

Are We There Yet?: The Development of a Corpus Annotated for Social Acts in Multilingual Online Discourse

Jonathan T. Morgan^a

{JMO25, WHAT, EBENDER, ZLIYI, GRACHEVA, ZACHRY}

Meghan Oxley^b

@UW.EDU

Emily M. Bender^bLiyi Zhu^bVarya Gracheva^bMark Zachry^a^a*Department of Human Centered Design & Engineering, Box 352315*^b*Department of Linguistics, Box 354340**University of Washington**Seattle, WA 98195 USA***Editors:** Stefanie Dipper, Heike Zinsmeister, Bonnie Webber

Abstract

We present the AAWD and AACD corpora, a collection of discussions drawn from Wikipedia talk pages and small group IRC discussions in English, Russian and Mandarin. Our datasets are annotated with labels capturing two kinds of social acts: alignment moves and authority claims. We describe these social acts, discuss our annotation process, highlight challenges we encountered and strategies we employed during annotation, and present some analyses of resulting the data set which illustrate the utility of our corpus and identify interactions among social acts and between participant status and social acts in online discourse.

Keywords: Annotation, Linguistics, Discourse, Computer-Mediated Communication

1 Introduction

The application of machine learning to natural language processing problems has been very effective at automatic annotation of morphosyntactic structure and, in at least some application contexts, extracting relational semantic information and general sentiment. The work described in this paper is motivated by the question of whether these successes can be extended to the automatic identification of the social acts of discourse participants. An important first step in developing automated annotation methods is the manual annotation of the target structures, which in turn requires the development of operationalizable definitions. To that end, we have defined two sets of social acts that can be annotated at the discourse turn level and carried out annotation of these acts in six corpora, representing three languages (English, Mandarin, Russian) and two genres (asynchronous online forum discussions and synchronous online chat).¹

Reliable identification of social goals, individual motivations and discursive strategies of conversational participants presents both theoretical and practical challenges. Few theoretical frameworks systematically account for a meaningful and coherent set of the kinds of things discussion participants “do” with language as they interact across varied communication situations. When research from different fields addresses what people do with language, it often addresses a limited range of field-specific concerns that do not readily align with potentially

¹ The social acts and their annotation within the English Wikipedia corpus were described in Bender et al. 2011, along with some preliminary analysis. This paper expands that description and analysis as well as extending both to multilingual and multigenre considerations.

complementary ideas from seemingly related work in another field. Across the work in different fields, there is little guidance in terms of defined categories that can be operationalized for annotation and thus allow for systematic investigation of the specific word, sentence, and turn-level structural properties of the language used in the social acts. Likewise, it is difficult to develop and apply annotation guidelines to reliably label such pragmatic-level phenomena because of their inherent variability and sensitivity to contingencies introduced by the medium, genre and language of the communicators.

This paper describes the development of such a theoretically-motivated annotation effort, resulting in the Authority and Alignment in Wikipedia Discussions (AAWD) corpus and the Authority and Alignment in Chat Discussions (AACD) corpus.² These corpora contain online discussions in English, Mandarin and Russian annotated for two types of social acts: authority claims and positive/negative interpersonal alignment moves. The AAWD corpus includes annotation of these social acts in asynchronous, language-based interactions among editors on Wikipedia article discussion pages. The AACD corpus includes annotation of the social acts in synchronous, textual interactions among small groups of individuals engaged in collaborative planning through the medium of Internet Relay Chat (IRC).

The annotations described here were produced with both engineering and scientific goals in mind. On the one hand, they can serve as training and test data for machine learning, that is, the automatic detection in text of the phenomena that we are annotating by hand (see for example Marin et al. 2011). To the extent that these social acts (and others of a similar granularity) are deployed in various combinations towards larger social goals of conversational participants, the ability to detect them in naturally occurring discourse can be an important step towards automatic recognition of those larger social goals. For example, automatic detection of alignment moves and authority claims could facilitate the detection of sections of multiparty interactions where the participants take sides regarding some particular issue (which in turn could help in the identification of contentious issues); interactions in which participants attempt to establish themselves as credible sources in order to make power plays (Kriplean et al. 2007); and in identifying potential trolls, scapegoats or other participants who tend to become the focus of a strong negative response from an entire group. Similarly, as social acts such as alignment moves and authority claims are means by which individuals perform roles (e.g., moderator, expert or novice), automatic detection of social acts can contribute to the automatic detection of fluid roles in conversation exchange. To the extent that the distribution and form of the social acts vary across genres, the inclusion of different genres in annotation guideline development and in the annotated corpus is critical to the development of robust detection algorithms.

In addition to the engineering goals of automatic detection of social acts, this annotated corpus holds scientific interest. We offer operational definitions of our social acts such that they can be annotated at the level of a turn (or sub-turn utterance), but not directly in terms of the linguistic form that they take. This approach allowed us to apply the same definitions across genres and across languages (while allowing the annotation guidelines to develop as we moved to these new corpora) and thus to compare the distribution and linguistic form of the social acts across genres and languages. In this article, we report some preliminary comparative analysis across Mandarin, English and Russian language Wikipedia discussions, and present some initial data and observations on the distribution and presentation of social acts in synchronous Internet Relay Chat discussions in English. However, these findings are intended to be illustrative rather than definitive, and certainly not exhaustive. There is still much more that could be done with these corpora, which we have made available for anyone who is interested in working with it.

In this paper, we define the social acts of interest, describe our two corpora and annotation scheme, and highlight strategies and challenges that emerged during our annotation process. Our

² Both available from <http://ssli.ee.washington.edu/projects/SCIL.html>

analysis considers the distribution of social acts across three languages and two genres, and explores hypotheses to illustrate potential interactions between social acts and related social phenomena. We draw on existing literature in linguistics, communication and rhetoric to define our two types of social acts, alignment moves and authority claims. From linguistics and organization studies, we draw on the concept of “identity work” to motivate our focus on these two types of social acts and guides our hypotheses regarding their potential characteristics and interactions. We describe how we generated and refined our annotation guidelines iteratively during annotation, drawing on observations and feedback from the annotators as they applied each version of the guidelines and we highlight effective strategies and unanticipated challenges we encountered. We focus primarily on those strategies and challenges that relate to achieving reliable annotation of discourse-level phenomena and on localizing a set of annotation guidelines to different languages and genres. We present a set of preliminary analyses that illustrate potential language- and genre-mediated differences in the expression and distribution of alignment moves and authority claims. Finally, we test hypotheses related to how the performance and reception of a social act on Wikipedia is shaped by aspects of the contributor’s identity, such as their official status within the community and their level of experience.

2 Social Acts

The annotation of social acts differs from other annotation efforts concerned with linguistic structure or even linguistic meaning. In the creation of a syntactic treebank (e.g., the Penn Treebank, Marcus et al. 1993), annotators are concerned with the linguistic form of the utterances they are annotating. The meaning is also relevant in that annotators use their understanding of each utterance in context to select among different possible syntactic structures to annotate. The annotations, in turn, make explicit aspects of the linguistic structure. Similarly, annotation efforts concerned with lexical and propositional semantics make explicit parts of the linguistic meaning (again, as disambiguated by context) closely aligned with sentence structure or with particular words (e.g., which phrases fill which semantic roles (Baker et al. 1998, Palmer et al. 2005), the scope of negation (Szarvas et al. 2008), or word sense (Snyder and Palmer 2004)).

In contrast, social acts concern not the form nor the linguistic meaning of utterances but what those utterances are used to accomplish. In this sense, annotation of social acts is akin to dialogue act annotations (Shriberg et al. 2004) or social events (Agarwal and Rambow 2010). Nonetheless, there are differences here as well. Dialog acts concern how the turns in a conversation connect to each other (e.g., question-answer pairs), whereas our social acts concern the social positioning of a discussant within a group. Indeed, dialog acts and social acts can be seen as orthogonal tiers of annotation: e.g., both a question and an answer to it can be positive or negative alignment moves. Our social acts differ from Agarwal and Rambow’s (2010) social events in that we are concerned with what discussants are accomplishing (or attempting to accomplish) with their utterances, whereas Agarwal and Rambow were looking to extract descriptions of interactions from narratives. In other words, our social acts are attributed to the interlocutors; their social events to entities described by an author.

Social acts, like dialog acts or social events, can be detected in part on the basis of the syntactic and semantic structure of the utterances that constitute (or, in the case of social events, describe) them. However, it is important to note that the linguistic structures available for performing social acts are highly varied, perhaps even more so than those available for performing dialog acts, which tend to be at least somewhat conventionalized (Searle 1975). Therefore, in creating annotated resources for social acts, it is neither feasible nor sufficient to annotate linguistic structures and then use these as the basis for annotating the social acts.

2.1 Identifying Social Acts on Wikipedia

Although Wikipedia, and wikis in general, are designed to make it easy for an individual to make content changes without consultation or prior approval (Cummings 2008), on many Wikipedia

Claim Type	Definition	Example
Credentials	Credentials claims involve reference to education, training, or a history of work in an area.	"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.	"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 AAWD or AACD corpora>
Forum	Forum claims are based on policy, norms, or contextual rules of behavior in the interaction.	"Do any of these meet wikipedia's [[WPRS 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.	"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.	"I think in the minds of most people, including the government, the word "war" and a formal declaration of war have come apart."

Table 1. Authority claims by type.

articles editors conventionally discuss and justify major changes that they make to article content, using other wiki pages as discussion forums (Viegas 2007). Clay Shirky has called Wikipedia "the product of unending argumentation... which grows not from harmonious thought but from constant scrutiny and emendation." (Shirky 2008) These discursive norms mean that Wikipedia editors must often perform complex negotiations about article content. They publicly align with or against other editors in the discussion by making statements that support or oppose proposals made by other editors, such as including a particular image in the article or re-writing the introductory section, or that express approval or disapproval of particular edits others have already made to the article. Editors must also justify changes they have already made, since any particular edit can be reverted or re-written by another editor at any time.

