Stylistic Accommodation on Reddit

Caitrin Armstrong
260501112

Sunyam Bagga
260777459

Abstract

We implement measures of linguistic entrainment and apply them to a large, diverse dataset of Reddit comments.

1 Introduction

The tendency to align features of speech to that of a conversational partner's is known as *accommodation, entrainment, convergence, or alignment*. A well-studied phenomenon, it has been shown to occur in many contexts across many different features (Levitan et al., 2015). Recently, research on entrainment has progressed in two directions. First, the use of computational models has allowed for large-scale investigation. Second, investigation has moved onto the online space, where users are lacking the usual nonverbal modalities of face-to-face interaction.

In (Danescu-Niculescu-Mizil et al., 2011), the authors challenged the theory of linguistic entrainment on a novel dataset: Twitter conversations. Using 14 linguistic features generated by LIWC (Pennebaker et al., 2015), their multi-stage investigation found that cohesion (the idea that tweets occurring in the same conversation are more close linguistically than those that don't) occurs for all linguistic features, and this could not be explained by homophily (the idea that tweets about the same topic will naturally be stylistically similar). By developing a measure of stylistic accommodation, they showed that a Twitter user is more likely to use a linguistic feature if the tweet they are replying to exhibited the same feature.

In this paper, we follow the methods of (Danescu-Niculescu-Mizil et al., 2011) to investigate the presence of entrainment on a huge conversational dataset derived from Reddit.com. It is interesting to test this theory on Reddit since it lacks the traditional social-media functionalities (users cannot create friends, and almost always post under pseudonyms). To a greater extent than Twitter we observe context collapse - the idea that users are interacting with those they have, and never will, meet face-to-face (Michael and Otterbacher, 2014). Reddit is divided into smaller communities, each one centered around a particular topic ranging from broad to very niche. Communities are self-governed and vary in terms of how cohesive they are (how much users feel committed to the group as a collection of individuals), and it has been shown that cohesiveness and entrainment are correlated (Gonzales et al., 2010). Reddit comments are, therefore, a rich dataset for testing this theory. With this in mind, we test:

1. *Hypothesis 1*: Using the same linguistic features and model as in (Danescu-Niculescu-Mizil et al., 2011), we will observe the presence of both stylistic cohesion and accommodation on Reddit.

2. *Hypothesis 2*: The overall amount of stylistic accommodation displayed in a subreddit will correlate with the amount of group cohesion as measured through responses to a multicomponent measure of group cohesion.

3. *Hypothesis 3*: We can improve on the measure of stylistic accommodation by filtering out short comments.

4. *Hypothesis 4*: We can improve on the measure of stylistic accommodation proposed in (Danescu-Niculescu-Mizil et al., 2011) by integrating the frequency of occurrence of the linguistic features.

2 Related Work

The linguistic features chosen to model entrainment in social media have varied wildly across studies, from parts of speech (Tran and Ostendorf, 2016) to emoticons (Michael and Otterbacher, 2014). Others simply measured the distance between internal and external word and word-ending usages (Tamburrini et al., 2015). We chose to focus only on (subconscious) stylistic features since we did not want to capture topic or conscious mimicry(Michael and Otterbacher, 2014).

We follow Gao et al. by modifying the measure of accommodation described in (Danescu-Niculescu-Mizil et al., 2011) to include the length of a post and the frequency of a linguistic feature occurrence. We do not implement the normalization measure described in (Jones et al., 2014) since that was not necessary for testing our hypotheses.

Welbers and de Nooy (Welbers and de Nooy, 2014) show that stylistic accommodation occurs on a special-interest forum, especially with those stylistic features that are related to a shared identity, supporting their thesis that accommodation is a part of the group bonding process. They do not, however, offer further qualitative evidence to corroborate this. We directly investigate this claim by comparing our subreddit-level group accommodation to a multi-component measure of group cohesion. As far as we are aware, this is the first time this correlation has been investigated in a social media setting, although (Gonzales et al., 2010) investigated this with a different measure in a controlled CMC experiment.

