Coordination and Social Network Centrality

Bill Noble
University of Amsterdam
winobes@gmail.com

Abstract

The ubiquitous use of the internet for social interaction provides a rich source of data for various kinds of social research. In this paper, we use techniques in social network analysis to examine trends in linguistic coordination among Wikipedia editors. A metric called linguistic style coordination is used to detect differences in coordination between different groups of editors. Speakers coordinate more towards targets in highly central social positions. The effect of social position is shown to be greater than the effect of whether or not the target is a Wikipedia administrator. Finally, we speculate that network effects on coordination may suggest that coordination plays a role in the development and evolution of linguistic communities, and suggest future research for exploring this connection.

1 Introduction

Perhaps the most distinguishing feature of the internet as a communication platform is its ability to support highly interactive communities among geographically and otherwise disconnected users. In the current age of prolific internet use, it is not uncommon for an individual to participate in many different internet communities with varying degrees of connection to each other and to offline communities. An important part of belonging to any community is understanding and adopting its linguistic norms, be they vernacular, syntactic patterns, or any other linguistic vector along which a community may differentiate itself. Given that an individual may belong to many distinct linguistic communities, conforming to the norms of any given community can be difficult. Making it even more challenging is the fact that, online communities have linguistic conventions that change over time (Danescu-Niculescu-Mizil et al., 2013). This dynamic nature of linguistic norms means that the challenge of conforming applies not just to new members, but also to long-time community member. In this paper, we investigate the effects of social structure on linguistic coordination in one online community and suggest that the goal of conforming to the norms of the community as a whole may explain those effects.

The widespread availability of social network data in the age of ubiquitous internet has contributed to a renaissance in network analysis as a tool for social research, but much of the data resulting from online activity contains interactions recorded in natural language, which is often overlooked in such structural analyses. A growing body of research has focused on using this linguistic data to provide insight into questions about social structures and about natural language itself. This paper attempts to combine these two trends by applying measures of network centrality to explain differences in linguistic coordination in discussions between editors on Wikipedia talkpages.

2 Related Work

The research presented in this paper is an extension of the results in *Echoes of Power* (Danescu-Niculescu-Mizil et al., 2012). There it is shown that in group discussions, the degree to which a speaker echoes the linguistic style of the person to whom they are responding reveals the power differential between the individuals involved. This effect is demonstrated for both static, status-based power and for the "situational power" that arises from dependence of some kind. We employ many of the same methods used in that paper (Section 5.1), but instead of power, we look at the effect of an individual's position in a social network context.

The ease with which social network structure can be extracted from online communities has made network analysis of such communities a very active method of research in many fields concerned with social interaction. That said, the effect of network structure on language communities had been observed long before the prevalence of the internet. There are a number of studies dealing with social structure and linguistic conformity.

Eckert (1988) investigated the effects of the social network of suburban Detroit area adolescents on their susceptibility to urban phonological innovations. She argues that traditional demographic data insufficiently describe the spread of linguistic change through their community, and that linguistic change can better be explained by looking at social network structure. Furthermore, she discusses linguistic style as an important factor in maintaining acceptance in a social group. Eckert claims that the incentive for linguistic alignment is especially strong among adolescents for two reasons: First, there is an increased need for markers of social differentiation in an environments of unstable, rapidly developing social structure. And second, the adolescent social context is insulated from the normalizing influence of the more linguistically stable adult world.

Online communities share a lot of the characteristics that make adolescent society rich with linguistic differentiation; they develop and change rapidly, and enjoy a certain degree of disattachment from the "real world" of face-to-face communication.

Danescu-Niculescu-Mizil, et al. (2013) study changes in linguistic norms change in online communities. New members of online communities adapt and use more of the language specific to the community as time goes on. During an initial learning phase, users are also more likely adopt changes in the linguistic norms of the community as a whole than are long-time users.

Kooti, et al. (2012) also looks at linguistic change in online communities. Specifically, they investigate the diffusion of conventions for source attribution on Twitter. They study find that early adopters of online linguistic conventions tend to be more active and better connected.

