Annotating Social Acts: Authority Claims and Alignment Moves in Wikipedia Talk Pages

Emily M. Bender*, **Jonathan T. Morgan**†, **Meghan Oxley***, **Mark Zachry**†,
Brian Hutchinson‡, **Alex Marin**‡, **Bin Zhang**‡, **Mari Ostendorf**‡
*Department of Linguistics, †Department of Human Centered Design and Engineering
‡Department of Electrical Engineering
University of Washington
{ebender,jmo25,what,zachry}@uw.edu, {brianhutchinson,iskander,binz,mo}@ee.washington.edu

Abstract

We present the AAWD corpus, a collection of 365 discussions drawn from Wikipedia talk pages and annotated with labels capturing two kinds of social acts: alignment moves and authority claims. We describe these social acts and our annotation process, and analyze the resulting data set for interactions between participant status and social acts and between the social acts themselves.

1 Introduction

This paper presents a new annotated resource: the Authority and Alignment in Wikipedia Discussions (AAWD) corpus (available from `http://ssli.ee.washington.edu/projects/SCIL.html`). The AAWD corpus contains discussions from English-language Wikipedia talk pages extracted from the 2008 Wikipedia data dump and annotated for two types of social acts: authority claims and positive/negative alignment moves. In brief, an authority claim is a statement made by a discussion participant aimed at bolstering their credibility in the discussion. An alignment move is a statement by a participant which explicitly positions them as agreeing or disagreeing with another participant or participants regarding a particular topic.

These annotations are intended to make accessible for automated processing two interesting and characteristic aspects of interaction in online discussion forums. As a dataset for computational and sociolinguistic analysis, the discussion pages within Wikipedia are valuable for several reasons. First, the interaction among the participants is nearly entirely captured within the dataset, and all of the "identity-work" (Bucholtz and Hall, 2010) done by Wikipedia discussion participants needs to be done directly in the text of their comments. Furthermore, the discussions tend to be task-driven, focused on the shared goal of improving the associated article. This leads the data to be a particularly rich source of linguistic expressions of authority and alignment.

Our annotations represent a kind of information which is rather different from that involved in NLP tasks such as POS tagging, morphological analysis, parsing and semantic role labeling. Such tasks involve recognizing information that is implicit in the linguistic signal but nonetheless part of its structure. Tasks such as named-entity recognition and word sense disambiguation are also close to the linguistic structure of the signal. Authority claims and alignment moves, on the other hand, are examples of communicative moves aimed at social positioning of a discussant within a group of participants, which may be specialized dialog acts but are referred to here as "social acts." We distinguish social acts from "social events" as described in (Agarwal and Rambow, 2010): social events correspond to types of interactions among people, whereas a social act is associated with a fine-grained social goal and reflected in the specific choices of words and orthographic or prosodic cues at the level of a turn.

The primary value of this new data set is in facilitating computational modeling of a new task type, i.e. the identification of fine-grained social acts in linguistic interaction. While there has been some prior work on detecting agreements and disagree-

Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 48–57,
Portland, Oregon, 23 June 2011.

ments in multiparty discussions (Hillard et al., 2003; Galley et al., 2004), which is related to detecting positive/negative alignment moves, most previous work on authority bids has involved descriptive studies, e.g. (Galegher et al., 1998). Computational modeling of these phenomena and automatic detection will help with understanding effective argumentation strategies in online discussions and automatic identification of divisive or controversial discussions and online trolls. We believe that these tasks also provide an interesting arena in which to study linguistic feature engineering and feature selection. As with tasks such as sentiment analysis, a simple "bag-of-words" model with word or even n-gram-based features is not sufficiently powerful to detect many instances of these social acts, where combinations of positive and negative words must be interpreted in context, e.g. *absolutely* is positive alone but amplifies a negative in *absolutely not*, and *yeah* in *yeah, I want to correct something John said of course* doesn't necessarily indicate agreement. The typical scenario where hand-annotated training data is limited presents a challenge for learning phrase patterns that discriminate social acts.