In apparently the only related study conducted on Reddit data, Tran and Ostendorf (Tran and Ostendorf, 2016) investigated the relative utility of language style versus topic models for subreddit identification. They found not only that style is a better indicator of subreddit identity, but that there is a positive correlation between the style similarity of a post to the community and the community's reception of that post. While this provides us with good evidence for the importance of style on Reddit, they worked only with parts of speech as a model of style, and did not explicitly investigate the presence of language accommodation on Reddit.

Finally, (Michael and Otterbacher, 2014) implemented a model of language herding similar to (Danescu-Niculescu-Mizil et al., 2011) but with higher-level features, finding that language herding does occur on a travel forum, with users adopting their language that of the previous posters'. This supports the idea that this phenomenon can occur on websites without support for social-media functions, and even on a website without direct inter-personal interaction.

3 Method

3.1 Dataset

Our dataset consists of Reddit comments made to posts in the following 18 subreddits: monarchism, DebateCommunism, socialism, SocialDemocracy, LibertarianSocialism, conservatives, GreenParty, PirateParty, democrats, Objectivism, moderatepolitics, christianancaps, futuristparty, DebateaCommunist, LibertarianDebates, paleoconservative, BullMooseParty, Liberal, between the dates of January 1 2015 and December 30th 2017. Sub-sampling is described in each experiment. Empty comments, comments made by deleted users, and self-replies were removed. In addition, all non-alphabetic characters were removed from the text.

3.2 Hypothesis 1: Occurrence

For defining cohesion and accommodation, we adopt the probabilistic framework defined in (Danescu-Niculescu-Mizil et al., 2011).

3.2.1 Stylistic Cohesion

For a given dimension C, they define cohesion as:

$$Coh(C) = P(T^C \wedge R^C | T \leftrightarrow R) - P(T^C \wedge R^C) \quad (1)$$

where T^C (and R^C) is the event in which a Reddit comment T (and R) exhibits C, and $T \leftrightarrow R$ is the condition that comments T and R form a conversational turn (R is a reply to T). In equation 1, the first term captures the probability that a stylistic dimension is exhibited in comments that are part of a conversation, and the second term is the probability that the same dimension is exhibited in unrelated comments. Hence, $Coh(C) > 0$ demonstrates that cohesion is observable in our dataset.

The first probability in Equation 1 is calculated as:

$$P(T^C \wedge R^C | T \leftrightarrow R) = \frac{|\{(t,r) | t \leftrightarrow r, t^C, r^C\}|}{|\{(t,r) | t \leftrightarrow r\}|} \quad (2)$$

where t^C is the condition that comment t exhibits C. To estimate the second probability, we first construct a set of "fake turns" by randomly pairing together comments within subreddits (regardless of their authors). Then, we can write:

$$P(T^C \wedge R^C) = \frac{|\{(t,r) | t \neq r, t^C, r^C\}|}{|\{(t,r) | t \neq r\}|} \quad (3)$$

where t^C is the condition that comment t exhibits C, and $t \neq r$ is the condition that comments t and r are part of a "fake" turn.

3.2.2 Stylistic Accommodation

For a stylistic dimension C and a pair of users (a, b), the accommodation of user b to user a is measured by how much the fact that user a exhibits C in a Reddit comment T_a increases the probability of b to also exhibit C in his reply to T_a:

$$Ac_{(a,b)}(C) = P(T_b^C | T_a^C, T_b \hookrightarrow T_a) - P(T_b^C | T_b \hookrightarrow T_a) \quad (4)$$

where T_a^C (and T_b^C) is the event in which a comment posted by user a (and user b) exhibits C, and $T_b \hookrightarrow T_a$ is the condition that T_b is a reply to T_a. The condition $T_b \hookrightarrow T_a$ models the temporal nature of accommodation: a user can accommodate to her conversational partner only after receiving her input. To calculate global accommodation for a given dimension C, expectation is taken over all possible conversing user pairs (a, b):

$$Ac(C) = E[Ac_{(a,b)}(C)] \quad (5)$$

Hence, $Ac(C) > 0$ demonstrates that accommodation is observable in our dataset.