3 Theoretical Background

3.1 Linguistic Style Coordination

It has been well established that when people interact, they tend to subconsciously converge in their behavior. Many of the vectors along which this convergence occurs are linguistic. Cognitive psychology explains this convergence by an unconscious mechanism known as *priming*.

In this paper, we are concerned primarily with linguistic alignment that occurs within dialogues. We call this kind of immediate dialogue-specific alignment *coordination*. The imitation of syntactic structure, phonology, rate of speech, and lexical choice are all examples of linguistic coordination. A measure of linguistic coordination will quantify the degree to which dialogue participants converge on their speech behavior. Since coordination is directed, (a speaker *coordinates with* another by echoing some of her speech behavior), measures of coordination may be asymmetrical.

Linguistic style coordination is one such measure. It considers how much a speaker immediately echoes certain markers of linguistic style present in the utterances to which they are replying.

We write $\mathcal{E}_u^m$ to mean that utterance u exhibits a marker m. For two individuals a and b, the linguistic style coordination of b (the *speaker*) towards a (the *target*) for a style marker m is defined as:

$$C^m(b, a) = P(\mathcal{E}_{u_b}^m \mid \mathcal{E}_{u_a}^m) - P(\mathcal{E}_{u_b}^m)$$

Where (u_a, u_b) belongs to the set of pairs utterances made by b in response to a. Note that if none of the utterances of a that b responds to exhibit m, then $C^m(b, a)$ is undefined. This notion is easily extended to measure the coordination of b towards a group of targets A by considering all of the utterances where b is responding to any of the members of A.

Following Danescu-Niculeascu-Mizil, et al. (2012), we use the presence of a word in a particular category of *function words* as linguistic style markers. Function words, as opposed to *content words* are words that have little semantic meaning outside of a sentence in which they appear. Function words have two desirable qualities for our purposes: First, they have little semantic meaning; a speaker who makes an utterance containing words from some category of function words could usually have conveyed the same meaning without using words in that category. Thus, the use of function words from a certain category is a genuinely *stylistic* choice and not dependent on the content of the utterance. Second, function words are processed and used unconsciously (Ireland et al., 2011), and thus subject to the effects of unconscious alignment influences.

3.2 Network Centrality

A social network is a graph model of a community whose nodes are individuals and whose edges represent links between those individuals. These edges are sometimes weighted to capture the strength of certain links or directed to represent asymmetrical relationships. Given a social network, it is natural to want to extract information about the importance of individuals in the community. Which nodes are important depends not only on their place in the network structure, but also on what it means to be important in that community and how these features are encoded in the network model. Network centrality is a family of measures that give each node a numerical value representing importance. Here we consider three measures of network centrality described below.

Eigenvector centrality tries to capture the notion that your importance in a network depends on the importance of your neighbors. Let $M(n)$ be the *neighborhood* of n; that is, the nodes in N that are connected to n. Then the eigenvector centrality of n is defined by

$$T^e(n_0) = \frac{1}{\lambda} \sum_{n \in M(n_0)} T^e(n)$$

where λ is a constant, the *eigenvalue*. There may be multiple values of λ for which the eigenvector centrality is defined, but taking the largest value provides a consistent measure across the network (Wasserman, 1994).

Betweenness centrality measures how much a node contributes to the overall connectivity of the network. Nodes who lie on more *shortest paths* between pairs of other nodes have higher betweenness centrality. Specifically it looks at all of the shortest paths between each pair of nodes, and counts how many of them contain the node in question. (Bolland, 1988).

$$T^b(n_0) = \sum_{n \neq m \in N} \frac{|\{Path(m,n) | n_0 \in Path\}|}{|\{Path(m,n)\}|}$$

3.3 Power, Centrality, and Importance

As discussed in Section 2, power differences may be static (arising from a formal or informal position of authority) or dynamic (arising from a situation where one person has power over something that another needs). In this paper, by *power* we generally mean to restrict attention to the static kind of power.