The two types of social acts described in this paper, authority claims and alignment moves, are therefore particularly relevant to understanding the dynamics of Wikipedia editorial discussions. However, we believe that the expression of authority and alignment is common across a variety of discursive contexts, especially contexts in which conversational participants are engaged in collaborative activities such as group decision-making, debate and production. Furthermore these social acts themselves exhibit a degree of regularity both within specific contexts (such as a Wikipedia discussion in English, with its singular jargon and conventions) and also across contexts. In the following sections, we describe the theoretical and empirical basis for authority and alignment, present the annotation schemes we developed for identifying them, and reflect on the annotation process.

2.2 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 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 their authority on the matter. These "arguments from authority" have long been recognized as an important component of informal logic by many language philosophers (Locke 1959 [1690], Liu 1997). The self-presentation of authority has also been empirically examined in a variety of spoken and written contexts by scholars from disciplines such as communication, rhetoric, health studies, sociolinguistics, linguistic pragmatics and political science (for instance, Galegher 1998, Jensen 2003, Mackiewicz 2010, Richardson 2003, Thompson 1993), providing a framework for understanding the strategies and conventions that communicators operating in different genres and media employ to establish themselves as credible discursive participants. Although the linguistic construction of authority claims can vary greatly within a single genre and across genres, the regularity in the types of claims that are made and the construction of claims facilitates empirical analysis.

Authority claims provide an interesting lens through which to view an authored text or a conversation transcript, as the overall frequency of claims can reflect the nature or purpose of the discourse (for example a task-oriented collaboration vs. an 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), discursive patterns related to the expression of authority 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 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 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. We classify authority claims into the following types: *Credentials*, *Experiential*, *External*, *Forum*, *Institutional* and *Social Expectations*. While our claim types were developed independently of existing classification schemes, several of our types do mirror categories used in previous research. Richardson's (2003) *warranting strategies* are particularly salient, particularly *Warranting by source* (similar to our 'External' claim type), *Warranting by reference to personal experience* (see 'Experiential', below) and *Warranting by reference to status* (see 'Credentials' and 'Institutional', below). Our codebook³ includes detailed definitions as well as positive and negative examples for each claim type. See **Table 1** for a list of our claim types, with brief definitions and examples.

2.3 Alignment Moves

In multiparty discourse, interpersonal 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

³ Available from <http://ssli.ee.washington.edu/projects/SCIL.html>

Move Type	Example	Baym (1996)
Positive Alignment		Markers of Agreement
Explicit Agreement	"I agree." "Yes." "That's right." "Correct." "Exactly." "Absolutely." "Of course." "I don't disagree." "I'll do that."	<i>Explicit Indicators of Agreement</i>
Praise/Thanking	"Great idea." "Good point." "Well said." "Thanks for your work on the article. It looks much better."	<i>Expressions of Gratitude</i>
Positive Reference to Previous Speaker's Point	"Like Bill was saying . . ." "As you mentioned earlier . . ." "As Mary was indicating . . ."	<i>Acknowledgment of Other's Perspective</i>
Other - Positive	Clear indicator of positive alignment that doesn't fit into one of the above categories	<i>Smiley Faces</i>
Negative Alignment		Markers of Disagreement
Explicit Disagreement	"I disagree." "No." "That's wrong." "That's false."	<i>Explicit Indicators of Disagreement</i>
Doubting	"I doubt that." "I don't think so." "That's questionable." "You can't be serious."	<i>Qualification</i>
Sarcastic Praise	"That's a GREAT plan. While you're at it, why not destroy the entire article?"	n/a
Criticism/Insult	"That's ridiculous." "That's a terrible idea." "You're nuts." "You fool."	n/a
Other - Negative	Clear indicator of negative alignment that doesn't fit into one of the above categories	n/a

Table 2. Alignment move types.

that participant while expressing alignment *against* another participant functions as a means of increasing 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* (Goffman 1981), may reflect long-term 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 (Wine 2008). This concept of alignment is a different phenomenon from Pickering and Garrod's (2004) alignment, in that it describes participants' attitudes towards one another rather than how they linguistically represent and align their mental models in order to perform successful communication.

Ways of expressing agreement and disagreement can vary according to a variety of social factors, including power relations among participants, gender, participant goals, and conversational context (Rees-Miller 2000). Research has suggested that expressions of agreement and disagreement in written language tend to be more explicit than oral expressions of agreement and disagreement (Mulkay 1985; Mulkay 1986). Text-based online discussions generally reflect this turn towards more explicit cues (Baym 1996) as participants compensate for the lack of the many non-linguistic communication mechanisms that are available in face-to-face interactions. In

asynchronous online group discussion forums such as those on UseNet and Wikipedia, the way agreement and disagreement are expressed is also mediated by the non-dyadic nature of the discussion. For instance, if an editor on a Wikipedia talk page disagrees with another editor, that disagreement is effectively public: it is equally visible to a large audience of (known and unknown) other conversational participants.

Depending on community norms about the acceptability of disagreeing, the public quality of interpersonal communication may lead a speaker to perform more explicit “facework” (Baym 1996, Brown and Levinson 1987) by either downplaying their disagreement (for instance, by hedging) or by exaggerating it, even to the point of rudeness or flaming. The presence of a large, participatory audience can also cue speakers to exhibit more “front stage” behaviors (Goffman 1959), exaggerating their (discursive) movements like an actor on a stage. Thus, a Wikipedia discussion participant is likely to exhibit a more regularized pattern of agreement and disagreement as they shape their language to adhere to local norms and adopt the jargon of their anticipated audience. Different languages also reflect differing conventions for expressing agreement and disagreement (see, for example, Mori 1999), which constitutes another mediating factor in how these social acts are performed in a given situation. Our annotation scheme contributes to existing research on the role of conversation context, medium and language in shaping online discourse by accounting for a range of alignment cues in two types of text-based, task-oriented online discussions across three languages.

We classify alignment moves into positive and negative types, according to whether the participant is indicating direct (explicit) or general (implicit) agreement or disagreement with the target: Positive alignment is annotated in cases of *explicit agreement*, *praise/thanking*, *positive reference* to another participant’s point or where other clear indicators of positive alignment are present. Negative alignment is annotated in cases of *direct disagreement*, *doubting*, *sarcastic praise*, *criticism/insult*, *dismissing*, or where other clear indicators of negative alignment (such as typographical cues) are present. We did not explicitly adopt our alignment sub-types from any previous researchers’ classification schemes. Instead, our categories were developed through iterative exploratory qualitative analysis of broadcast talk show transcripts from the GALE broadcast speech corpus⁴ and a sample of Wikipedia discussion pages. However, like our authority claims, many of our alignment sub-types are similar to those defined by other researchers. In particular, Baym (1996) recognized a set of alignment cues that mirror our own. A summary of our alignment move types and related markers identified by Baym is presented in **Table 2**. A more detailed list of alignment types, cases and definitions are available in the codebook included with our corpus at the URL previously specified.

2.4 Social Acts and Identity Work

“Identity work,” a sociological concept common to both organization studies and sociolinguistics (Alvesson and Willmott 2002, Bucholtz and Hall 2010), describes the way in which individuals’ social identities are constructed, reflected and transformed through their own communication practices and those of the people with whom they interact. Both Wikipedia talk pages and IRC channels are contexts in which participants often interact with others whom they only know within that context, or with whom they have never interacted before. Online identities are often fluid and transitory even in cases where they are persistent within the system (for instance, in the form of a user name, handle or pseudonym) since a user may use multiple identities, and identifying features (such as a users’ profile information) can be hidden or altered at any time. In these spaces participants also lack many common mechanisms used in face-to-face contexts to communicate attention, emotions, roles, and goals. For instance physical gestures, eye contact, facial expressions, posture, and vocal inflection are not available in these textually-mediated

⁴ <http://projects.ldc.upenn.edu/gale/data/Catalog.html>

environments. Additionally, each participant’s identity is less likely to be pre-established and persistent in Wikipedia and IRC discussions than in discussions among known others, such as in business meetings. In such spaces we expect the discursive practices associated with “identity work” to be more explicit, as discussion participants constantly re-present aspects of their personal values and their social status, their attitudes and allegiances in the text they type.

Given the few visible markers of identity on Wikipedia and the fact that editors are constantly interacting with new collaborators, Wikipedians perform authority by adopting insider language and other community-specific norms of interaction related to the task of collaboratively writing an encyclopedia (see for example Kriplean et al. 2007). Supporting arguments with specific references is one such norm. In order to investigate the relationship between social acts and identity work, we explore the extent to which Wikipedia editors’ authority claims reflect certain socially salient aspects of their identity as Wikipedians. We use two particular social identity measures, user role and their degree of experience, which are manifested in specific ways within Wikipedia. We hypothesize that as editors become more integrated into Wikipedia, they will make more authority claims. In order to test this hypothesis, we leverage metadata about individual Wikipedia editors that is captured in our dataset but is not immediately visible to other participants: user roles (for instance, administrator, registered editor, anonymous editor), total lifetime edits, and length of membership.

We also describe “v-index,” a measure of an editor’s degree of integration, investment or “veteran status” within the Wikipedia community at a particular point in time. 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 edit (in this case, a conversational turn) is the greatest v such that the editor has made at least v edits to Wikipedia within the past v months (28-day periods). Our tests of these hypotheses about the relationship between identity and social acts are described in Section 5.