The second probability in equation 4 is estimated as the fraction of b's replies to a in which b's comment t_b exhibits C:

$$P(T_b^C | T_b \hookrightarrow T_a) = \frac{|\{(t_a, t_b) | t_b \hookrightarrow t_a, t_b^C\}|}{|\{(t_a, t_b) | t_b \hookrightarrow t_a\}|} \quad (6)$$

Similarly, the first probability is:

$$P(T_b^C | T_a^C, T_b \hookrightarrow T_a) = \frac{|\{(t_a, t_b) | t_b \hookrightarrow t_a, t_b^C, t_a^C\}|}{|\{(t_a, t_b) | t_b \hookrightarrow t_a, t_a^C\}|} \quad (7)$$

To calculate each of these values, we took a random sub-sample of 1000 comment-pairs from each subreddit. We only included user pairs where the users had conversed at least 5 times. Subreddits BullMooseParty, christianancaps, LibertarianSocialism, PirateParty, paleoconservative, and futuristparty had less than 1000 but greater than 100 eligible sampled conversations over the specified daterange. They were included despite small values because of their utility in providing insights into small subreddits.

We used the LIWC 2015 software to analyse the language text in Reddit comments. The following linguistic features were considered: Analytic, Clout, Authentic, Tone, all Pronouns, Articles (article), Prepositions (prep), Conjunctions (conj), Negations (negate), Quantifiers (quant), Discrepancies (discrep), Tentative (tentat), Certainty (certain), and Differentiation (differ) (Pennebaker et al., 2015).

3.3 Hypothesis 2: Correlation with Group Cohesion Measure

The author CA was in the process of conducting a survey of members of the subreddits included in this investigation. This survey includes a well-known measure created by Leach et al. to capture the degree of cohesion felt by members of a group (Leach et al., 2008). The multi-component measure results in a single value, or an overall group cohesion score. The number of responses per subreddit range from $7 - 172$, but the response value is not significantly linearly correlated with the number of responses (Pearson's $R = 0.1151$).

3.4 Hypothesis 3: Length Restriction

Our implementation is the same as in Hypothesis 1 (Section 3.2), but here we only included conversations where each comment had more than 10 words.

3.5 Hypothesis 4: Integrating Frequency of Feature Occurrence

We modify the measure of accommodation implemented to test Hypothesis 1 by incorporating frequency of the LIWC features. Instead of just counting the number of replies (as in Equations 6 and 7), we used the percentage of total words that match the dictionary category for that linguistic feature. Equation 6 is straightforward, but to ensure that Equation 7 stays a probability distribution, we took the minimum count of t_b^C and t_a^C.

4 Results

4.1 Hypothesis 1: Occurrence

Cohesion was present across all subreddit and linguistic feature pairs except: LibertarianSocialism & "pronoun", Objectivism & "pronoun", BullMooseParty & "you", paleoconservative & ("ppron", "we",

"quant", "certain"). Figure 1 shows the estimates of the two probabilities for the subreddit 'Conservatives' for each of the 17 linguistic features considered.

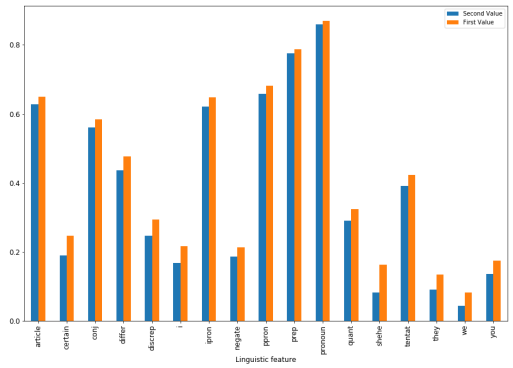

Figure 1: The effect of cohesion observed in the subreddit 'Conservatives'

Now that we have established that cohesion is exhibited in our dataset, we can address accommodation. The presence of accommodation was confirmed for the linguistic features and subreddits described in Table 1 (Two tailed Paired t-test where $p < 0.01$). Due to limited space, we limit our visualizations to three linguistic features across four subreddits. As can be seen in Figure 2, users in the Liberal subreddit accommodate equally across all three linguistic features. Whereas, users in Objectivism accommodate more for 1st person plural pronoun (we), but less for 'differ' and Conjunction (conj). A similar behavior is seen for the users conversing in the Democrats subreddit. Accommodating the most for 1st person plural pronoun seems to be the general trend across all subreddits.