We also sometimes refer to an individual's *position* in a community. Here we are referring specifically to position in a social network as defined by one of the centrality measures described above. Position might easily be seen as a source of power. In fact, this is often true, but it is important to note that network position does not necessarily equate to power. Take the example of exchange networks: An exchange network is one where social relations involve the exchange of valued commodities, be they physical, like good and services, or less tangible, like affection or information. In such networks, power often increases with access to non-central individuals who have less choice in partners for exchange. In such situations it actually presents a power advantage *not* to be centrally located (Cook et al., 1983).

Importance is a broader notion than that of either power or centrality. When we talk about the importance of an individual in a community, we mean the degree to which that individual shapes the identity, norms, and interests of the community. Clearly importance intimately related to the other two notions; an individual may be important because of her central social position, and importance may either result from or contribute to an individual's power. That said, there might easily be important individuals who don't occupy a particularly central position or possess a lot of power.

4 Hypothesis

This paper will investigate three hypotheses that relate the results in *Echoes of Power* (Danescu-Niculescu-Mizil et al., 2012) to the concept of network centrality.

$H1$: Speakers coordinate more with those in more central social positions.

$H2$: People in more central social positions are more likely to possess power.

$H3$: The effect of $H1$ holds independently of any correlation observed in $H2$.

5 Investigation

5.1 Experimental Method

The Wikipedia talkpages corpus consists of a collection of conversations from Wikipedia editors' talk pages. Each utterance (or *post*) is annotated with metadata that includes the username of the editor who made the post, the username of the editor on whose page the post was made, and information about which previous utterance it is a reply to (if any). There is also metadata on users, including whether or not the user is a Wikipedia administrator (an *admin*) and how many edits she has made. The corpus was collected in August 2011 (Danescu-Niculescu-Mizil et al., 2012).

A network structure was imposed on the corpus as follows: An node was created for each user with metadata information. An (undirected) edge between two users was formed whenever one user made a post on the talkpage of another. The edges in this model are intended to be interpreted as a mutual recognition of one another's status as a member of the Wikipedia editors' community. The talkpage is the locus of an editors involvement in Wikipedia as a community. The rationale for our edge definition is that posting on another user's talkpage (even if the utterance is not made to that user directly) is evidence the speaker is interacting with that user *as a member of the Wikipedia community* rather than merely as in interlocutor in an isolated discussion.

The resulting network contains 25826 nodes and 85731 edges. There were 555 unconnected components totaling 571 users which were left out of the final analysis. Eigenvector and betweenness centrality were measured for each user in the network (Hagberg et al., 2008). Of the 25826 users in the network, 1822 are admins. In order to compare network centrality and adminship directly, we consider a user a to be *highly central* for some centrality measure T if $T(a)$ ranks among the highest 1822 for any user in the network. Likewise we consider a user to be *highly active* if she ranks among the highest 1822 users by edit count.

Linguistic style coordination was calculated for eight categories of commonly used function words. For each category, a list of words was compiled using frequency counts from the British National Corpus (Leech et al., 2014). Coordination towards some target group, was calculated per-user for each marker in the same manner as described in section 3.1. Per-user aggregate measures were also calculated across markers. Independent t-tests on these distributions were used to establish significance of coordination differences across target groups.

In addition to coordination towards a group, we define the coordination towards an individual from a group of users, $C^m(B, a)$ analogously. This notion was used to calculate the coordination that each user receives from the group of all users, *coordination received*. Using coordination received, we were able to correlate coordination with non-binary characteristics on individual users.

The measure $C^m(b, A)$ is only defined if there are occurrences of utterances exhibiting m made by individuals in A to which b has replied. Three aggregate measures of an individual's coordination across markers result from the three ways to deal with these undefined values.

Aggregate 1 Only consider individuals for whom all coordination is defined for all markers. The aggregate coordination is the mean coordination across all markers.

Aggregate 2 Replace undefined coordination measures with the average value that marker receives across all individuals for whom it is defined.

Aggregate 3 Take the average coordination for the markers that are defined for the individual, ignoring missing values.