3 Data collection

We gathered English language Wikipedia data first, and then gathered similar samples from the Russian and Mandarin Wikipedias. The AACD corpus was gathered later, and is intended to offer meaningful comparisons and contrasts with the AAWD corpus. We gathered the AACD by facilitating a set of task-based group IRC discussions using participants we recruited locally.

3.1 Wikipedia Talk Page discussions

Wikipedia talk pages (also called discussion pages) are editable pages on which Wikipedia editors can take part in threaded, asynchronous discussions about the content of other pages, particularly the article pages that most visitors to Wikipedia are familiar with. Every article page in Wikipedia has an associated talk page wherein editors can discuss and collaboratively plan editing actions on that article; any editor interested in a given article can join the conversation on that article’s talk page. Conversational exchanges on the talk pages may take the form of a polite deliberation aimed at achieving a final consensus-based decision or a heated argument as editors advocate different ideas about matters concerning 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 forums (such as IRC channels 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 other 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

English	Counts
Total annotated discussions	185
Total turns	3361
Turns w/ external claim	459
Turns w/ experiential claim	77
Turns w/ forum claim	260
Turns w/ credentials claim	3
Turns w/ social expectations claim	21
Turns w/ any claim	703
Mandarin	
Total annotated discussions	225
Total turns	1517
Turns w/ external claim	60
Turns w/ experiential claim	24
Turns w/ forum claim	77
Turns w/ credentials claim	7
Turns w/ social expectations claim	3
Turns w/ any claim	155
Russian	
Total annotated discussions	122
Total turns	893
Turns w/ external claim	67
Turns w/ experiential claim	7
Turns w/ forum claim	19
Turns w/ credentials claim	2
Turns w/ social expectations claim	4
Turns w/ any claim	82

Table 3. Total authority claims in AAWD corpus by language.

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).

Our English-language Wikipedia dataset is drawn from a publicly-available 2008 Wikipedia XML data dump⁵ and is composed of 365 discussions associated with 47 talk pages. These articles were selected based on a list of topic keywords extracted from a set of English-language

⁵ http://en.wikipedia.org/Wikipedia:Data_dumps

	N	%
English		
Total turns	2890	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 making alignment moves	315	31.9
Russian		
Total turns	1806	100
Turns w/ positive alignment	558	31
Turns w/ negative alignment	142	8
Turns w/ any alignment	645	36
Mandarin		
Total turns	2767	100
Turns w/ positive alignment	808	29
Turns w/ negative alignment	153	6
Turns w/ any alignment	925	33

Table 4. Total alignment moves in AAWD corpus by language.

broadcast news transcripts in the GALE broadcast speech corpus. A Python script was used to query the Google search engine to identify English language Wikipedia articles related to those keywords. Because the Russian and Mandarin language editions of Wikipedia have fewer users and therefore less talk page discussion over all, our Mandarin and Russian datasets consist of discussions from the talk pages on those language editions of Wikipedia that had the most edits, rather than articles specifically related to those in the English language dataset.

All the selected discussions contain at least 5 conversational turns and at least 4 human participants.⁶ We set this minimum threshold for discussion length and number of participants because we expected that our social acts would be most evident in discussions where there was a higher level of interactivity. Because not all Wikipedia articles are highly collaboratively created, contested, or frequently updated, many talk page threads do not meet these minimum criteria. Of the 365 discussions in our final English dataset, 185 were annotated for both alignment moves and authority claims. The Mandarin and Russian-language versions of Wikipedia were annotated for both authority and alignment in 225 and 122 discussions, respectively. See **Table 3** for a breakdown of authority claim annotation across the three languages, and **Table 4** for a breakdown of alignment move annotation in English.

All annotation was completed by paid university students, and took place between 2009 and 2011. We anticipated that it would be difficult to perform annotation on both types of social acts (alignment and authority) simultaneously, so annotators were instructed to only annotate a discussion for one social act at a time.

3.2 IRC discussions

Our chat data is based on a set of 12 textual, synchronous exchanges among four-person groups chatting in a private IRC channel. These exchanges were all facilitated by the researchers in an effort to develop a corpus of language use examples that would share some characteristics with the Wikipedia discussion page corpus described above.

In each of the 12 sessions, 4 individuals interact through Internet Relay Chat (IRC) for approximately 45 minutes. These sessions were all saved as time-stamped transcripts. The chat

⁶ Some of the turns in Wikipedia discussions are actually contributed by automated agents, called “bots.”

datasets include four 45-minute exchanges in English, four in Mandarin, and four in Russian, for a total of 48 total participants across languages. All participants were native speakers of the language in which the session was conducted and were comfortable typing their language on a standard QWERTY keyboard using standard Microsoft Windows 7 language packs to remap keys to non-English character sets, as appropriate. All sessions occurred between August 2010 and August 2011. Participants were primarily recruited through university list-serves and by word-of-mouth. Additional participants were recruited through paid advertisements in a university daily newspaper, both in print and online. All participants received compensation for their participation in the form of a \$25 gift card.

Given that our chat participants were primarily recruited through word of mouth and were therefore not sampled randomly, there are notable demographic differences between the three language groups. The mean ages of our English, Mandarin, and Russian language groups were 20, 26, and 30, respectively. Education levels varied between groups; 13% of English participants, 100% of Mandarin participants, and 81% of Russian participants had earned a postsecondary degree.

Participants also differed somewhat in their use of online chat systems. While 94% of the English participants and 100% of the Mandarin participants used online chat systems on a daily or weekly basis, the Russian participants used online chat systems less frequently, with only 69% of participants reporting that they used online chat systems daily or weekly. Participants in all three groups reported that they primarily used online chat for communicating with people that they know offline (94%), and slightly under half of the participants also used online chat for conducting meetings (42%).

Session Number	1	2	3	4	Total Claims
Session Code	5ne	4ne	3ne	2ne	
Authority					
External Authority	8	9	4	2	23
Experiential Authority	7	5	2	5	19
Social Expectations Authority	2	1	2	0	5
Forum Authority	1	0	0	1	2
Credentials Authority	0	2	0	0	2
Column Totals	18	17	8	8	51
Alignment					
Positive Alignment	59	56	73	53	241
Negative Alignment	29	38	7	7	81
Totals	88	94	80	60	322
Total Turns in Chat Session	646	654	534	395	2229

Table 5. Authority claim types and alignment moves across chat sessions for English.

	AAWD	AACD
Computer-mediated Communication Type	Asynchronous interaction	Synchronous interaction
Time Constraints	Open-ended	45 Minutes
Previous Interaction Among Participants	Primarily Online	Primarily Offline
Conversational Roles	Loosely-structured: unregistered (IP), registered editor (username), administrator (username)	Pre-assigned: project manager, publicity coordinator, secretary, accountant
Task Type	Collaborative writing	Collaborative event planning
Number of Participants per Conversation	Four or more	Four
Motivation for Participating	Uncompensated	Compensated
Conversation Topic	Determined by related article topic	Pre-assigned

Table 6. Genre and Task Differences in the Wikipedia and chat datasets.

A spreadsheet containing all demographic information collected on the chat participants is provided with the AACD at the URL previously specified.

The sessions were stimulated by the researchers, using a common scenario for the participants across all sessions. At the beginning of each session, participants were assigned to one of four different discussion roles: *project manager*, *accountant*, *publicity coordinator*, or *secretary*. A brief description of the responsibilities incumbent on each role were provided in the scenario prompt, which is available with our dataset. Our chat scenario was thus designed to reflect in a synchronous forum some of the features of Wikipedia talk page discussions, which often address specific considerations and decisions about Wikipedia articles and involve multiple editors with different knowledge sets, motivations and points of view. In each chat session, the participants were asked to work together toward a shared goal (the planning of a party for students in a large university lecture course), where each participant’s tasks and concerns differed according to their assigned role.

Each of the discussions in this collection was independently annotated for authority claims and alignment moves by a native speaker researcher. Because these IRC discussions were not dually annotated, these data and the preliminary analyses we provide below should be considered provisional and illustrative of the potential for cross-genre comparison.

3.3 Genre Similarities and Differences

Together the AAWD and AACD corpora represent productive resources for exploring social acts in online language use. We present these corpora together in order to provide opportunities to explore the way genre, medium and language shape social acts. Both corpora include interactions among individuals who are loosely bound by a collective orientation to accomplishing the same task (either creating an encyclopedic article or planning an event). The shared task nature of these interactions provides some constraint on the topical focus of each exchange set though individual participant contributions are not further constrained. In both corpora the interactants rely on

online language to make their individual intentions known to others and to affect the group’s thinking about the task at hand. Still, important differences also distinguish the two corpora. The AAWD data represents asynchronous interactions, while the AACD data is synchronous interactions. Related to this, in the AAWD data there is no time constraint related to the accomplishment of the motivating task, but in the AACD data that participants had a set time (45 minutes) to work toward their task goal. Genre- and task-mediated differences between the two datasets are summarized in **Table 6**.

4 Annotation Process

In this section we both present the results of our annotation and analysis, and reflect on the benefits and drawbacks of our framework and process. We find that while the social acts we chose to capture exhibit some regularity, even after extensive guideline revision and annotator training we were not able to achieve Cohen’s Kappa scores of higher than 0.6 for most of our social act sub-types. One of the greatest challenges in our annotation effort was to reliably identify instances where a social act was present, and to set meaningful boundaries that presented annotators with as few problematic edge cases as possible. Encouragingly, however, we find that in instances where multiple annotators annotated a particular utterance as containing a social act, we saw much higher agreement on the type of social act presented in that utterance, suggesting that in more prototypical cases, elements of each social act were relatively consistent in their presentation.

We will also discuss the process of localizing our guidelines, originally developed by native English speakers with reference to English language data, to Mandarin and Russian, and how this localization process was guided by differences in the manifestations of authority and interpersonal alignment in different languages but within the same genre.