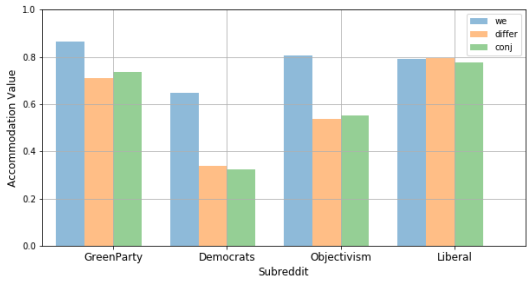

Figure 2: Accommodation across subreddits

Table 1: Presence of Features

Subreddit	Features
christian_ancaps	we, they, shehe, tentat, quant, conj, i, differ, you, discrep, certain, ipron, negate, article, prep, ppron, pronoun
conservatives	they
DebateaCommunist	article, certain, conj, differ, discrep, i, ipron, negate, ppron, prep, pronoun, quant, shehe, tentat, they, we, you
democrats	certain, conj, differ, discrep, i, ipron, negate, quant, shehe, tentat, they, we, you
futuristparty	shehe
GreenParty	article, certain, conj, differ, discrep, i, ipron, negate, ppron, prep, pronoun, quant, shehe, tentat, they, we, you
Liberal	article, certain, conj, differ, discrep, i, ipron, negate, ppron, prep, pronoun, quant, shehe, tentat, they, we, you
LibertarinDebates	article, certain, discrep, i, quant, shehe, tentat, they, we, you
LibertarianSocialism	article, certain, conj, differ, discrep, i, ipron, negate, ppron, prep, pronoun, quant, shehe, tentat, they, we, you
moderatepolitics	certain, conj, differ, discrep, i, negate, ppron, prep, pronoun, quant, shehe, they, we, you
monarchism	shehe, we, you
Objectivism	certain, conj, differ, discrep, i, negate, quant, shehe, tentat, they, we, you
PirateParty	shehe, you
SocialDemocracy	shehe, they, we
socialism	shehe, they, we

While we do not provide a statistical analysis, we note that in (Danescu-Niculescu-Mizil et al., 2011) cohesion and accommodation were found across all dimensions; comparatively cohesion appears to occur less on Reddit than on Twitter.

4.2 Hypothesis 2: Correlation with Group Cohesion Measure

We measured the correlation between the cohesion survey results and the averaged linguistic accommodation value for each subreddit to be Pearson's $R = 0.398$. While this result shows some correlation, it is not very strong overall. The plot of the relationship between the values is shown in Fig. 3.

4.3 Hypothesis 3: Length Restriction

Using a two tailed paired t test across all accommodation results for subreddit and linguistic features, we find that restricting the length does not result in a significant increase in the accommodation values (paired $t(17) = -1.1808$, $p = 0.119072$).

This length threshold also marginally decreases the correlation with the group cohesion survey results (Pearson's $R = 0.393$), indicating that it is not necessarily an improved proxy for group cohesion.

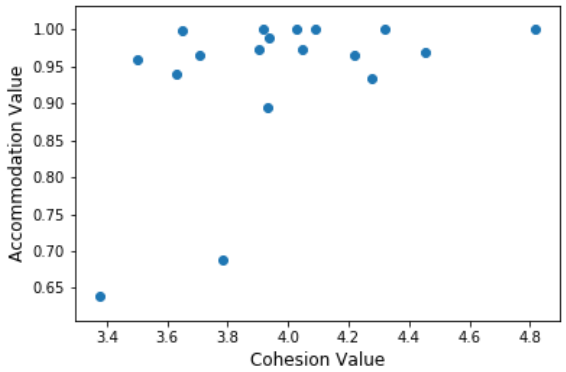

Figure 3: The cohesion score and average accommodation value per subreddit for the base implementation of accommodation

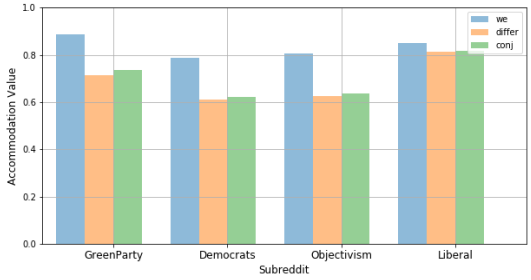

Figure 4: The difference in accommodation observed across our three conditions

The implications of this change are illustrated in Fig. 4 and described in the next section below.