In the remainder of this paper, we further describe the social acts and annotation schemata (Section 2), provide details of the AAWD corpus (Section 3), and analyze the distribution of the social acts (Section 4). This analysis describes the distribution of the social acts and tests hypotheses about their interactions with each other and with user status.

2 Annotation Schemata

2.1 Authority Claims

The ability to persuade others to believe in one's statements or the soundness of one's judgments is a necessary component of human social interaction. In order to establish the necessary credibility to secure the belief or assent of others, communicators will often couch their statements in some broadly-recognized basis for authority. These "arguments from authority" have been recognized as an important component of informal logic by many language philosophers (Liu, 1997), including John Locke (1959 [1690]). In recent decades the self-presentation of authority has been studied in a variety of spoken and written contexts by scholars from disciplines such as communication, rhetoric, health studies, sociolinguistics, linguistic pragmatics and political science in order to understand the strategies that communicators operating in different genres and media employ to establish themselves as credible discursive participants. Studies of online product reviews (Mackiewicz, 2010), online political deliberation (Jensen, 2003), scientific publications (Thompson, 1993), online forum posts (Galegher et al., 1998; Richardson, 2003) and radio talk-shows (Thornborrow, 2001) have revealed that considerations of genre, medium and social context all shape the ways interactants attempt to claim the authority to be listened to and taken seriously.

From the perspective of discourse analysis, authority claims provide an interesting lens through which to view a text, as the overall frequency of claims can reflect the nature or purpose of the discourse (e.g. task-oriented collaboration vs. undirected conversation) and the distribution of claim types can reveal features of the social context in which they are made, such as shared norms, practices and community values. For example, since certain bases for authority may be seen as more credible than others in certain contexts (such as citation of peer-reviewed publications in academic scholarship, or references to personal experience in online support groups), the prevalence and distribution of different types of claims in a written text or a conversation transcript can illuminate the shared values of speakers and audiences in a given genre (Galegher et al., 1998). Although the linguistic construction of authority claims can vary greatly according to the genre of the communication, within a single genre there is often great regularity in the ways claims are made, such as the common *I'm a long-time listener* introduction used by radio talk-show call-in guests. Even across genres, recognizable types emerge: references to personal credentials (such as education or profession) are found to be important in newsgroup messages (Richardson, 2003), product reviews (Mackiewicz, 2010) and online scientific article comments (Shanahan, 2010).

Our taxonomy of authority claims was iteratively developed based on our empirical analysis of conversational interaction in two different genres: political talk shows and Wikipedia discussion pages (Oxley et al., 2010), with reference to

the literature cited above. Our codebook (available from `http://ssli.ee.washington.edu/projects/SCIL.html`) includes detailed definitions as well as positive and negative examples for each claim type.

We classify authority claims into the following types (examples are drawn from our data):

Credentials: Credentials claims involve reference to education, training, or a history of work in an area. (Ex: *Speaking as a native born Midwesterner who is also a professional writer...*)

Experiential: Experiential claims are based on an individual's involvement in or witnessing of an event. (Ex: *If I recall correctly, God is mentioned in civil ceremonies in Snohomish County, Washington, the only place I've witnessed one.*)

Institutional: Institutional claims are based on an individual's position within an organization structure that governs the current discussion forum or has power to affect the topic or direction of the discussion. (Not attested in our corpus.)

Forum: Forum claims are based on policy, norms, or contextual rules of behavior in the interaction. (Ex: *Do any of these meet wikipedia's [[WP:RS | Reliable Sources]] criteria?*)

External: External claims are based on an outside authority or source of expertise, such as a book, magazine article, website, written law, press release, or court decision. (Ex: *The treaty of international law which states that wars have to begin with a declaration is the Hague Convention relative to the Opening of Hostilities from 1907.*)

Social Expectations: Social Expectations claims are based on the intentions or expectations (what they think, feel or believe) of groups or communities that exist beyond the current conversational context. (Ex: *I think in the minds of most people, including the government, the word "war" and a formal declaration of war have come apart.*)

2.2 Alignment Moves

In multiparty discourse, relationships among participants manifest themselves in social moves that participants make to demonstrate alignment with or against other participants. Expressing alignment with another participant functions as a means of enhancing solidarity with that participant while expressing alignment against another participant maintains social distance between conversational participants, particularly in situations where participants may be previously unacquainted with each other (Svennevig, 1999). Changes in the alignment of participants toward one another or "shifts in footing" may reflect changes in interpersonal relationships or may be more transitory, demonstrating minor concessions and critiques embedded within larger, more stable patterns of participant agreement and disagreement (Goffman, 1981; Wine, 2008).