4.1 Guideline Development

We developed our initial authority claim and alignment categories through qualitative examination of English language Wikipedia discussions outside of our dataset, as well as transcripts of political talk show broadcasts in the GALE corpus. Two of the researchers developed an initial data-driven set of authority claim types, and a third researcher expanded this typology through additional analysis of talk pages, guided by established constructs related to the self-presentation of authority from rhetorical theory and informal argumentation. Our authority claims are not intended to map directly onto categories established in previous work (for instance, classical rhetorical appeals to *ethos*, *logos* and *pathos*). Rather they are intended to capture the ways in which these universal discursive moves manifest in topic-focused debates and task-based deliberation across a variety of online contexts. Furthermore, although our goal was not to apply existing categorization schemes to our data, some of the authority claim types we developed are similar or complementary to categorizations employed by other researchers (for example Mackiewicz 2010, Richardson 2003). See Section 2 above for additional citations.

We identified candidate sub-types of alignment moves through a review of the literature on conversational agreement and disagreement, especially the work of Brown and Levinson (1987) and Baym (1996). This candidate list was vetted and supplemented and through sample annotation by two of the researchers, and in response to student annotators’ observations. Alignment sub-types proved less easily discriminable than our authority claim types, leading us to treat them as illustrative cues rather than an exhaustive and mutually-exclusive set of categories. Alignment cues were used to identify the presence of alignment in a turn, and to discriminate between instances of positive and negative alignment. Our process for refining these guidelines is described in further detail below.

Our initial English guidelines included simple descriptions of the phenomena to be annotated, but only a few examples. As coding progressed, more examples were added to provide necessary clarification to the social acts categories as annotators raised questions during weekly annotation

meetings. We developed both positive and negative examples in order to clarify the boundaries between social act categories (for instance, to distinguish between an external and an experiential authority claim), and to establish heuristic thresholds to help annotators decide what qualified as a social act (for instance, whether the use of the word “yeah” always counted as positive alignment, or whether other contextual features were necessary for determining whether or not it was merely a backchannel utterance.). We summarize and illustrate some of these adaptations below.

4.1.1 Differentiating alignment from similar or contrary opinion

While testing out our guidelines, we observed that it can be difficult to differentiate between a turn in which a speaker expressed a similar or opposing opinion to a previous turn, and one where the turn-taker was actually aligning with or against a previous speaker. In many cases on Wikipedia, a speaker will express a similar idea as a previous talk page participant, or make an alternative, contradictory proposal without explicitly orienting their statement towards (or making any reference to) the person they are supporting or contradicting. In order to increase consistency in annotation, we required annotators to find substrings within the utterance which explicitly mark it as alignment and identify them with keyword tags, in order to mark a turn as containing alignment.

4.1.2 Identifying sarcasm

Sarcastic statements can be difficult to recognize reliably. During both guideline development and refinement, we encountered cases where one annotator labeled an utterance as sarcastic and another did not, and they were not even able to successfully reconcile their opinions during open discussions at annotator meetings. To address this issue, we limited the types of sarcasm which we labeled as a potential alignment cue to *sarcastic praise*. We also prompted our annotators to look for typographical cues related to sarcasm, such as bolded or italicized words or the use of CAPS (for example, “Oh, sure, that’s a GREAT idea.”)

4.1.3 Specifying personal experience

Identifying experiential claims reliably also proved difficult initially. We found that discussion participants often related events that they had experienced, or things that had happened to them outside of the context of establishing authority. In order to aid in identifying references to personal experience that were specifically linked to authority claims, we limited our annotators to experience-related utterances that contained a first or second-person pronoun, such as “because I was living there at the time” or “we never used to say it that way.”

4.1.4 Naming social groups

Social expectations claims, though infrequent in our data, proved difficult to label consistently. One way we attempted to address this difficulty was to require our annotators to only label a claim as “social expectations” when it contained a reference to a named group, such as “Wikipedia readers”, “The GOP” or “Iowa voters.” This helped guide annotators in situations where the group being named was ambiguous or implied, such as in “some editors think” or “they won’t be convinced by mere facts.” We also found that these discussions included many statements about what named groups were doing or had done in the past, which were often not associated with claims of authority, such as “the media overemphasized his role” or “the administration is acting swiftly.” We addressed this by emphasizing with additional examples that our definition of ‘Social Expectations’ required a claim that made a statement about what groups *thought*, *believed* or *desired*.

4.1.5 Connecting moves with targets

It is difficult to identify whether an utterance contains an alignment move without a clear indication of who is being addressed. In order to help our annotators differentiate clear alignment

moves from more ambiguous cases, we required that each labeled move meet basic criteria related to identifying a target of the utterance. Alignment moves had to include either a named target (for example, “I doubt that will work, Tom”) or some other unambiguous personal reference such as a second person pronoun *and* be situated in the discussion thread in such a way that the person the speaker was referring to was clear from the context (such as the editor who made the immediately previous post, or the editor who started the thread). Research on the role of ‘facework’ in interpersonal communication, as presented in Brown & Levinson (1987) and discussed extensively in Baym (1996), provides support for our personal pronoun requirement: using personal references, such as attributing an idea to a person through use of a personal pronoun, is a stronger indicator of negative alignment than making an oblique comment on an idea because it constitutes an explicit threat to the target’s *positive face*.

All social acts were annotated at the “turn” level, with each “turn” representing a single message in a discussion thread. Each turn could contain multiple instances of a social act: for instance, an author could make an experiential and an external claim within the same turn. In order to capture these phenomena, all utterances that contained claims were labeled with keyword tags, as in **Example 1**:

Example 1.

“<claim1=experiential>*I’ve read up on this*</claim1><claim2=external”>*and the most recent New York Times op-ed says that Biden was right.*</claim2>”

Individual authority claims were identified and marked by contiguous keyword spans within a single sentence. We found that cues to alignment with or against a specific target tended to be less regularly expressed, and were often scattered across multiple sentences. In such cases, our use of keyword tags to capture words and phrases indicative of alignment allowed us to label our data accurately and flexibly. Multiple alignments with or against multiple targets could be annotated at the level of the entire turn, and one or more keyword spans associated with each alignment move could be marked across different sentences within the turn. For example:

Example 2.

“<k1 polarity=pos target=Speaker2>*That’s right, Speaker2*</k1>. <k2 polarity=neg target=Speaker3>*Speaker3’s way off base*</k2>, *but* <k1 polarity=pos target=Speaker2>*you seem to have a good solution*</k1>. <k3 polarity=neg target=Speaker2>*However disagree with your name for the section*</k3> – *‘Iraq War’ is used in the United States media and should be used here as well.*”

Therefore, marking alignment moves at the turn level, allowing multiple alignment moves per turn and identifying a single alignment move by multiple keyword spans, allowed annotators to capture instances where an author expressed both positive and negative alignment towards the same target within the same turn, as in **Example 2** above.

Rounds of Mandarin Annotation	Authority Agreement (Before Review)	Alignment Agreement (Before Review)
Annotator group 1, final round (5/2011)	0.56	0.61
Annotator group 2, 1st round (7/2011)	0.13	0.30

Table 7. Comparison of inter-annotator agreement between Mandarin annotator groups.

4.2 Annotation tool

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. For authority, claims in separate sentences of an annotation unit counted as separate even if they were of the same type.

In some cases, the imperfectly re-created threaded structure of our data in XTRANS made the target of an utterance difficult to identify, so we also included an optional hyperlink to the text of each turn on the live Wikipedia website to allow annotators to view the turn in context.

4.3 Guideline Refinement

We refined our social act categories iteratively, through weekly group annotator meetings and by setting up an annotator email list to which the annotators could send questions as they worked. One of the primary difficulties these meetings addressed was to circumscribe ambiguous or hazy categories by creating firm rules about what did and what did not count as an instance of a social act. These rules were based on common patterns we saw in the data, and generally took the form

Mandarin Annotation Rounds (Group 2)	Authority (Before Review)	Alignment (Before Review)
First round, 7/2011	0.13	0.30
Second round, 7/2011	0.44	0.42
Third round, 8/2011	0.72	0.64
Russian Annotation Rounds	Authority	Alignment
Fourth round, 6/2011	0.56	0.52
Fifth round, 7/2011	0.52	0.58

Table 8. Longitudinal comparison of inter-annotator agreement for second Mandarin annotator group (top) and Russian annotator group (bottom).

of structural or linguistic cues (see examples above). For instance, in order to remove ambiguity about the target of an alignment move, annotators were only allowed to specify a target if that user was mentioned by name, if the alignment target was the author of the previous post in the thread, or if the target was the author of the first post in the thread. This narrowing of the possible targets available to annotators increased consistency in target identification and greatly simplified the process of target identification for our annotators.

Annotator turnover and guideline refinement both affected inter-agreement (measured as averaged Cohen’s Kappa). To illustrate this challenge, we present inter-annotator agreement data from the Mandarin annotation project carried out between May and August of 2011. In May 2011, two of the three Mandarin annotators left the project. In late June, two new annotators were trained and joined the project. Comparisons of the inter-agreement rates for the last round of the old annotators and the first round of the new annotators are provided in **Table 7**.

The new Mandarin annotator team witnessed the second Mandarin guideline refinement and other minor guideline modification. As noted, these changes were based on the annotators’ questions and comments. As the Mandarin guidelines were refined, the inter-annotator agreement rates improved. In the case of Russian annotation however, discussion and guideline refinement did not necessarily lead to increased inter-annotator agreement rates. Longitudinal comparisons of annotator consistency in Mandarin and Russian are presented in **Table 8**.

