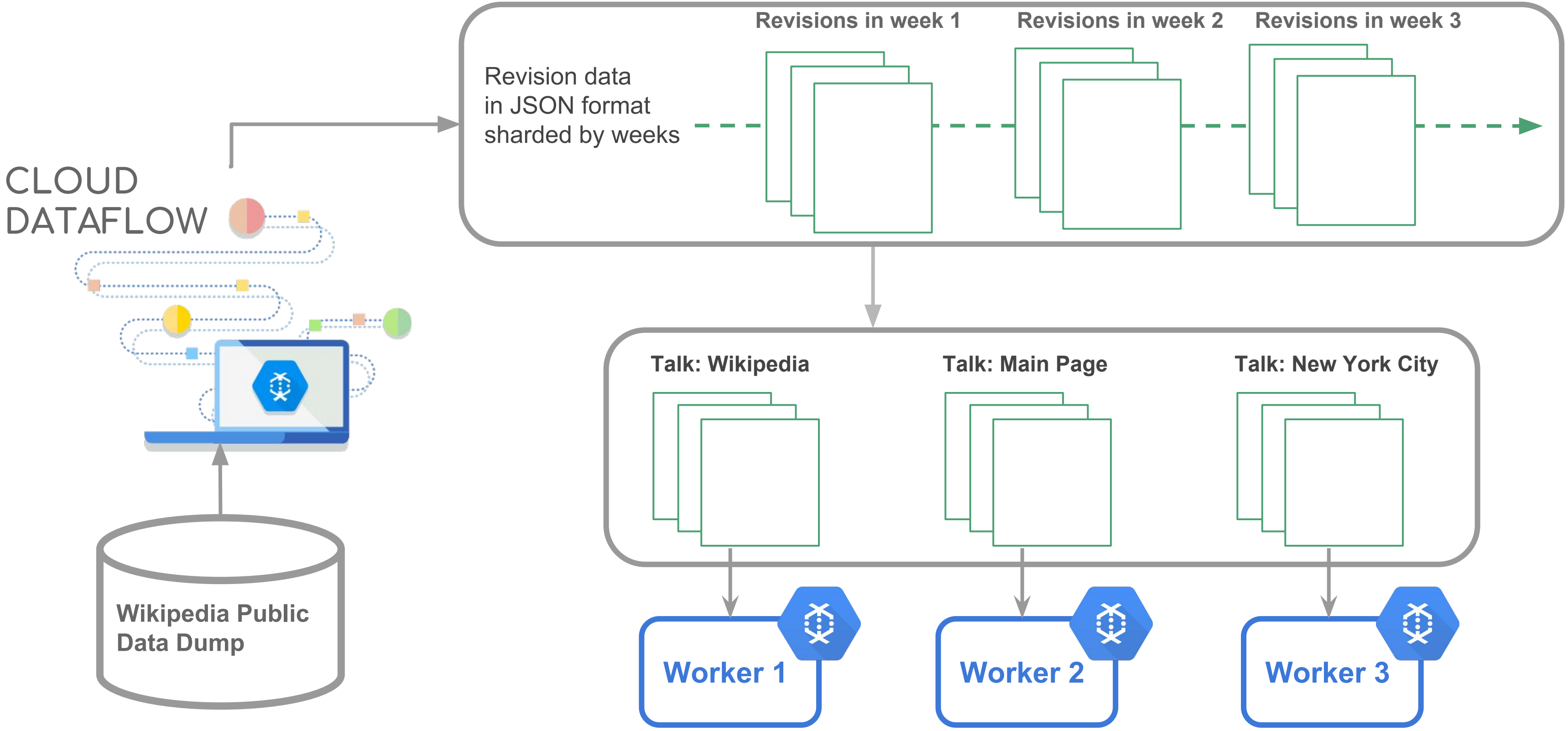
**Revision R4**
**Author: U3**

**== Improving the