5.2 Results

5.2.1 Centrality and Coordination

	agg. 1	agg. 2	agg. 3
eigenv.	0.2068	0.1205	0.2038
between.	0.2257	0.1348	0.2291
# edits	0.1763	0.0790	0.1736

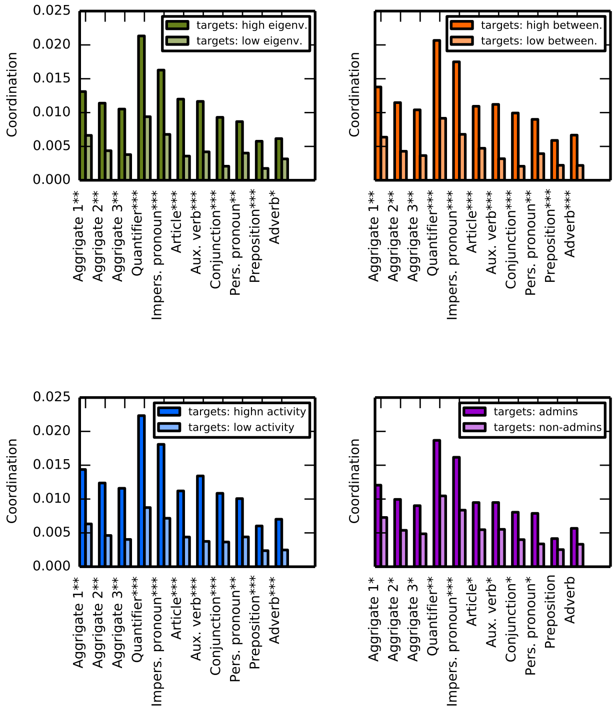
Table 1: Spearman correlation between centrality/activity and coordination received ($p < 0.001$ for all values).

Spearman correlations to *coordination received* were calculated for both measures of centrality and for level of activity (Table 1).

Next, per-marker and aggregate coordinations were calculated for each binarized centrality measure and for adminship (Figure 1). Here the focus is the effect of membership in a certain class of influential users has coordination.

5.2.2 Adminship and Centrality

To investigate the connection between adminship and network centrality, we looked at the average centrality of admins versus non-admins. It was observed that admins are more central on average than non-admins (Table 2). We also looked at what percentage of admins belong are also in the "highly central" categories created by binarizing the centrality measures to match adminship.



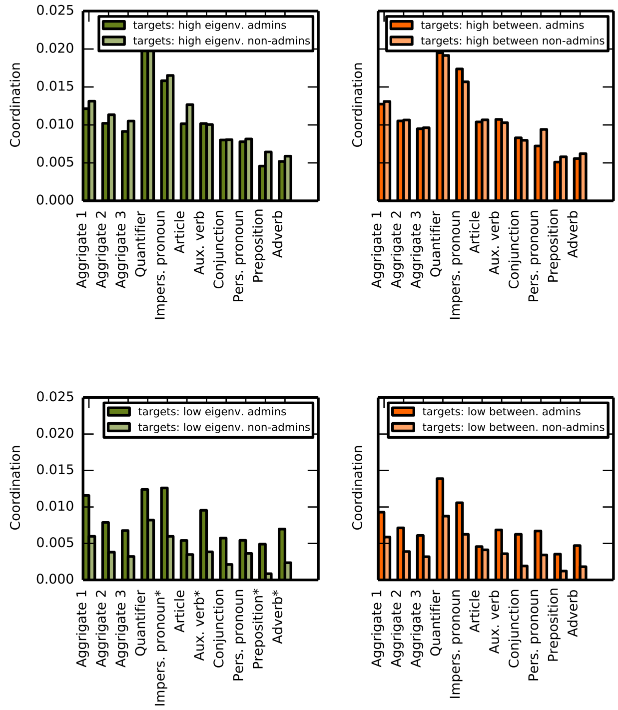


Figure 1: Linguistic style coordination towards targets with high/low eigenvector/betweenness/edit counts and towards admins/non-admins. All measures are marked for significance by independent t-test as follows: $* = p < 0.05$, $** = p < 0.01$, $*** = p < 0.001$.

Figure 2: Linguistic style coordination towards admin/non-admin targets within high/low centrality groups.