As Wikipedia editors negotiate about article content, they make statements that support or oppose propositions suggested by other editors and thereby publicly align either with or against other editors in the discussion. Although ways of expressing agreement and disagreement vary according to power relations between participants, participant goals, and conversational context (Rees-Miller, 2000), previous research has suggested that expressions of agreement and disagreement in written language are more explicit than oral expressions of agreement and disagreement (Mulkay, 1985; Mulkay, 1986) and that statements of agreement are particularly explicit in online discussions (Baym, 1996).

We classify alignment moves into positive and negative types, according to whether the participant is agreeing or disagreeing with the target:

Positive alignment moves express agreement with the opinions of another participant. Positive alignment is annotated in cases of explicit agreement, praise/thanking, positive reference to another participant's point (e.g. *As Joe pointed out...*), or where other clear indicators of positive alignment are present.

Negative alignment moves express disagreement with the opinions of another participant. Negative alignment is annotated in cases of explicit disagreement, doubting, sarcastic praise, criticism/insult, dismissing, or where other clear indicators of negative alignment (such as typographical cues) are present.

Based on our experience using the types of authority claims to diagnose and correct sources of inter-annotator disagreement (see §3.3 below), we developed subtypes of positive and negative alignment. While these do not have the same theoretical grounding as the types of authority claims, they did serve the same purpose of improving our annotation

over time.

We annotate a target for each alignment move, which may be one or more specific other parties in the conversation, the group as the whole, or someone outside the conversation. In addition, we include a category labeled "unclear" for cases where there is an alignment move, but the annotators are not able to discern its target. Again, the codebook includes example subtypes as part of detailed definitions as well as positive and negative examples for each alignment type.

3 The Corpus

3.1 Source Data

Wikipedia talk pages (also called discussion pages) are editable pages on which editors can take part in threaded, asynchronous discussions about the content of other pages. All editors potentially interested in a given article can join the conversation on that article's talk page. Sometimes these conversations take the form of a deliberative exchange or even a heated argument as editors advocate different ideas about such things as the content or form of an article. Each edit to the talk pages is recorded as a unique revision in the system and thus becomes part of the permanent record of system activity.

Wikipedia constitutes a particularly valuable natural laboratory for studies such as this one, for several reasons. First, the interaction among the participants is almost entirely captured within the Wikipedia database: while some Wikipedians might interact with each other in person or in other online fora (such as IRC or mailing lists), this is the exception rather than the rule. Furthermore, while participants often maintain persistent identities (usernames for registered users; IP addresses for unregistered ones) there are no cues to social identities available to the participants beyond what is captured in the digital record. Therefore all of the effort that participants put into constructing their online identities is in the record for analysis. Second, the discussions on Wikipedia talk pages tend to be goal-oriented, as the discussion topic is the Wikipedia article that the participants are collaboratively editing. This goal-orientation motivates participants to explicitly align with each other in the course of discussions and buttress their arguments with authority claims. Finally, the Wikipedia dataset contains rich metadata, such as the date and time of each edit (identified by revision id) to every article or talk page; the editor responsible for the edit (identified by username or IP address, depending on registration status); and markup such as hyperlinks and formatting used in the textual content of each edit. These metadata allow for sophisticated data analysis at the editor level (e.g. how many edits made by one editor in a given span of time) and the page level (e.g. how many editors have participated in a talk page discussion).