4.4 Adaptation of guidelines to Mandarin and Russian

The Mandarin and Russian annotation guidelines were adapted based on the English guidelines. The Mandarin guidelines went through two major rounds of refinement. In the first round, a researcher who was a native speaker of Mandarin adapted the English language guidelines and applied the preliminary guidelines to 5 sample annotations. Adaptations included adding examples of emphatic markers that are used to indicate disagreement in Mandarin. The researcher also removed the statement (included in the English guidelines) that “no” can indicate agreement if the preceding statement is negative because “no” always indicates disagreement in Mandarin. The second round of refinement took place after the Mandarin annotators finished their first round of annotation. General questions and observations that were brought up in the annotation meeting were added to the guidelines, such as the use of rhetorical questions/paired conjunctions to indicate negative alignment. Beyond these two major rounds of refinement, guidelines were iteratively refined based on annotator feedback and regular “spot checks” of annotated data.

When the first drafts of the Mandarin and Russian guidelines were finished, the editors annotated 5 randomly selected Mandarin and Russian Wikipedia discussions. This process quickly revealed further issues with the guidelines. For instance, it was found that people sometimes disagree with a Wikipedia item, but not a particular person. In the initial guidelines, there was no clear stipulation that stated whether attacking a Wikipedia item should be coded as a negative alignment or not. It was decided then to add a stipulation to the Mandarin and Russian guidelines that: “any alignment move that agrees/disagrees with an article, or Wikipedia, should NOT be coded.”

As with the English annotation guidelines, problems uncovered in the sample annotation pass were addressed through guideline modification. Additional modifications were made during the annotation process, with earlier discussions re-annotated to maintain consistency. Care was taken to keep the overall alignment move types consistent across all three languages, even as modifications were made to the types of alignment cues.

The researchers working on the guidelines for Mandarin and Russian found that some of the behaviors described were common between English and Mandarin/Russian. Therefore, they first translated the culturally shared keyword strings from English to Mandarin/Russian. For example, all three languages have very similar repository of explicit agreement/disagreement words, like “I agree 我同意 Я согласен”, “Yes 是的 Да”, “I disagree 我不同意 Я не согласен”,

“No 不是 Hei.” There do exist differences, however. For example, in the English alignment guidelines, there is a note that “no” can indicate agreement if the preceding statement is negative. But in Mandarin, the situation is the opposite. Specifically, in English, the answer “no” is fact-oriented, while in Mandarin, it is speaker-oriented. A simple example is:

Example 3.

(Suppose Mary is not a student.)

English

A: Mary is not a student.

B: **No.** She is not.

Mandarin

A: Mary is not a student.

B: **Yes.** She is not.

Therefore this note was removed from the Mandarin alignment guidelines.

Another example that shows language differences is that in Mandarin, rhetorical questions are frequently used in daily speech to indicate negative opinions. Such rhetorical questions often have explicit lexical markers either at the beginning or inserted in the middle of the sentence, like “难道...? How can/Can't...?” (at the beginning of a sentence), and “怎么就...? How come...?” (in the middle of a sentence). Below is an example:

Example 4.

A:	如果 汉族的发源地	只有 黄河中下游的话,
	If Han's birthland	only Huang River's lower and middle reaches,
	那 只能 说明	今天的 中国 辽阔 领土 来源于
	then only indicate	today's China's large territory comes from
	中央政权	对 其它 民族的 战争 与 征服。
	central government	to other peoples' war and conquering.

‘If the Han people’s birthland is only Huang River’s lower and middle reaches, it could only indicate that China’s current large territory comes from the wars launched by the central government and their conquering towards other minorities.’

B:	不征服	怎么	来	土地?
	No conquering	how come	get	land?

‘If they didn’t win through conquering, how could they get land?’

As a result, one extra category named “rhetorical question” was added under the Negative Alignment Cues in the Mandarin alignment guidelines while there is no such category in the English guidelines.

In order to make the Mandarin and Russian Guidelines more comprehensive and representative, more examples were added from early rounds of Mandarin/Russian annotation. Guidelines were written in English, with examples in Russian/Mandarin, each with English translations, to ensure that they were easily accessible to the researchers and annotators across all

three languages. Like English, Mandarin went through several ‘test’ rounds of annotation starting from early 2010 using an early version of the social act guidelines. This annotation was based on samples drawn from both Wikipedia discussion pages and political talk shows in the GALE corpus. When the Mandarin guidelines were revised and evaluated in preparation for the primary annotation task in 2011, the Mandarin-speaking researcher viewed roughly 30 files that had been annotated in 2010 and selected a batch of keywords that seems to occur frequently in Mandarin Wikipedia discussions. These keywords were added to the Mandarin guidelines as examples.

The Russian-speaking researcher followed a similar process, examining data from randomly selected Russian Wikipedia discussions and basing the keywords and examples for Russian annotators on this data. As a result, the Russian guidelines used two sources of keywords and examples: Russian data examined by the Russian guidelines editor and examples that were either translations of English examples or were created by the researcher who led the Russian annotation effort in order to account for conversation topics that were not encountered in the initial set of Wikipedia examples, but were likely to appear in the Russian data.

4.5 Annotation Quality

In complicated annotation tasks, such as those conducted in this work, evaluating annotation quality 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

	N	κ	A
Authority Claims			
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
<i>Overall</i>	1157	0.59	0.86
Alignment Moves			
Explicit Agreement	379	0.62	0.94
Praise/Thanking	117	0.6	0.98
Positive Reference	86	0.2	0.98
Explicit Disagreement	453	0.29	0.92
Doubting	198	0.23	0.96
Sarcastic Praise	38	0.3	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
<i>Overall</i>	1378	0.5	0.8

Table 9. Agreement summary for authority claims and alignment moves in the English language Wikipedia dataset. N denotes the number of turns of the given type that at least one annotator marked.

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 more than two annotators have labeled a set of instances, we compute the average of pairwise agreement.

κ scores for authority claim and alignment move agreement in English are presented in **Table 9**. The counts, N , presented in **Table 9** reflect the total number of turns marked as containing claims of a given type by *any* annotator, rather than the total counts of claims that were independently annotated by two annotators (which are presented in **Table 3**). For authority, the most common types of claims, forum and external, are also two of the most reliably identified. For alignment, better agreement was demonstrated for the positive alignment sub-types than for the negative sub-types, but agreement was lower in general. The difficulty of making fine distinctions between the types of negative alignment move (for instance, discriminating between *sarcastic praise* and *criticism/insult*) appears to be 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 0.8 cut-off tentatively suggested by Artstein and Poesio (2008).⁷ 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 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.

The AAWD corpus includes both social acts that were identified by at least two annotators working independently, as well as those that were identified by only one annotator. We expect our multiply-annotated authority claims and alignment moves to correspond to more blatant or prototypical examples and our singly-labeled moves and turns, while sometimes being genuine noise, to pick out more subtle examples. The AACD corpus contains only singly annotated social acts. A single native-speaker researcher annotated each IRC discussion.

4.6 Lessons Learned

Through the process of creating and testing our annotation guidelines, we developed several strategies which were effective for resolving ambiguity within the guidelines, minimizing the cognitive load of the annotation task for our annotators, and increasing the overall quality of the annotation produced. These strategies are outlined below.

4.6.1 Treat guideline development and annotation as iterative processes

No researcher is omnipotent, and issues with the annotation guidelines are bound to arise in any annotation project, particularly when a set of guidelines developed on one language and genre is applied to new data. Our researchers met with annotators on a weekly basis to identify ambiguities within the guidelines and to reach consensus on how those ambiguities should be resolved. We also set up a mailing list for annotators to email us questions as they arose during annotation and for the researchers to provide comments and feedback. This iterative approach necessarily entails revisions to the guidelines to improve clarity and updates to the annotation after points of contention have been ironed out. Although this process can be time consuming, we agree wholeheartedly with MacQueen et al. (1998, p.36) that “re-coding should not be viewed as a step back; it is always indicative of forward movement in the analysis”.

4.6.2 Where possible, reduce cognitive load for the annotators

Although most of our annotators performed annotation of both of our social acts, alignment and authority, they were not asked to annotate both simultaneously. Our annotators found it easiest to

⁷ Artstein and Poesio also note that it may not make sense to have only one threshold for the field.

perform one type of annotation for several weeks rather than alternating frequently between different tasks. The annotation tool was set up so that when annotators chose to annotate a particular turn as containing a social act, they were automatically prompted with a list of possible acts to choose from, and were then provided with the next possible annotation labels based on the category chosen in the previous step. For example, if an annotator chose to mark a turn as containing positive alignment, they would then be prompted to categorize that positive alignment move as an instance of either “explicit agreement,” “praise/thanking,” “positive reference to a previous speaker’s point,” or “other.” Building this sequence of options into the tool simplified the annotation task and eliminated some sources of accidental error, such as an annotator forgetting to indicate the type of positive alignment after marking its presence.

4.6.3 Provide the annotators with specific criteria for deciding whether a turn should be annotated, and include both positive and negative examples

Although the goal of our annotation was to code for social acts and not for particular linguistic structures, we were able to simplify the annotation task by restricting the annotation of some categories to turns containing particular linguistic cues or criteria. For example, to narrow the range of what might be coded as an “experiential” authority claim, annotators were instructed to look for first person pronouns such as “I,” “me,” “my,” or “we.” When coding “external” authority claims, annotators were instructed that the turn must name a specific source as the basis for that claim (a book, website, article, etc.). Although these decisions may have reduced the overall quantity of data coded with these categories, the addition of these criteria allowed annotators to apply these labels more consistently and with less deliberation (further decreasing cognitive load, as discussed above). Based on sources of confusion identified at regular annotator meetings, we expanded our guidelines to include not only positive examples (MacQueen et al. 1998, “inclusion criteria”) but negative examples as well (MacQueen et al. 1998, “exclusion criteria”). These negative examples enabled annotators to more readily reject a turn that was not appropriate for annotation and to more easily distinguish between our annotation categories.