	Mean Closeness			
	Admin	Non-Admin	Corr.	% Central
eigenv.	0.0106	0.0163	0.3946	43.7
between.	0.0852	0.0063	0.2238	39.0

Table 2: Relationship between adminship and centrality across three measures. Correlations coefficients are point-biserial measures between adminship and centrality as a scalar value. Independence t-tests and p-values for all means and correlation are < 0.001. The final column shows the percentage of admins who are also highly central (as defined in section 5.1.)

5.2.3 Separating Effects on Coordination

Per-marker and aggregate coordination were calculated for admins and non-admins within each class of highly-central users. No significant effects of adminship were found (Figure 2). Analogous calculations were made for the classes of users with low-centrality. Among users with low centrality, adminship has some observable positive effect on coordination, which is significant for most markers in the low eigenvector class, but not significant for betweenness (Figure 2).

5.3 Discussion

We observe that speakers coordinate significantly more with targets in the high betweenness and high eigenvector centrality classes than with targets in their low counterparts across all markers (Figure 1). Furthermore, aggregate coordination received is weakly, but significantly correlated with centrality (Table 1). The evidence is overwhelmingly in support of $H1$. We also found that the correlation between activity level (number of edits) and coordination received is weaker than for either of the centrality measures (Table 1). Further investigation would have to be conducted to conclude that there is not some other non-network-structural confound that can explain observed the connection between network centrality and coordination.

With respect to $H2$, there *is* a connection between network centrality and adminship, but it is not as strong as one might expect: less than half of admins are also highly central (Table 2). One explanation for this would be that admins are somewhat evenly distributed through the network on a macro level even though they are quite central locally. For example, it is natural to suppose that be that the corpus contains sub-communities which roughly map to various areas of editor expertise, and although some of these sub-communities are less central in the network as a whole, they still need administrators.

The correlation between eigenvector centrality and adminship is higher than that of betweenness centrality (Table 2). This is not surprising since eigenvector centrality is the measure more typically associated with power (Bonacich, 1987).

On $H3$, we can say with a high degree of confidence that network centrality correlates with coordination independently of the effects of adminship. First, membership in a high/low centrality class predicts a greater, difference in coordination (and with more significance) than does adminship (Figure 1). Secondly the centrality measure most correlated with *coordination received* (betweenness) is also the centrality measure with the weakest correlation to adminship. Finally, no effect of adminship was found on coordination within either class of highly central users, and the effect of adminship within the class of low-centrality users was largely insignificant. All of this leads to the conclusion that not only is the effect of centrality on coordination independent of its correlation to adminship, but the correlation of adminship with coordination may largely, if not entirely, be explained by the effects of network centrality.

This is not to say that the results in *Echoes of Power* (2012) correlating coordination with power in the Wikipedia corpus are unjustified; after all, network centrality itself is often used as a proxy for power. It does, however provide evidence against the idea that linguistic coordination follows *explicit status-granted* power since that is exactly what network centrality is not.

6 Conclusion

In this paper we have put forward the hypothesis that speakers coordinate more with targets who occupy a more central position in the social context of their dialogue. We have provided evidence for this claim by measuring linguistic style coordination in the Wikipedia talkpages corpus.

6.1 Coordination in Linguistic Communities

The interactive alignment model gives an account of dialog in which priming is the mechanism by which interlocutors align their representations of a discourse at various linguistic levels (Pickering and Garrod, 2004). In the interactive alignment model, alignment at a lower level enhances and contributes to alignment at higher levels.

Analogously, one could imagine a model of linguistic communities where coordination is the mechanism by which norms are propagated and where conforming to norms at a lower level (dialogue) enhances and contributes to conformity at higher levels (the features of a linguistic community). Such a model would make a connection between two well established linguistic phenomena; that of coordination on the level of dialog, and the organic emergence of distinct linguistic norms in different communities.