Article ==**

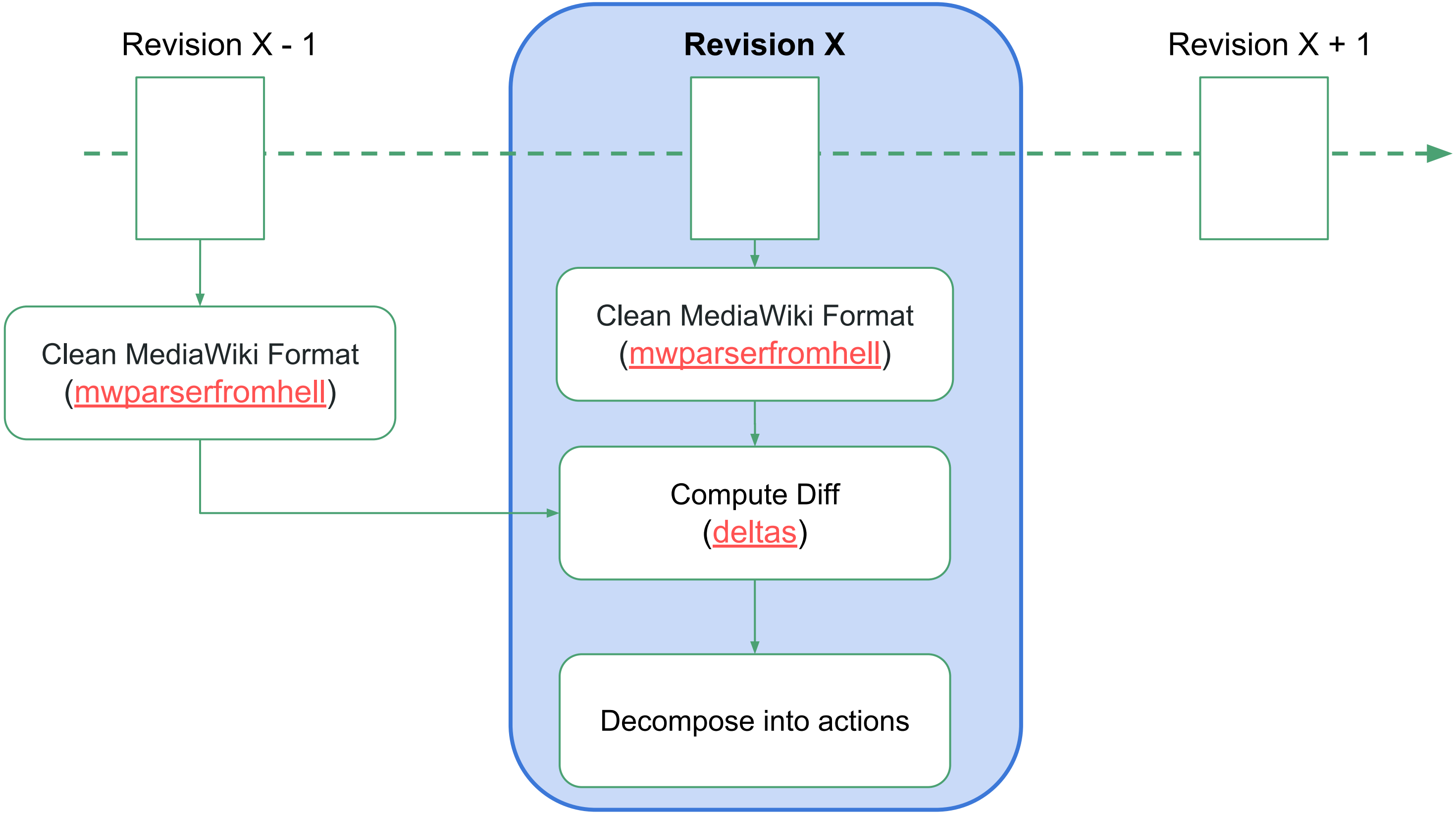
Let's discuss how to improve this article!