4.6.4 Perform regular spot checks on the annotation to identify both low-level and high-level disagreements between annotators

In early phases of guideline development, the discussions at our annotator meetings centered on feedback that we provided to annotators on small segments of the data that they had annotated. Questions raised by the annotators at these meetings also allowed us to discuss how to interpret the guidelines in unusual cases. This process worked well for identifying major points of ambiguity in our guidelines, but was limited by the fact that we could only discuss a small portion of the data that had been annotated. We realized that some consistent disagreements between annotators might also go unnoticed by the annotators themselves. As the set of annotated data grew, we were able to examine overall patterns in the use of annotation categories between annotators. This higher-level examination of annotation disagreements worked well for uncovering inconsistencies in annotation due to, for example, over- or under-use of a particular annotation category by one annotator. These patterns of disagreement could then become the topic of discussion at future annotation meetings.

5 Analysis

In this section, we present quantitative analysis of social act prevalence and usage among discussion participants with different roles (such as Wikipedia administrators, or accountants in the chat scenario). In the Wikipedia discussion corpus, we also analyze differences in authority self-presentation among participants with different levels of experience within the editor community, as measured by v-index, months editing, and total edits. While these analyses are meant to be illustrative rather than definitive, they suggest intriguing associations between the self-presentation of authority and interpersonal alignment, and between these social acts and other

features of discussions and discussants such as social role and experience level. The analyses below are based on a subset of the full AAWD corpus released by the project. Specifically, using the version of the files provided in the "merged" directory, we used only files which had been annotated by two annotators (working independently) and counted as authority claims/alignment moves only those that were identified as such by both annotators (where the annotators furthermore agreed on the type).

5.1 Differences in expression of alignment across the Wikipedia samples for the three languages

Speakers of different languages express agreement and disagreement in different ways (see for example Mori 1999). During the process of adapting our English language annotation guidelines to Mandarin and Russian, we identified several potential differences in the expression of alignment, which we addressed by altering our guidelines. During weekly meetings with our annotators, who were all bilingual in English and their native language, we asked them to reflect on other potential differences between the expression of alignment in those two languages. Based on their observations, as well as the observation of the researchers who led the annotation process in Mandarin and Russian, we formed several preliminary hypotheses regarding the ways in which Mandarin and Russian Wikipedia editors differ from English-speaking editors in some aspects in their language uses. For example, Hypothesis (ii) below is based on recurring observations by annotators that Mandarin Wikipedia editors tended to not use any explicit negators, but instead used paired conjunctions to indicate the negation. These preliminary hypotheses and findings are not definitive and are provided to illustrate opportunities for further research.

Hypotheses:

- (i) Negative alignment is more likely to be implicit in Mandarin than in English or in Russian.
- (ii) Mandarin speakers are more likely to use paired conjunctions and partial agreement to express negative alignment.
- (iii) Mandarin and Russian speakers are less likely than English speakers to use the names of their addressees in alignment moves.

(i) Negative alignment is more likely to be implicit in Mandarin than in English or in Russian

This hypothesis is supported by our data. When Mandarin speakers contradict someone else, they are less likely to use explicit disagreement words; instead, they tend to indirectly explain what they believe to be the fact without saying “no/not” or “disagree.” In a sample of 100 Mandarin negative alignment moves and 100 English negative alignment moves, we found that 75 English negative alignment moves contained either “no” or “not,” among which 14 negative alignment moves started with “no.” Also, in English, 7 out of 100 negative alignment moves contained “disagree.” In Mandarin, however, only 20 out of 100 negative alignment moves contained “不是 no/not,” among which 5 moves started with “不是 no/not.” No negative alignment moves contained “不同意 disagree.”

These results support the hypothesis that Mandarin speakers in this dataset use less explicit negative utterances when they are disagreeing with others. In other words, Mandarin speakers negate their statements more indirectly.

The same hypothesis was tested in Russian, and it was found that in our data the usage of explicit agreement markers in Russian is much higher than for Mandarin, but lower than for English. 258 negative alignment examples in Russian were analyzed. Of those examples, 153 moves (or 59.3%) were made using explicit disagreement markers, such as “не/no,” “ни/no” (treated by annotators as a marker of dismissal), “нет/no,” and “я не согласен/I do not agree.”

(ii) Mandarin speakers are more likely to use paired conjunctions and partial agreement to express negative alignment

Chinese frequently employs paired subordinate conjunctions, where the subordinate conjunction introduces the subordinate clause and another discourse connective introduces the main clause. These connectives are generally clause-initial or clause-medial (Xue 2005). Examples of paired conjunctions in Mandarin include: “虽然...但是... Although...but...,” “即使...也... Even though...still,” “不是...而是.. not...but,” “不但...而且 not only...but also...,” “不是...也不是...neither...nor,” “是...但... True it is..., but...” Below are two examples which were tagged as negative alignment in the sample files:

Example 5.

(B is annotated as Negative alignment.)

A: 你 用 “简体夷字”， 那便 是
You use “simplified tribal characters”, then ~~(you)~~ are

夷人， 不配 是 汉 人
tribal man, not qualified as ~~a~~-Han people.

‘You use simplified tribal characters, then you are a barbarian, not an authentic Chinese.’

B: 虽然 我 使用 简体字， 但是 我 赞成
Although I use simplified characters, *but* I support

恢复 正体 字。
revive orthodox characters.

‘Although I use the simplified characters, I support the revival of the traditional Chinese characters.’

Example 6.

(B is annotated as Negative alignment):

A: 既然 我们 汉族的 疆域 主要 是靠 战争
Since our Han’s territory mainly depends on wars

与 征服 而来， 我们 有 什么 资格
and conquering come from, we have what qualification

指责 少数民族 夺权 就是
criticize the Minority’s seizing the power as

“非正义”呢？
“injustice”?

‘Since China’s territory mainly comes from wars and conquering, how can we be qualified to condemn that the minority’s power seizure is injustice?’

B: 即使 是 武装 战争， 如果 是 为 取得 更多的
Even if (≠) is armed war, if is for get more

国土
 land,

仍然 是 侵略 战争， 也 很难说 就是 正义的。
 still is invading war, *still* very hard to say as justice.

‘Even if it is an armed war, if it is aimed for getting more land, it is still invading, and it is hard to say it is justice.’

We hypothesized that when making negative alignment moves, Mandarin speakers would tend to use more paired conjunctions and partial agreement than English speakers in our data set. Partial agreement means that agreement is made with reservation especially when there is doubt or feeling of not being able to accept something completely. A commonly used partial agreement structure in Mandarin is “adj is adj, but...,” meaning the speaker admits one aspect of the good quality, but denies another, with the focus usually on the latter. While the differences are not as stark as for Hypothesis (i), the data do provide support for this hypothesis as well. In a sample of 100 Mandarin negative alignment moves, there are a total of 9 moves (9%) that contain partial agreement and 19 moves (19%) that contain paired conjunctions. On the other hand, in a sample of 100 English negative alignment moves, there are 5 moves (5%) that contain partial agreement and 4 moves (4%) containing paired conjunction (specifically, sentences with the structure “ ... not only ... but (also) ...”).

(iii) Mandarin and Russian speakers are less likely than English speakers to use the names of their addressees in alignment moves

We hypothesized that Mandarin speakers would tend to use the names of their addressees in alignment moves less frequently than English speakers do. The statistics show, however, that English and Mandarin Wikipedia editors are very close with each other in terms of using names when making alignment moves. In a sample of 170 Mandarin alignment moves, there are 26 alignment moves that contain direct names of their addressees, which accounts for 15.29%. On the other hand, in a sample of 200 English alignment, there are 26 alignment moves that contain direct names of their addressees, which accounts for 13%. Pronouns are excluded in the calculation.

Russian speakers, however, included the names of their addressees less often when aligning. Out of 258 negative alignment moves in Russian, only 4 were made with a specific addressee name, which constitutes just 1.55%. Compared to 13% in English or 15.29% in Mandarin, this number suggests that Russian speakers tend to omit specific addressee names when disagreeing with another Wikipedia participant. Moreover, the same pattern was found in positive alignment in Russian. Out of 79 positive alignment moves, only 1 was made using a specific addressee name, which constitutes 1.26% of all positive alignment moves. This data suggests that in both types of alignment, negative and positive, Russian speakers usually do not include the name of the person they are agreeing or disagreeing with.

Alignments made without a specific addressee name also had some practical implications for the annotation process across all languages, i.e., it was difficult for the annotators to identify the target of (dis)agreement (the person with whom the speaker was (dis)agreeing) in the absence of an addressee name.

Initial Turn	Alignment in Next 10 Turns
No Authority Claim	0.52
Any Authority Claim	0.63

Table 10. Average number of alignment moves targeted at participant in 10 following turns.

5.1.5 Summary

While we were able to define social acts in a way that allowed us to annotate them across languages and thus explore differences in both their distribution and realization across those languages, these differences should not be taken as direct evidence for specific differences between national cultures. Wikipedia draws editors from across the globe, and the languages we are working with are spoken in communities in many different nations, both as a native language and as a second language. More generally, it would be overly simplistic to generalize from the differences in expression of social acts that we found to differences between cultures. Rather, the differences in social acts merely suggest that there is room to explore what cultural conventions might help shape the ways in which these social acts are carried out.