If coordination is the mechanism by which a linguistic community propagates its norms, we would expect to see higher levels of coordination towards individuals who are more important in the sense that they influence the community's linguistic norms. The focus of this paper has not been to establish such a connection, but the results of our investigation in the Wikipedia talkpages corpus do provide some motivation for speculation. We have found that coordination follows network centrality more closely than it does power granted by formal status. The previously mentioned study of the propagation of attribution conventions on Twitter (Kooti et al., 2012) demonstrates that centrally located individuals are influential in the establishment of certain linguistic norms. Thus, network centrality supplies at least a modicum of evidence connecting high levels of coordination to linguistic influence.

Viewed in one light, the claim would seem almost trivial: peoples' use of language is influenced most when talking to interlocutors who are linguistically influential. In another view, these ideas propose not only a new model for the formation and evolution of linguistic communities, but also claim that linguistic accommodation has motivations subtly different from what has previously been proposed: that higher levels of coordination towards an individual may have a social goal that doesn't have anything to do with that individual in particular, but rather with adapting to the linguistic norms of a community.

6.2 Future Work

The results discussed suggest that, at least in the case of the Wikipedia talkpages corpus, there really is something to be gained by considering network structure in the analysis of linguistic coordination. This gives rise to a plethora of other questions about network structure and coordination: What are the characteristics of social networks that exhibit structural influence on coordination? In those communities where network structure does have linguistic effects, what are the best network models (i.e., edge criteria) for observing those effects? Are there characteristics of social networks as a whole (e.g., average number of neighbors, frequency of totally connected sub-graphs, average shortest path length) that can predict certain linguistic features of a community?

In this paper, we generated a social network independently any linguistic data, and then looked at the effects of network structure on certain linguistic features, but there is a lot of potential for models that use linguistic data explicitly to define a network structure. For example, if coordination is indeed connected to the adoption of new behaviors, one might use a flow network whose edges are weighted for observed linguistic style coordination to predict the dispersion of new linguistic norms.

Biran et. al, (2012), develop a method for identifying discourse participants who are likely to be influencers in online communication. If coordination is the means by which community alignment occurs, one would expect to also see greater alignment with influencers. Their method identifies influencers using coarse dialogue patterns such as percentage of posts in a thread, but a fine-grained analysis that includes measures of linguistic style coordination may improve their results.

Finally, further research is needed to support any of the above speculation that coordination at the dialog level is the mechanism by which speakers learn the linguistic norms of a community. This is not an easy claim to provide evidence for, but research on this topic is made plausible by the vast quantities of linguistic data from online communities, and by the techniques of social network analysis.

References

John M Bolland. 1988. Sorting out centrality: An analysis of the performance of four centrality models in real and simulated networks. volume 10, pages 233–253. Elsevier.

Phillip Bonacich. 1987. Power and centrality: A family of measures. *American journal of sociology*, pages 1170–1182.

Or Brian, Sara Rosenthal, Jacob Andreas, Kathleen MjKeown, and Owen Rambow. 2012. Detecting influencers in written online conversations. In *Proceedings of the Second Workshop on Language in Social Media*, pages 37–45. Association for Computational Linguistics.

Karen S. Cook, Richard M. Emerson, Mary R. Gillmore, and Toshio Yamagishi. 1983. The distribution of power in exchange networks: Theory and experimental results. *American journal of sociology*, pages 275–305.

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 WWW*, pages 699–708.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovecc, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of WWW*.

Penelope Eckert. 1988. Adolescent social structure and the spead of linguistic change. volume 17, pages 183–207. Cambridge University Press.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, August.

Molly E. Ireland, Richard B. Slatcher, Paul W. Eastwick, Lauren E. Scissors, Eli J. Finkel, and James W. Pennebaker. 2011. Language style matching predicts relationship initiation and stability. *Psychological Science*, 22(1):39–44.

Farshad Kooti, Haeryun Yang, Meeyoung Cha, Krishna P. Gummadi, and Winter A Mason. 2012. The emergence of conventions in online social networks. In *ICWSM*.

G. Leech, P. Rayson, and A. Wilson. 2014. *Word Frequencies in Written and Spoken English: based on the British National Corpus*. Taylor & Francis.

Martin J. Pickering and Simon Garrod. 2004. The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, 27:212–225, 4.

Stanley Wasserman. 1994. *Social network analysis: Methods and applications*, volume 8. Cambridge university press.