| ID | Type | Content | ReplyTo | Action Parent | Author | Revision |
|---|---|---|---|---|---|---|
| 1 | Creation | == Improving the Article == | Null | Null | U1 | R1 |
| 2 | Addition | Let's discuss how to write [...] | 1 | Null | U1 | R1 |
| 3 | Deletion | Let's discuss how to write [...] | Null | 2 | U2 | R2 |
| 4 | Addition | You are an IDIOT. | 1 | Null | U2 | R2 |
| 5 | Restoration | Let's discuss how to write [...] | Null | 2 | U1 | R3 |
| 6 | Deletion | You are an IDIOT. | Null | 4 | U1 | R3 |
| 7 | Modification | Let's discuss how to improve [..] | 1 | 2 | U3 | R4 |

# Reconstruction Pipeline

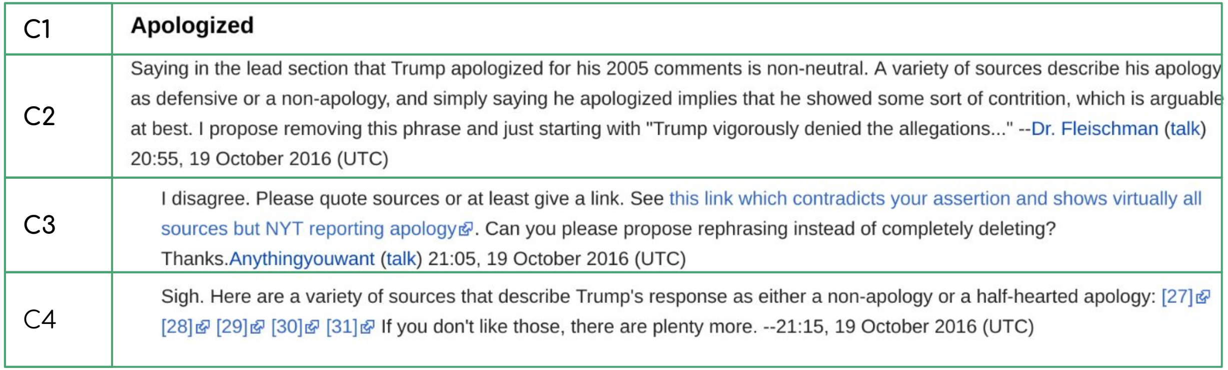


# Reconstruction Pipeline -- Worker Step


Revision X - 1
Revision X
Revision X + 1
Clean MediaWiki Format (mwparserfromhell)
Clean MediaWiki Format (mwparserfromhell)
Compute Diff (deltas)
Decompose into actions

# Reconstruction Pipeline -- Data Structures

**Page State**
*Records offsets of comments of those that are present on the page*
*To distinguish comment additions from modifications*