5.2 Comparison of Social Act expression in English Wikipedia and IRC

Two researchers qualitatively examined the English language AACD dataset, taking notes on potential differences between the AACD and AAWD English datasets as well as other salient medium variables such as turn length and conversation structure. The researchers then met to share notes and discuss their findings, and identify possible high-level trends. See **Table 5** (Section 3.2 above) for a breakdown of authority claims and alignment moves in the English language IRC data.⁸ Below we share several observations that illustrate the ways in which genre may interact with the presence and form of social acts between the chat and Wikipedia data, and which could be productively investigated in future work. These findings also illustrate some of the challenges posed by differences in conversation structure on the application of coding schemes and annotation processes developed on one genre to another.

5.2.1 Alignment tends to be explicit in IRC data

Unsurprisingly given the limited overall timeframe and synchronous nature of the chat genre, IRC chat participants tended to take shorter turns and to express alignment with other participants more explicitly and with less extensive argumentation than Wikipedia editors.

5.2.2 Alignment moves are common in IRC data

However, overall shorter turns, more-rapid turn-taking, and more explicit agreement/disagreement in IRC can make it difficult to distinguish ‘true’ alignment from backchannels (such as “yeah”). Rapid turn-taking and a lack of nested turn threading also makes it difficult to identify alignment targets.

5.2.3 Negative alignment is less prevalent in IRC data than in Wikipedia data

The relative infrequency of negative alignment moves in IRC may reflect “facework” considerations, as the participants in our study often knew each other offline, and even when they did not, the fact that participants met face-to-face before and after the chat session may have made them less willing to be seen as disagreeable. In addition, the artificial nature of the task scenario may have made them less invested in a particular outcome and therefore less likely to dispute others’ proposals.

⁸ Comparative analysis of alignment and authority claims in the Russian and Mandarin data is beyond the scope of this paper, but represents an intriguing opportunity for future research.

Initial Turn	Positive	Negative	Overall
External Authority Claim	0.26	0.49	0.74
Forum Authority Claim	0.22	0.2	0.42

Table 11. Average number of positive, negative and overall alignment moves targeted at claim-making participant in 10 following turns.

5.2.4 Authority claims are comparatively rarer in IRC data

Authority claims are rarer on IRC than on Wikipedia, but are also harder to identify because they're often less formally structured. Claims may be incomplete or partially implied, or spread across multiple turns. For instance "Safeway has grapes for 80 cents a pound" could be an authority claim, or just an observation. It becomes hard to tell without including evidence from the speaker's previous and subsequent turns.

5.2.5 Authority claim distribution is different in IRC data and Wikipedia data

The distribution of authority claim types differs between Wikipedia and IRC discussions, in a way that likely reflects the medium (less time to craft a complex empirical or logical argument) and the genre (a short term collaboration among immediate peers, with less at stake) and the artificiality of the scenario (assigned roles, made-up task). Our English IRC data exhibits a higher proportion of experiential claims, and social expectations claims (for instance, when a participants assert which kind of pizza their fellow students will prefer).

5.3 Interactions between Authority and Alignment on Wikipedia

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.⁹ 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 (Students' t-test, $t=-2.086$, $df=772$, $p=.037$; **Table 10**)

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 (Students' t-test, $t=3.189$, $df=411$, $p=.002$) and more negative alignment moves than did forum claims (Students' t-test, $t=3.839$, $df=415$, $p<.001$). However, external claims did not elicit significantly more positive alignment moves than forum claims (Students' t-test, $t=0.695$, $df=309$, $p=.488$). This is illustrated in **Table 11**.

5.4 Social Acts and Identity Work

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

⁹ The findings presented in sections 5.3 and 5.4 are slightly revised and expanded from Bender (2011).

# Editors	% Forum Claims	% External Claims	% Turns with Claims	User Type
44	47.1	45.1	19.6	Administrator
192	29.1	63.6	22.3	Registered
55	18.3	70.6	19.8	Unregistered
291	29.8	62.5	21.6	All Types

Table 12. Percent of turns by different types of Wikipedia editors that contain authority claims.

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.

Several possible metrics could be used to represent an editors' level of integration into Wikipedia. In order to test our hypothesis that more integrated editors would make authority claims with greater frequency, we analyzed our data using several of these measures. First, we analyzed whether editors with different official roles (anonymous editors, registered editors, and administrators) exhibited different degrees of claim-making. We then analyzed whether total edits by an editor (under a particular user name), the number of months an editor had been active on Wikipedia, or our v-index measure showed a positive correlation with claim-making. We present our rationale for developing the v-index measure, our sampling criteria, and the findings from our evaluation of v-index, total edits and months active in Section 5.4.2 below.

5.4.1 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.

This is borne out in our data. 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, $n=3164$, $df=2$, $\chi^2=2.367$, $p=.306$) users of different statuses did use significantly different proportions of forum claims and external claims (chi square test for independence, $n=973$ turns, $df=8$), which were the most frequent claim types in the sample overall. **Table 12** presents a breakdown of the claim-making behavior of different user types.

5.4.2 Authority claims by Level of Experience

Evaluating the level of experience of a particular Wikipedia editor presents numerous challenges. Although registered editors (who are identified by their username) are often more experienced than unregistered editors (who are identified by their IP address), this is not a given: a veteran Wikipedian who has thousands of edits' worth of editing experience may edit a page while not logged in. And it is possible for an editor to work for years on Wikipedia without ever creating a username. Status as an administrator is generally thought to be an unambiguous signal of editing experience, since administrators are 'elected' by the community in recognition of extensive work. However, administrators make up only a small fraction of all registered Wikipedia editors, and many non-administrators have comparable levels of experience, but never apply for adminship. Other measures based on community recognition of quality work, such as 'Barnstars' (Kriplean et

Total Edits (log)	Total Turns	Turns with Claims	% Claim Turns
1	78	14	18%
10	1178	263	22%
100	1419	288	20%
1000	234	54	23%
10000	9	0	0%

Table 13. Proportions of claim-bearing turns by participants with different edit counts, log scale.

al. 2008) may serve as robust signals of experience, reputation or investment, but are variably distributed and can also be hard to interpret.

Two measures that are commonly cited by Wikipedia community members as indicators of experience or status are the number of months (or years) an editor has been active, and the total number of edits (to articles, talk pages, policy pages, etc.) an editor has made. However, these measures may not accurately reflect an editor’s level of participation. For instance, is an editor of five years and 20,000 edits who has not made an edit since 2008 as invested as an editor who joined in 2011, but has since made 3,000 edits? However, length of ‘tenure’ is still important: it takes time to integrate into the community and become a “Wikipedian,” due to both the technical complexity of the software and the dizzying variety of rules and conventions (Butler et al. 2008).

The v-index score is designed to account for an editor’s level of investment within the community at a specific point in time by taking into account recent edits over recent months. The longer an editor has demonstrated a high level of sustained participation, the higher their v-index will be. If they are generally a low-level participant their v-index be lower, and if they were a highly active in the past, but their recent participation shows a decrease or has become erratic, their v-index will drop. We therefore hypothesize that v-index will show a stronger correlation with claim-making behavior than either the number of months since the editor joined Wikipedia or the total number of edits by that editor, because it more accurately reflects an editor’s level of engagement and expertise at the point at which they are making a particular utterance.

To evaluate the claim that v-index is a better measure of engagement than simple edits or time counts, we replicated the v-index finding from Bender (2011), and then calculated claim-frequency by total months active and total edits for comparison. We assigned a v-index, months-editing and total-turns value to every turn in our English dataset made by a registered editor or an administrator.¹⁰ Then we sorted these turns into buckets with one bucket for each v-index and month editing, and one bucket per 100 edits. This resulted in 40 v-index buckets, with a top value of 46; 53 months-editing (as 28-day periods) buckets with values with a top value of 58, and 223 total-edits buckets, with a top value of 1580, indicating that one editor in our sample had made over 150,000 edits.¹¹

For each of the three measures, the number of turns per bucket value declined rapidly and steeply, although at different rates. In order to assure an adequate sample size for each bucket, and to avoid a single editor’s claim-making behavior disproportionately influencing the correlation among the higher bucket values (which were represented by far fewer turns), we set a threshold for each dataset at the *last* bucket (ascending) where total turns was greater than 50 *and* total unique editors represented by that bucket was greater than 20. Our data are presented in **Table 13**.

We performed a one-sided Spearman’s rho rank correlation on each of our three metrics against the percentage of turns that contained any type of authority claim. V-index showed a strong, significant positive correlation with claim-making, confirming our hypothesis (one-sided,

¹⁰ See **Appendix A** for v-index sampling considerations and potential sources of error.

¹¹ As of February 29th, 2012 the top lifetime edit count for a single editor was 968,000.

correlation coefficient = .596, $n=14$, $p=0.012$). Total months editing did not show a significant correlation (correlation coefficient = .253, $n=14$, $p=0.192$). Total edits showed a marginally significant correlation (correlation coefficient = .533, $n=9$, $p=0.07$).

Because the results from total edits seemed suggestive, we calculated these values on a log scale as well, with buckets for editors with 1-10 edits, 11-100 edits, 101-1000 edits and 1001-10,000 edits. While this sample is too small to yield a significant correlation, the results (**Table 13**) do not show a clear increase in the number of claim-bearing turns for editors with higher edit counts. We present these results merely to illustrate that for these data the way samples are grouped can affect trends observed. We also caution that choosing a higher or lower “cutoff” value for these sample buckets, or using an alternate statistic, may affect the significance of the resulting correlation.

While additional studies, ideally correlating v-index to other kinds of editor behavior, would be required to establish v-index as a reliable measure of editor engagement, we find these initial results promising.

6 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 participants’ “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. Though this data set is small compared to many that are used in machine learning, it has already been used in research on the automatic detection of forum claims. (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.¹² We hope to see similar approaches applied to the automatic detection of other types of authority claims and of alignment moves in future.