The Wikimedia Foundation frequently releases the database dump of the Wikipedia pages in the form of XML (available at `http://download.wikimedia.org`). The database dumps are categorized into languages, and for each language, there are XML files corresponding to different levels of detail in terms of the information they contain. To get the information on all revisions, we used the largest database dump, which contains all Wikipedia pages and complete edit history. The XML file was parsed and a database created locally with all the revision information for both main pages and talk pages. We then constructed queries to retrieve the main pages and corresponding talk pages based on a list of topics for which extensive discussions are likely to occur.

Our data is drawn from a set of 365 discussions from 47 talk pages. The discussions were selected to contain at least 5 turns and at least 4 human participants.[1] The earliest edit in our data set is from January 29, 2002 and the latest is from January 6, 2008. A total of 1,509 editors collectively make 6,066 turns in this data. Of the 365 discussions, 185 were annotated for both alignment moves and authority claims. An additional 26 were annotated for alignment only and an additional 154 were annotated for authority only. The numbers of editors and turns in these sets are shown in Table 1.

3.2 Annotation Units

A Wikipedia talk page is in itself a wiki-style document. Thus, each modification to a talk page by an editor can modify multiple sections of the page. We define a "turn" as a contiguous body of text on the

[1] Wikipedia discussions may also include contributions by automated "bots".

	Annotated for		
	authority	alignment	both
pages	47	36	36
discussions	339	211	185
editors	1,417	988	896
turns	5,636	3,390	2,960

Table 1: Pages, discussions, editors and turns in annotated data

corresponding page that was modified as part of a single revision. Thus, a single revision may result in multiple turns being added. Each turn may include one or more paragraphs of text, either existing but modified, or new additions. We annotated authority claims at the paragraph level and alignment moves at the turn level. The larger unit is used for alignment moves because the phenomenon as defined can span a larger section of text.

The annotation tool (a modified version of LDC's XTrans (Glenn et al., 2009)) allowed annotators to indicate the presence and type of claims or moves in each annotation unit, in addition to selecting spans of text corresponding to each social act. For alignment moves, within a turn, alignment of the same type (positive or negative) with the same target was annotated as a single alignment move, even across multiple sentences. Where the type or target differed, we annotated up to three separate alignment moves per annotation unit. For authority claims, we also annotated up to three claims per annotation unit, with each claim identified by a single span of text. Claims in separate sentences of an annotation unit counted as separate even if they were of the same type. Figure 1 gives an example from our codebook of a turn with multiple alignment moves.

3.3 Annotation Process

Each discussion thread was annotated independently by two or more annotators. Inter-annotator agreement was calculated at weekly intervals to assess annotation progress and identify areas of disagreement. Adjudicators also performed "spot checks" of annotated data weekly and provided feedback when there were disagreements among annotators or when codes seemed to be inconsistently or erroneously applied. The codebooks for authority claims and alignment moves were also iteratively refined with the addition of positive and negative examples and specific linguistic cues commonly associated with particular move or claim types based on spot-check results and annotator feedback.

Two strategies that proved useful in maintaining consistency in the frequency and reliability of coding across annotators were the computation of average agreement and comparison of overall counts of each codable unit on a weekly basis. Computing average agreement allowed adjudicators to identify particular categories that were proving especially difficult to code consistently, and to better focus their efforts on re-training annotators and updating the relevant sections of the annotation guidelines. Comparing counts of the number of times two annotators had coded a particular category over the same number of discussions also proved useful for identifying potential problems with under- or over-coding of a category by a particular annotator.

3.4 Reconciliation

The manual annotation process was completed independently by each annotator, resulting in multiple sets of labels. To create a single copy of the data that can be used in learning experiments, an algorithm was designed to merge the annotations into a single, "master" version. The algorithm balances annotation consistency and simplicity of the merging process. We treat the annotations for each unit in a file as a set with respect to type: Multiple labels of the same type are treated as a single label for purposes of reconciliation, with only one label of each type allowed for each annotation unit.

We mark each social act which had been identified by at least two annotators as having "high confidence." If a social act was identified by only one annotator in that annotation unit, it is marked as having "low confidence." This procedure yields two sets of social act types found in each annotation unit, one consisting of the high confidence labels, and another of the low confidence labels. The labels from each set are kept distinct, i.e. for each label in the high confidence set, the corresponding label in the low confidence set has the suffix "_single" appended to the high confidence label.