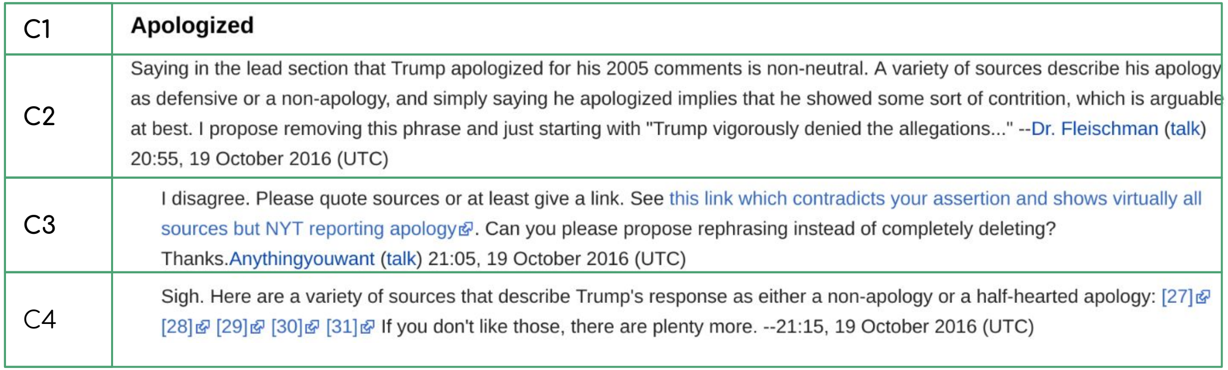


| | |
|---|---|
| C1 | **Apologized** |
| C2 | Saying in the lead section that Trump apologized for his 2005 comments is non-neutral. A variety of sources describe his apology as defensive or a non-apology, and simply saying he apologized implies that he showed some sort of contrition, which is arguable at best. I propose removing this phrase and just starting with "Trump vigorously denied the allegations..." --Dr. Fleischman (talk) 20:55, 19 October 2016 (UTC) |
| C3 | I disagree. Please quote sources or at least give a link. See this link which contradicts your assertion and shows virtually all sources but NYT reporting apology. Can you please propose rephrasing instead of completely deleting? Thanks.Anythingyouwant (talk) 21:05, 19 October 2016 (UTC) |
| C4 | Sigh. Here are a variety of sources that describe Trump's response as either a non-apology or a half-hearted apology: [27] [28] [29] [30] [31] If you don't like those, there are plenty more. --21:15, 19 October 2016 (UTC) |

# Reconstruction Pipeline -- Data Structures

**Page State**

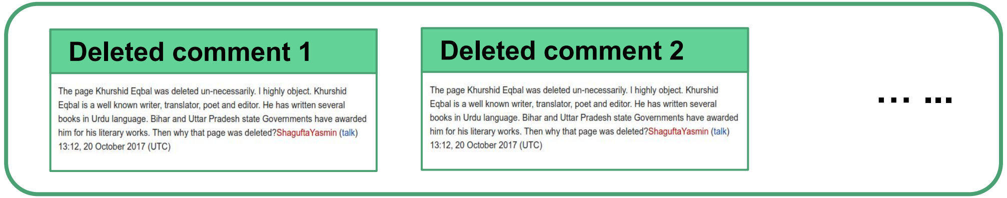
*Records offsets of comments of those that are present on the page*
*To distinguish comment additions from modifications*



| | |
|---|---|
| C1 | **Apologized** |
| C2 | Saying in the lead section that Trump apologized for his 2005 comments is non-neutral. A variety of sources describe his apology as defensive or a non-apology, and simply saying he apologized implies that he showed some sort of contrition, which is arguable at best. I propose removing this phrase and just starting with "Trump vigorously denied the allegations..." --Dr. Fleischman (talk) 20:55, 19 October 2016 (UTC) |
| C3 | I disagree. Please quote sources or at least give a link. See this link which contradicts your assertion and shows virtually all sources but NYT reporting apology. Can you please propose rephrasing instead of completely deleting? Thanks.Anythingyouwant (talk) 21:05, 19 October 2016 (UTC) |
| C4 | Sigh. Here are a variety of sources that describe Trump's response as either a non-apology or a half-hearted apology: [27] [28] [29] [30] [31] If you don't like those, there are plenty more. --21:15, 19 October 2016 (UTC) |

**Deleted Comments**

*To identify restorations*
*Each comment expires in 2 weeks*



**Deleted comment 1**

The page Khurshid Eqbal was deleted un-necessarily. I highly object. Khurshid Eqbal is a well known writer, translator, poet and editor. He has written several books in Urdu language. Bihar and Uttar Pradesh state Governments have awarded him for his literary works. Then why that page was deleted?ShaguftaYasmin (talk) 13:12, 20 October 2017 (UTC)

**Deleted comment 2**

The page Khurshid Eqbal was deleted un-necessarily. I highly object. Khurshid Eqbal is a well known writer, translator, poet and editor. He has written several books in Urdu language. Bihar and Uttar Pradesh state Governments have awarded him for his literary works. Then why that page was deleted?ShaguftaYasmin (talk) 13:12, 20 October 2017 (UTC)

# The Resulting Data Structure  (again)

**Revision R1**
**Author: U1**

**== Improving the Article ==**

Let’s discuss how to write this article!

**Revision R2**
**Author: U2**

**== Improving the Article ==**

You are an IDIOT!

**Revision R3**
**Author: U1**

**== Improving the Article ==**

Let’s discuss how to write this article!

**Revision R4**
**Author: U3**

**== Improving the Article ==**

Let’s discuss how to improve this article!