We have also described and reflected on the iterative process through we developed our annotation guidelines. The original drafts of the guidelines were developed on the basis of an initial pass through sample data paired with theoretical reflections, and attempted to map out the space of possible variations within the social act types were annotating. These guidelines were then used by other annotators to annotate more data. Measuring inter-annotator agreement and examining specific cases of disagreement led us to tighten up the guidelines. In general, our strategy was to make the guidelines more restrictive rather than to cast a wider net, based on our observation that “core” or “prototypical” examples of our social acts were easier for the annotators to recognize and agree on. This was driven in part by the fact that our field uses inter-annotator agreement as a measure of consistency of annotations and consistency of annotations in turn as a proxy for the degree to which annotations represent ground truth. Our experience annotating social acts brings into focus and problematizes this proxy relationship: by tightening our guidelines in order to achieve better consistency, it could be argued that we increased the number of false negatives (unlabeled social acts) in our annotated corpus. On the other hand, as with many annotation projects, our labels did not have well-established *a priori* definitions. Thus

¹² An anonymous reviewer notes that the AAWD and AACD corpora are not large enough for certain approaches to machine learning. One of our goals in publishing our code books along with the data we annotated was to enable future work extending the annotations to larger collections of text.

one product of this annotation effort has been the creation of the social act typologies themselves, which, of necessity, focused on overt, explicit variants within a larger space. While similar observations likely pertain to many types of annotation efforts, we believe they are particularly prominent in this domain because we are working at such remove from the linguistic structure of the utterances we are annotating.

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, we expect that the distribution and presentation of our social acts be manifestly different in online genres other than Wikipedia discussions and Internet Relay Chat. Wikipedia discussions are shaped by a set of well-defined, local communication norms that are closely tied to the task of distributed, collaborative writing and the culture of open-source software. Internet Relay Chat and related instant message protocols present their own constraints, and furthermore a comparison between our facilitated discussions and ‘organic’ IRC data would help shed light on the ways our protocol and testing environment may have shaped the discourse.

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 hope that the online communication genres captured in the AAWD and AACD corpora prove to be valuable resources for social scientific analyses of communication behaviors as well as a resource for the development of NLP systems which can automatically identify these social acts, on 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.

We wish to thank our colleagues: Brian Hutchinson, Alex Marin, Mari Ostendorf, and Bin Zhang. We also gratefully acknowledge the contribution of the annotators: Wendy Kempsell, Kelley Kilanski, Robert Sykes and Lisa Tittle. We also wish to thank the anonymous reviewers for their valuable feedback on this manuscript.

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.

Appendix A: Sources of Inconsistency in Computing v-Index

The basic algorithm for calculating the v-index values and the data included in the AAWD corpus are the same as that used in Bender et al. (2011). However, the calculation of v-index requires access to the rest of Wikipedia as well. The v-index values reported in Bender et al. (2011) were calculated against the 2008 Wikipedia snapshot. In this paper, we instead use a live mirror of the Wikipedia database. As v-index values only reflect revisions made before the turns in question, this should in principle lead to the same results. However, the live database differs from the snapshot in that revisions that have been permanently deleted from Wikipedia after the snapshot was taken are no longer reflected in the database. Because of Wikipedia’s built-in version control system, any content ever entered into Wikipedia is perpetually available and potentially viewable by default. Employees of the Wikimedia Foundation (but not individual editors) may therefore occasionally completely delete individual revisions of a page if making it available may have legal ramifications (for instance contain potentially libelous content), or revisions which contain particularly sensitive information (such as a users social security

number).¹³ While this is a comparatively rare occurrence, one of the discussions we annotated (comprising 58 turns) was associated with such a deleted page.

We were able to calculate v-index values for the turns in this deleted discussion by looking up the timestamp of those turns which were available to us because one of the authors (a contractor with Wikimedia) possessed special data access privileges. However, there may still be slight differences as the deletion of other pages not captured in our database may have lowered the v-index value for some proportion of our turns.

In addition, in the 2011 work, we calculated v-index values for unregistered users by treating all edits from the same IP address as belonging to the same user. The mapping from IP addresses to users is not reliable, however, and accordingly we have chosen to exclude turns by unregistered users in the current analysis. This removes 443 turns from consideration.

We did not evaluate potential correlations between v-index, months editing and total edits and authority claims for Russian and Mandarin because of a significant reduction in sample size.

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 (EMNLP '10)*, Association for Computational Linguistics, pages 1024-1034, Stroudsburg, Pennsylvania.
- Mats Alvesson and Hugh Willmott. 2002. Identity regulation as organizational control: Producing the appropriate individual. *Journal of Management Studies*, 39(5):619-644.
- Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555-596.
- Collin F. Baker, Charles J. Fillmore and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Association for Computational Linguistics, pages 86-90, Stroudsburg, Pennsylvania.
- 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.
- Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages. In *Proceedings of the ACL-HLT Workshop on Language in Social Media (LSM 2011)*, pages 48-57, Portland, Oregon.
- Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge.
- Mary Bucholtz and Kira Hall. 2010. Locating identity in language. In Carmen Llamas and Dominic Watt, editors, *Language 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*, Association of Computing Machinery, pages 27-36, San Diego, California.
- Brian Butler, Elisabeth Joyce, and Jacqueline Pike. 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 Twenty-sixth Annual SIGCHI Conference on Human Factors in Computing Systems (CHI'08)*, Association of Computing Machinery, pages 1101-1110, New York, New York.

¹³ See http://en.wikipedia.org/wiki/Wikipedia:Office_actions

- Robert E. Cummings. 2008. What was a Wiki, and why do I care? A short and usable history of Wikis. *Wiki Writing: Collaborative Learning in the College Classroom*, pages 1-16. University of Michigan Press, Ann Arbor.
- Jolene Galegher, Lee Sproull, and Sara Kiesler. 1998. Legitimacy, authority, and community in electronic support groups. *Written Communication*, 15(4):493–530.
- Meghan Lammie Glenn, Stephanie M. Strassel, and Haejoong Lee. 2009. XTrans: A speech annotation and transcription tool. In *Proceedings of Interspeech 2009*, pages 2855–2858, Brighton, UK.
- Erving Goffman. 1959. *The Presentation of Self in Everyday Life*. Doubleday, Garden City, New York.
- Erving Goffman. 1981. *Forms of Talk*. University of Pennsylvania Press, Philadelphia.
- Jakob L. Jensen. 2003. Public spheres on the internet: Anarchic or government sponsored; a comparison. *Scandinavian Political Studies*, 26(4):349–374.
- Travis Kriplean, Ivan Beschastnikh, David W. McDonald, and Scott A. Golder. 2007. Community, consensus, coercion, control: Cs*w or how policy mediates mass participation. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work (GROUP '07)*, Association of Computing Machinery, pages 167-176, New York, New York.
- Travis Kriplean, Ivan Beschastnikh, and David W. McDonald. 2008. Articulations of wikiwork: Uncovering valued work in Wikipedia through barnstars. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08)*, Association of Computing Machinery, pages 47-56, New York, New York.
- J. Richard Landis and Gary G. Koch. 1977. Measurement of observer agreement for categorical data. *Biometrics*, 33(1):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.
- Kathleen M. MacQueen, Eleanor McLellan, Kelly Kay, and Bobby Milstein. 1998. Codebook development for team-based qualitative analysis. *Cultural Anthropology Methods*, 10(2):31-36.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Detecting forum authority claims in online discussions. In *Proceedings of the ACL-HLT Workshop on Language in Social Media (LSM 2011)*, pages 39-47, Portland, Oregon.
- Junko Mori. 1999. *Negotiating Agreement and Disagreement in Japanese: Connective Expressions and Turn Construction*. John Benjamins Publishing Company, Amsterdam.
- 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*, Association for Computing Machinery, article 26, 2 pages, New York, New York.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71-105.
- Martin J. Pickering and Simon Garrod. 2004. The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, 27(2):212-225.
- Janie Rees-Miller. 2000. Power, severity, and context in disagreement. *Journal of Pragmatics*, 32(8):1087-1111.
- Kay Richardson. 2003. Health risks on the internet: Establishing credibility online. *Health, Risk and Society*, 5(2):171-184.
- John R. Searle. 1975. Indirect Speech Acts. In Peter Cole and Jerry L. Moran, editors, *Syntax and Semantics Volume 3: Speech Acts*, pages 59-82. Academic Press, New York.
- 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.
- Clay Shirky. 2008. *Here Comes Everybody: The Power of Organizing Without Organizations*. Penguin Press, New York.
- Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, Association for Computational Linguistics, pages 97-100, Cambridge, Massachusetts.
- Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Association for Computational Linguistics, pages 41-43, Barcelona, Spain.
- Jan Svennevig. 1999. *Getting Acquainted in Conversation: A Study of Initial Interactions*. John Benjamins Publishing Company, Amsterdam.
- György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The BioScope corpus: Annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, Association for Computational Linguistics, pages 38-45, Stroudsburg, Pennsylvania.
- Dorothea K. Thompson. 1993. Arguing for experimental “facts” in science. *Written Communication*, 10(1):106.
- Fernanda B. Viegas, Martin Wattenberg, Jesse Kriss, and Frank van Ham. 2007. Talk before you type: Coordination in Wikipedia. In *Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS '07)*, page 78, IEEE Computer Society, Washington, D.C.
- Linda Wine. 2008. Towards a deeper understanding of framing, footing, and alignment. *Columbia University Working Papers in TESOL and Applied Linguistics*, Teachers College, 8(3):1-3.
- Nianwen Xue. 2005. Annotating discourse connectives in the Chinese Treebank. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 84-91, Ann Arbor, Michigan.