Aggregated social act labels are propagated to the sentence level by using a dynamic programming algorithm to match sentences (determined by automatic segmentation) with the keyword spans

speaker	turn	transcript	alignment1	alignment2	alignment3
S1	3	<k1>S2, I think you're right</k1>. <k2>S3's idea is way off base </k2>, but <k1> you seem to have a good solution</k1>. <k3>But I disagree with your name for the section</k3> — Iraq War is used in the United States media and should be used here as well.	positive:S2: :explicit_ agreement	negative:S3: :explicit_ disagreement	negative:S2: :explicit_ disagreement

Figure 1: Example from alignment codebook

based on overlap. A sentence could have multiple positive labels if one or more annotators labeled it for different types in the high or low confidence set. Sentences in turns with a marked social act but not aligned to text spans are labeled as "unused" due to the ambiguity associated with a limit on the number of social acts annotated per unit. All sentences in an annotation unit for which no annotator found any positive labels are labeled with the negative class. The data distributed at `http://ssli.ee.washington.edu/projects/SCIL.html` include both the underlying per-annotator files as well as the files output by the reconciliation process.

3.5 Annotation Quality

In complicated annotation tasks, such as those conducted in this work, establishing reliable ground truth is a fundamental challenge. The most popular approach to measuring annotation quality is via the surrogate of annotation consistency. This assumes that when annotators working independently arrive at the same decisions they have correctly carried out the task specified by the annotation guidelines. Several quantitative measures of annotator consistency have been proposed and debated over the years (Artstein and Poesio, 2008). We use the well-known Cohen's kappa coefficient κ, which accounts for uneven class priors, so one may obtain a low agreement score even when a high percentage of tokens have the same label. We also report the percentage of instances on which the annotators agreed, A, which includes agreement on the absence of a particular label. When a set of instances have been labeled by more than two annotators, we compute the average of pairwise agreement.

Scores for authority claim and alignment move agreement are presented in Tables 2 and 3.[2] For authority, the most common types of claims, forum and external, are also two of the most reliably identified. For alignment, the positive type has much better agreement scores than the negative type. Interestingly, it appears that the fine distinctions between the types of negative alignment move are a large factor in the low agreement scores. When all of the negative categories are merged, agreement is higher, although still less than for positive alignment moves.

Our κ values generally fall within the range that Landis and Koch (1977) deem "moderate agreement", but below the .8 cut-off tentatively suggested

Claim Type	N	κ	A
forum	451	0.52	0.92
external	715	0.63	0.91
experiential	185	0.33	0.96
social expectations	78	0.13	0.98
credentials	6	0.57	0.99
Overall	1157	0.59	0.86

Table 2: Agreement summary for authority claims. N denotes the number of turns of the given type that at least one annotator marked.

Move Type	N	κ	A
explicit agreement	379	0.62	0.94
praise/thanking	117	0.60	0.98
positive reference	86	0.20	0.98
explicit disagreement	453	0.29	0.92
doubting	198	0.23	0.96
sarcastic praise	38	0.30	0.99
criticism/insult	556	0.32	0.91
dismissing	396	0.16	0.91
All positive	509	0.66	0.94
All negative	1092	0.45	0.85
Overall	1378	0.50	0.80

Table 3: Agreement summary for alignment moves. N denotes the number of turns of the given type that at least one annotator marked.

[2] Institutional claims are exceedingly rare in our data, appearing in only three labels. This is not sufficient for proper κ computation, and so we do not include them in Table 3.

by Artstein and Poesio (2008).[3] One possible reason is that the negative class is not as discrete as it might be in other tasks: both alignment moves and authority claims can be more or less subtle or explicit. We have designed our annotation guidelines to emphasize the more explicit variants of each, but the same guidelines can sometimes lead annotators to pick up more subtle examples that other annotators might not feel meet the strict definitions in the guidelines. Thus we expect our "high-confidence" labels to correspond to the more blatant examples and the "low-confidence" labels, while sometimes being genuine noise, to pick out more subtle examples.