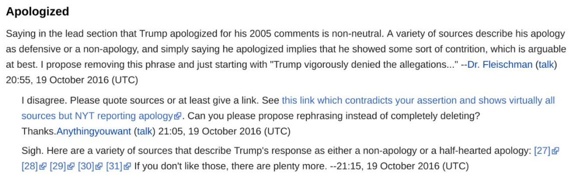
| ID | Type | Content | ReplyTo | Action Parent | Author | Revision |
|---|---|---|---|---|---|---|
| 1 | Creation | == Improving the Article == | Null | Null | U1 | R1 |
| 2 | Addition | Let’s discuss how to write [...] | 1 | Null | U1 | R1 |
| 3 | Deletion | Let’s discuss how to write [...] | Null | 2 | U2 | R2 |
| 4 | Addition | You are an IDIOT. | 1 | Null | U2 | R2 |
| 5 | Restoration | Let’s discuss how to write [...] | Null | 2 | U1 | R3 |
| 6 | Deletion | You are an IDIOT. | Null | 4 | U1 | R3 |
| 7 | Modification | Let’s discuss how to improve [..] | 1 | 2 | U3 | R4 |

# Dataset Statistics

3.6M Users

16M Talk Pages

72M Revisions



**Apologized**

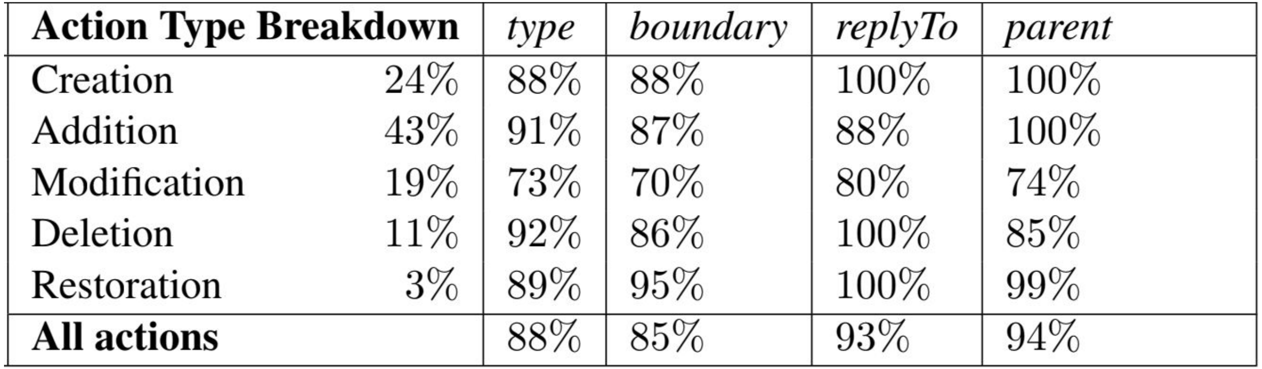
Saying in the lead section that Trump apologized for his 2005 comments is non-neutral. A variety of sources describe his apology as defensive or a non-apology, and simply saying he apologized implies that he showed some sort of contrition, which is arguable at best. I propose removing this phrase and just starting with "Trump vigorously denied the allegations..." --Dr. Fleischman (talk) 20:55, 19 October 2016 (UTC)

> I disagree. Please quote sources or at least give a link. See this link which contradicts your assertion and shows virtually all sources but NYT reporting apology. Can you please propose rephrasing instead of completely deleting? Thanks.Anythingyouwant (talk) 21:05, 19 October 2016 (UTC)
>
> Sigh. Here are a variety of sources that describe Trump's response as either a non-apology or a half-hearted apology: [27] [28] [29] [30] [31] If you don't like those, there are plenty more. --21:15, 19 October 2016 (UTC)