4.4 Hypothesis 4: Integrating Frequency of Feature Occurrence

Using a two tailed paired t test across accommodation results for all subreddits and linguistic features, we find that the new accommodation calculation does result in a significant increase in the accommodation values (t = -17.01042, $p < .00001$.). This is depicted in Figure 4.

The new method also decreases the correlation with the group cohesion survey results (Pearson's $R = 0.33$), indicating that it is not an improved proxy for group cohesion. Overall, our two conditions both increased our observed accommodation values and increased the number of subreddit feature pairs that were significant, but because they do not result in an increased correlation with our group cohesion survey results, we cannot make any speculation about their true utility.

5 Discussion and Conclusion

In this paper, we implemented measures of stylistic cohesion and accommodation as described by (Danescu-Niculescu-Mizil et al., 2011). Our original hypothesis was confirmed, as we found that stylistic cohesion and accommodation occurred on Reddit, to varying degrees per subreddit.

We attempted to further improve upon these measures by implementing two common-sense improvements: requiring a minimum length of comment for analysis, and modifying the accommodation calculations such that the frequency of linguistic feature is considered.

While these two modifications did statistically increase the amount of accommodation and thus also the occurrence of statistically significant subreddit and linguistic feature combinations, they did not result in an increased correlation with the group survey results. It is impossible to say whether these modifications offered any improvement on the original method without having a comparative metric. Given the decreased correlation with the surveyed group cohesion, it seems that they did not improve. However, the correlation was already moderate and non-significant.

Due to the dataset chosen, we were limited in the amount of data available. Future investigations should work with subreddits that have more available data. Correlations for a number of different measures such as subreddit distinctiveness and dynamicity (Zhang et al., 2017), network features and inter-user power relations (Danescu-Niculescu-Mizil et al., 2012)) should also be explored in addition to the group cohesion survey. This will allow us to better explore and compare measures of linguistic accommodation.

6 Statement of Contributions

Sunyam implemented all methods, wrote the methods section of the paper, parts of the results section, edited, and generated all figures.

Caitrin designed the study, wrote and ran experiments to generate results and wrote all other paper sections.

References

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754. ACM.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708. ACM.

Amy L Gonzales, Jeffrey T Hancock, and James W Pennebaker. 2010. Language style matching as a predictor of social dynamics in small groups. *Communication Research*, 37(1):3–19.

Simon Jones, Rachel Cotterill, Nigel Dewdney, Kate Muir, and Adam Joinson. 2014. Finding zelig in text: A measure for normalising linguistic accommodation. Coling.

Colin Wayne Leach, Martijn Van Zomeren, Sven Zebel, Michael LW Vliek, Sjoerd F Pennekamp, Bertjan Doosje, Jaap W Ouwerkerk, and Russell Spears. 2008. Group-level self-definition and self-investment: a hierarchical (multicomponent) model of in-group identification. *Journal of personality and social psychology*, 95(1):144.

Rivka Levitan, Stefan Benus, Agustın Gravano, and Julia Hirschberg. 2015. Entrainment and turn-taking in human-human dialogue. In *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction*.

Loizos Michael and Jahna Otterbacher. 2014. Write like i write: Herding in the language of online reviews. In *ICWSM*.

JW Pennebaker, RJ Booth, RL Boyd, and ME Francis. 2015. Linguistic inquiry and word count: Liwc 2015 [computer software]. pennebaker conglomerates.

Nadine Tamburrini, Marco Cinnirella, Vincent AA Jansen, and John Bryden. 2015. Twitter users change word usage according to conversation-partner social identity. *Social Networks*, 40:84–89.

Trang Tran and Mari Ostendorf. 2016. Characterizing the language of online communities and its relation to community reception. *arXiv preprint arXiv:1609.04779*.

Kasper Welbers and Wouter de Nooy. 2014. Stylistic accommodation on an internet forum as bonding: Do posters adapt to the style of their peers? *American Behavioral Scientist*, 58(10):1361–1375.

Justine Zhang, William L Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. *arXiv preprint arXiv:1705.09665*.