4 Analysis

While the main goal of this paper is to document the AAWD corpus, we also performed several statistical analyses of authority and alignment, in order to demonstrate the relevance of these social acts as markers of user identity and social dynamics within our corpus. In this section we present the overall distribution of authority claims and alignment moves, compare the prevalence of authority claims across user types, and show how a participant's claim-making behavior may affect how others subsequently align with them. In doing so, we consider only high-confidence labels from files which were annotated by at least two annotators. This subset includes 186 discussions annotated for alignment moves and 200 discussions annotated for authority claims. Of those, 149 discussions were annotated for both types of social acts.

4.1 Distribution of Social Acts

We find that 25% of the turns in our alignment data contain alignment moves and 21% of the turns in our authority data contain authority claims. In addition, 35% and 32% of the editors in each set make alignment moves and authority claims, respectively. The breakdown by alignment move and authority claim type is given in Table 4. Note that any given turn might contain both positive and negative alignment moves or multiple types of authority claims.

[3] Artstein and Poesio also note that it may not make sense to have only one threshold for the field.

	N	%
Alignment data		
total turns	2,890	100
turns w/positive alignment	330	11.4
turns w/negative alignment	467	16.2
turns w/any alignment	710	24.6
total editors	905	100
editors w/alignment moves	315	34.8
Authority data		
total turns	3,361	100
turns w/external claim	459	13.7
turns w/forum claim	260	7.7
turns w/experiential claim	77	2.3
turns w/soc. exp. claim	21	0.6
turns w/credentials claim	3	0.1
turns w/institutional claim	0	0
turns w/any claim	703	20.9
total editors	930	100
editors w/authority claims	297	31.9

Table 4: Summary of high-confidence alignment moves and authority claims

4.2 Authority Claim Types by User Status

Wikipedia distinguishes three different statuses: unregistered users (able to perform most editing activities, identified only by IP address), registered users (able to perform more editing activities, edits attributed to a consistent user name) and administrators (registered users with additional 'sysop' privileges). Participants of different statuses tend to do different kinds of work on Wikipedia, with administrators in particular being more likely to take on moderator work (Burke and Kraut, 2008), such as mediating and diffusing disputes among editors. Because conflict mediation requires a different kind of credibility than collaborative writing work, and because unregistered users are likely to be newer and therefore less likely to be incorporating references to Wikipedia-specific rules and norms into their projected identities (and, therefore, their conversation), we hypothesized that editors of different statuses would use different kinds of authority claims.

Indeed, this is borne out. While no user group was significantly more or less likely than any other to include authority claims overall in their posts (chi square test for independence, n=3164, df=2, χ^2=2.367, p=.306) users of different statuses did use significantly different proportions of each type of claim (chi square test for independence, n=973, df=8

Participant type	# users	% forum	% external	% claim-bearing turns
admin	44	47.1	45.1	19.6
reg	192	29.1	63.6	22.3
unreg	55	18.3	70.6	19.8
all	291	29.8	62.5	21.6

Table 5: Percentage of authority claims of forum and external types, and percentage of total turns which contained claims, across user statuses

χ^2=38.301, p<.001). As illustrated in Table 5, administrators are more likely than the other groups to make forum claims and less likely to make external claims, unregistered users make more external claims and fewer forum claims, and registered users exhibit a claim distribution that more closely reflects the overall distribution of claim types.

4.3 Authority Claim Prevalence by V-Index

Given the few visible markers of status on Wikipedia and the fact that editors are constantly interacting with new collaborators, Wikipedians perform authority by adopting insider language and norms of interaction. Supporting arguments with specific references is one such norm. Thus we hypothesized that as editors become more integrated into Wikipedia, they will make more authority claims. In order to test this hypothesis, we developed "v-index" as a proxy measure of degree of integration or "veteran status" within the community. Inspired by Ball's (2005) "h-index" of scholarly productivity, v-index balances frequency of interaction with length of interaction. Specifically, an editor's v-index at the time of a particular revision is the greatest v such that the editor has made at least v edits within the past v months (28-day periods).