48M Conversations

155M Actions

# Evaluation Result



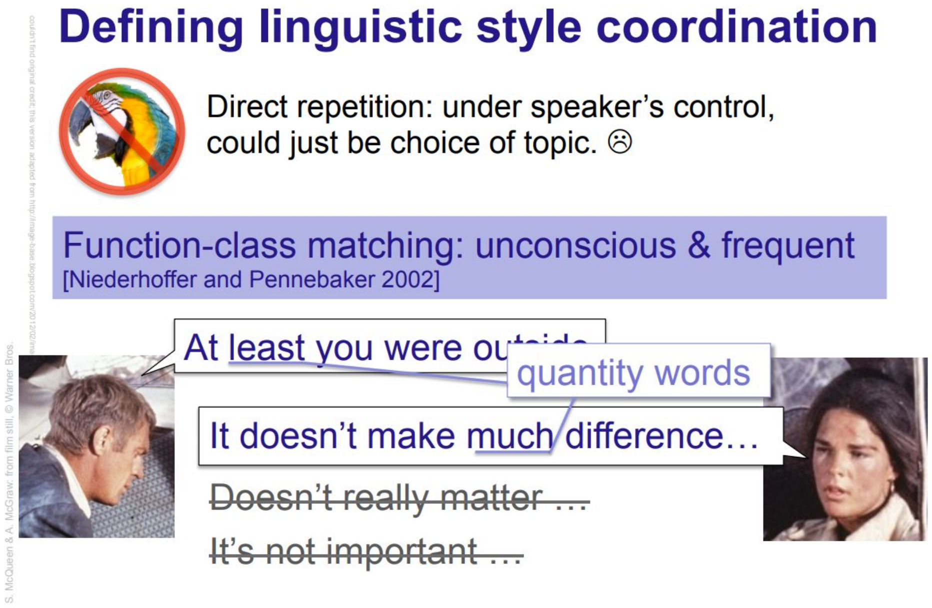
| **Action Type Breakdown** | | *type* | *boundary* | *replyTo* | *parent* |
|---|---|---|---|---|---|
| Creation | 24% | 88% | 88% | 100% | 100% |
| Addition | 43% | 91% | 87% | 88% | 100% |
| Modification | 19% | 73% | 70% | 80% | 74% |
| Deletion | 11% | 92% | 86% | 100% | 85% |
| Restoration | 3% | 89% | 95% | 100% | 99% |
| **All actions** | | 88% | 85% | 93% | 94% |

**Error type breakdown:**

52%  HTML parsing / ambiguous diff
18%  Complicated user behavior
30%  Other

# Case Study -- User Coordination

*Coordination:*
*In a conversation between A and B, to what extent B adopts A's language patterns.*



## Defining linguistic style coordination

Direct repetition: under speaker's control, could just be choice of topic. ☹

Function-class matching: unconscious & frequent
[Niederhoffer and Pennebaker 2002]

At least you were ou

quantity words

It doesn't make much difference…

~~Doesn't really matter …~~

~~It's not important …~~

couldn't find original credit; this version adapted from http://image-base.blogspot.com/2012/02/ima

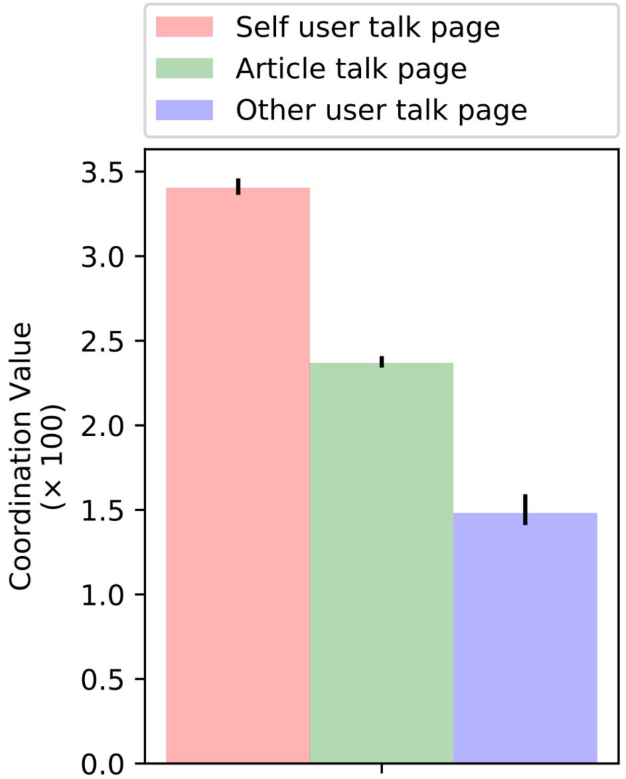
S. McQueen & A. McGraw: from film still, © Warner Bros.