We measured the v-index for each revision in our dataset, using all edits to Wikipedia in order to calculate v (not just edits to the discussions we have annotated). The v-index values for edits within our dataset range from 1 to 46.[4] We measured the proportion of turns with authority claims (of any type) for each v-index. The proportion of turns with authority claims is in fact positively correlated with v-index, confirming our hypothesis (one-sided Pearson's correlation coefficient, n=29 v-indices, r=0.371, p=0.024).

[4]The data becomes very sparse for v-indices above 29, with every v-index in this range represented by < 10 turns, so the v-indices of 30-46 were not included in this analysis.

Initial turn	Alignment in next 10 turns
no auth. claim	0.52
any auth. claim	0.63

Table 6: Average prevalence of alignment moves targeted at participant in 10 following turns

4.4 Interaction of Social Phenomena

Thus far, we have been addressing our social acts independently, but of course no social act occurs in a vacuum. Alignment moves and authority claims are only two types of social acts; many other types of social acts are present (and could be annotated) in this same data set. Even with only these two types (and their subtypes), however, we find interactions.

We hypothesized that authority claims would be likely to provoke alignment moves. That is, although participants may make alignment moves whenever someone else has expressed an opinion or taken action (e.g. edited the article attached to the discussion), we hypothesized that by making an authority claim, a participant becomes more likely to become a focal point in the debate. To test this, we calculated, for every turn, the number of alignment moves targeted at the author of that turn within the next 10 turns. We then divided the turns into those that contained authority claims and those that did not. Making an authority claim in a given turn made the participant significantly more likely to be the target of an alignment move within the subsequent 10 turns compared to turns that did not contain any claims (t=-2.086, df=772, p=.037; Table 6))

Furthermore, we find that different types of authority claims elicit different numbers of subsequent alignment moves. Specifically, turns that contain either external claims or forum claims (the two most prevalent claim types in our sample) interact differently with alignment. External claims elicited more alignment overall (t=3.189, df=411, p=.002) and more negative alignment moves than did forum claims (t=3.839, df=415, p<.001). However, external claims did not elicit significantly more positive alignment moves than forum claims (t=0.695, df=309, p=.488). This is illustrated in Table 7.

Initial turn	Alignment in next 10 turns		
	positive	negative	overall
external claim	0.26	0.49	0.74
forum claim	0.22	0.20	0.42

Table 7: Average prevalence of alignment moves targeted at participant in 10 following turns

5 Conclusion

We have presented the Authority and Alignment in Wikipedia Discussions (AAWD) corpus, a collection of 365 discussions drawn from Wikipedia talk pages and annotated for two broad types of social acts: authority claims and alignment moves. These annotations make explicit important discursive strategies that discussion participants use to construct their identities in this online forum. That "identity work" is being done with these social acts is confirmed by the correlations we find between proportions of turns with authority claims and external variables such as user status and v-index, on the one hand, and the interaction between authority claims and alignment moves on the other.

As an example of a social medium, Wikipedia is characterized by its task-orientation and by the fact that all of the interactants' "identity work" with respect to their identity in the medium is captured in the database. This, in turn, causes the data set to be rich in the type of social acts we are investigating. The dataset was used for research in automatic detection of forum claims, as presented in a companion paper (Marin et al., 2011). That work focused on using lexical features, filtered through word lists obtained from domain experts and through data-driven methods, and extended with parse tree information. Automatic detection of other types of authority claims and of alignment moves is left for future research.

We believe that, as social acts, authority claims and alignment moves are broadly recognized communication behaviors that play an important role in human interaction across a variety of contexts. However, because Wikipedia discussions are shaped by a set of well-defined, local communication norms which are closely tied to the task of distributed, collaborative writing, we expect authority claims and alignment moves will manifest differently in other genres. Future work could explore the range of variation among the linguistic cues associated with authority and alignment categories across genres, cultures and communication media, as well as the possible role of additional categories or social acts not discussed here. We believe that the communicative ecology of Wikipedia discussions, combined with the rich metadata of the Wikipedia database, presents a highly valuable natural laboratory in which to explore social scientific analyses of communication behaviors as well as a resource for the development of NLP systems which can automatically identify these social acts, in Wikipedia and beyond.

Acknowledgments

This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI or the U.S. Government.

The original Wikipedia discussion page data for this study was made available from a research project supported by NSF award IIS-0811210. We thank Travis Kriplean for his initial assistance with scripts to process this data dump.

We also gratefully acknowledge the contribution of the annotators: Wendy Kempsell, Kelley Kilanski, Robert Sykes and Lisa Tittle.

References

Apoorv Agarwal and Owen Rambow. 2010. Automatic detection and classification of social events. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1024–1034, Cambridge, MA. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Philip Ball. 2005. Index aims for fair ranking of scientists. *Nature*, 436:900–900.

Nancy Baym. 1996. Agreements and disagreements in a computer-mediated discussion. *Research on Language and Social Interaction*, 29:315–345.

Mary Bucholtz and Kira Hall. 2010. Locating identity in language. In C. Llamas and D. Watt, editors, *Lan-*

guage and Identities. Edinburgh University Press, Edinburgh.

Moira Burke and Robert Kraut. 2008. Mopping up: Modeling wikipedia promotion decisions. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, pages 27–36. Association of Computing Machinery.

Jolene Galegher, Lee Sproull, and Sara Kiesler. 1998. Legitimacy, authority, and community in electronic support groups. *Written Communication*, 15(4):493–530.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 669–676, Barcelona, Spain.

Meghan Lammie Glenn, Stephanie M. Strassel, and Haejoong Lee. 2009. XTrans: A speech annotation and transcription tool. In *INTERSPEECH-2009*, pages 2855–2858.

Erving Goffman. 1981. *Forms of Talk*. University of Pennsylvania Press, Philadelphia.

Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. 2003. Detection of agreement vs. disagreement in meetings: training with unlabeled data. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, pages 34–36.

Jakob L. Jensen. 2003. Public spheres on the internet: Anarchic or government sponsored; a comparison. *Scandinavian Political Studies*, 26:349–374.

J. Richard Landis and Gary G. Koch. 1977. Measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Yameng Liu. 1997. Authority, presumption and invention. *Philosophy and Rhetoric*, 30(4):413–427.

John Locke. 1959 [1690]. *An Essay Concerning Human Understanding*. Dover Publications, New York.

Jo Mackiewicz. 2010. Assertions of expertise in online product reviews. *Journal of Business and Technical Communication*, 24(1):3–28.

Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Detecting forum authority claims in online discussions. *Proceedings of the Workshop on Language in Social Media (LSM 2011)*.

Michael Mulkay. 1985. Agreement and disagreement in conversations and letters. *Text*, 5(3):201–227.

Michael Mulkay. 1986. Conversations and texts. *Human Studies*, 9(2-3):303–321.

Meghan Oxley, Jonathan T. Morgan, Mark Zachry, and Brian Hutchinson. 2010. "What I know is...": Establishing credibility on Wikipedia talk pages. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, Gdansk, Poland. Association for Computing Machinery.

Janie Rees-Miller. 2000. Power, severity, and context in disagreeement. *Journal of Pragmatics*, 32(8):1087–1111.

Kay Richardson. 2003. Health risks on the internet: Establishing credibility on line. *Health, Risk and Society*, 5(2):171–184.

Marie-Claire Shanahan. 2010. Changing the meaning of peer-to-peer? Exploring online comment spaces as sites of negotiated expertise. *Journal of Science Communication*, 9(1):1–13.

Jan Svennevig. 1999. *Getting Acquainted in Conversation: A Study of Initial Interactions*. John Benjamins Publishing Company, Amsterdam.

Dorothea K. Thompson. 1993. Arguing for experimental "facts" in science. *Written Communication*, 10:106.

Joanna Thornborrow. 2001. Authenticating talk: Building public identities in audience participation broadcasting. *Discourse Studies*, 3(4):459–479.

Linda Wine. 2008. Towards a deeper understanding of framing, footing, and alignment. *Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics*, 8(3):1–3.