[Danescu-Niculescu-Mizil 2012]
Slide from Lillian Lee at the "New Directions in Analyzing Text as Data" conference

# Case Study -- User Coordination

*Coordination:*
*In a conversation between A and B,*
*to what extent B adopts A’s language patterns.*

We computed coordination value using methodology proposed by previous work* over a subset of users on different locations.



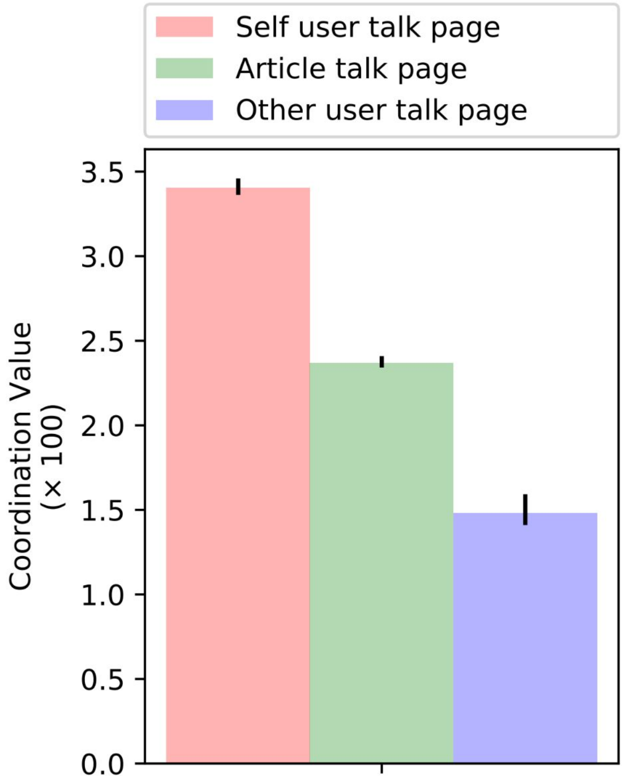
*Danescu-Niculescu-Mizil, Cristian, et al. "Echoes of power: Language effects and power differences in social interaction." Proceedings of the 21st international conference on World Wide Web. ACM, 2012.

# Case Study -- User Coordination

*Coordination:*
*In a conversation between A and B,*
*to what extent B adopts A's language patterns.*

We computed coordination value using methodology proposed by previous work* over a subset of users on different locations.

***Rich** context of conversational interactions.*

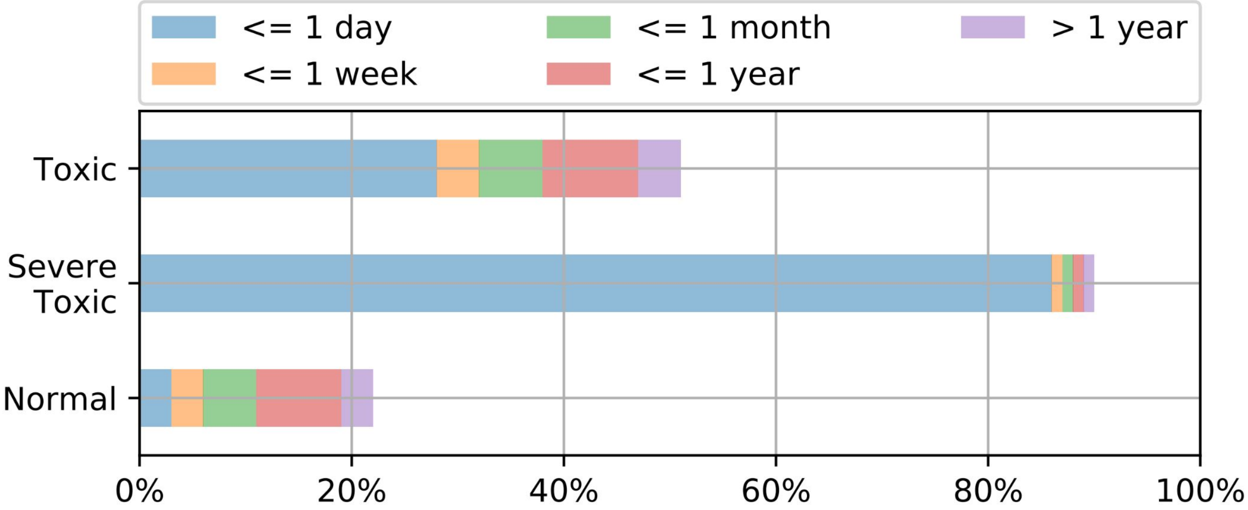


*Danescu-Niculescu-Mizil, Cristian, et al. "Echoes of power: Language effects and power differences in social interaction." Proceedings of the 21st international conference on World Wide Web. ACM, 2012.

# Case Study -- Moderation of Toxic Behavior

Addition and creation contents are labeled as Perspective API in terms of toxic/non-toxic, severe toxic/non-severe toxic.

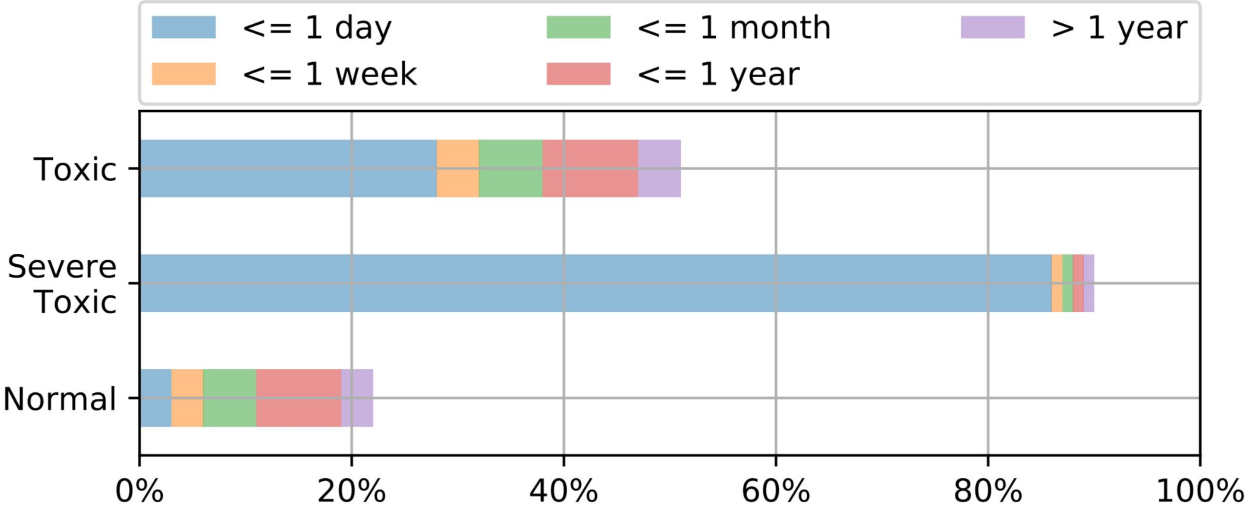
We measure the speed of deletion of these contents, compared to that of normal contents.



# Case Study -- Moderation of Toxic Behavior

Addition and creation contents are labeled as Perspective API in terms of toxic/non-toxic, severe toxic/non-severe toxic.

We measure the speed of deletion of these contents, compared to that of normal contents.



***Rich** detail of conversational interactions.*

# Next Steps

- Improve Pipeline Quality and Efficiency
- Live System of the Wikipedia Talk Pages conversations
- Release of WikiConv (with scores) & Github code

# Questions?