A robust framework for estimating linguistic alignment in social media conversations

Gabriel Doyle
Department of Psychology
Stanford University
Stanford, CA 94305
gdoyle@stanford.edu

Dan Yurovsky
Department of Psychology
Stanford University
Stanford, CA 94305
yurovsky@stanford.edu

Michael C. Frank
Department of Psychology
Stanford University
Stanford, CA 94305
mcfrank@stanford.edu

ABSTRACT

When people talk, they tend to adopt the behaviors, gestures, and language of their conversational partners. This "accommodation" to one's partners is largely automatic, but the degree to which it occurs is influenced by social factors, such as gender, relative power, and attraction. In settings where such social information is not known, this accommodation can be a useful cue for the missing information. This is especially important in web-based communication, where social dynamics are often fluid and rarely stated explicitly. But connecting accommodation and social dynamics on the web requires accurate quantification of the different amounts of accommodation being made.

We focus specifically on accommodation in the form of "linguistic alignment": the amount that one person's word use is influenced by another's. Previous studies have used many measures for linguistic alignment, with no clear standard. In this paper, we lay out a set of desiderata for a linguistic alignment measure, including robustness to sparse and short messages, explicit conditionality, and consistency across linguistic features with different baseline frequencies. We propose a straightforward and flexible model-based framework for calculating linguistic alignment, with a focus on the sparse data and limited social information observed in social media. We show that this alignment measure fulfills our desiderata on simulated data. We then analyze a large corpus of Twitter data, both replicating previous results and extending them: Our measure's improved resolution reveals a previously undetectable effect of interpersonal power in social media interactions.

Categories and Subject Descriptors

[**Applied Computing**]: Psychology; [**Applied Computing**]: Sociology; [**Human-centered computing**]: Social media; [**Human-centered computing**]: Social networks

General Terms

social media, conversation, social networks, language, communication, coordination, social status, power, psychology

1. INTRODUCTION

When people interact, they tend to act similarly, adopting similar postures, speaking in similar ways, and using similar words. This *communication accommodation* [20] is a pervasive part of human interactive behavior, arising in many different dimensions of interaction, including gesture, posture, tone, and language use [11, 3, 21, 32, 26, 9]. From a scientific perspective, greater degrees of accommodation can signal power relationships or affiliation [46, 23, 14], and from an engineering perspective, interactive agents that accommodate are seen as friendlier and more human [34]. One of the most important and well-studied forms of accommodation is *linguistic alignment*, in which conversational partners align aspects of their communicative style and content to one another. Roughly speaking, linguistic alignment can be measured as the change in likelihood of using a given "marker" – most often a word (e.g., *you*) or word category (e.g., prepositions) – based on its use by a partner.

But while the basic idea of linguistic alignment has been used in a range of studies across fields, it has been quantified using a variety of substantially different measures [27, 12, 17]. Some measures conflate influences that others separate, some combine features that others do not, and some account for individual speaker differences that others do not. Further, it is unclear whether these measures are appropriate for sparse interactional data observed in social media and other web-based settings (though see [47]).

Our goal in this paper is to address the issue of inconsistent measures across studies of linguistic alignment. We begin by describing a set of desiderata for linguistic alignment measures. We then report simulations showing that existing measures of alignment fail when faced with sparse linguistic data of the type that are common on the web. We propose a new model-based alignment metric and show that it fulfills our desiderata. We end by using this new metric to analyze Twitter data, and show that it succeeds in detecting the the influence of power dynamics on accommodation behavior in sparse data where no effect had been detected previously.

2. PRIOR WORK ON ACCOMMODATION AND ALIGNMENT

Accommodation is a general and deeply-ingrained human behavior. Children as young as 12 months old accommodate to their parents on pitch [33], and fictional dialogues show similar alignment to real ones. These two datapoints suggest that accommodation is a crucial and "unmediated" mechanism [41, 13]. Accommodation can even influence human-computer interactions, with people rating interactions with accommodating computer systems as more satisfying even when the conversant is known to be a computer [34, 44, 5].

Linguistic accommodation, which we will refer to as *alignment*, has been one of the key domains in which hypotheses about accommodation have been tested. A major branch of this work is known as Linguistic Style Matching (LSM) [36, 27]. The focus of LSM is on "stylistic" accommodation, as opposed to content accommodation; practically, LSM examines the reuse of function words (prepositions, pronouns, articles, etc.) that carry little inherent semantic information, as opposed to content words (nouns, verbs, etc.). This focus on function words is motivated by the argument that function words represent a stylistic choice, as a speaker can choose between many different function words in composing a message without substantially changing its meaning. Function word use has been shown to vary between people, but to remain fairly consistent within a single person's writing [40]. As such, it has been fruitfully applied to authorship attribution as well [4].[1]

Accommodation, especially linguistic alignment, can be a critical part of achieving social goals. Performance in a variety of cooperative decision-making tasks has been positively related to the participants' linguistic convergence [17, 29]. Match-matching in speed dating as well as stability in established relationships have been linked to increased alignment [27]. Alignment can also improve persuasiveness, encouraging listeners to follow good health practices [30] or convincing children to share more [6].

Alignment typically is convergent, making the conversants more similar, but the degree and direction of alignment differs from situation to situation. In some situations people may diverge, intentionally or not; this divergence is often tied to a particular social goal, such as maintaining an appropriate power dynamic between doctors and patients [16]. In addition, different dimensions or features may exhibit convergence at different strengths [42, 1, 12] and/or different time-scales [16]. Lastly, alignment and accommodation are usually incomplete, in that people become more similar but not the same. For instance, [20, 22] show that near-complete accommodation can come off as cloying or derisive.

Variability in accommodation behavior can be sociologically and psychologically meaningful. Power relationships are an important source of differential accommodation, with less powerful conversants generally accommodating more to more powerful conversants. Prominent examples of such asymmetric accommodation include interviews and jury trials [46, 23, 14]. Additionally, factors such as gender, likability, respect, and attraction all interact with the magnitude of accommodation [1, 35]. Such differences in accommodation can also be indicative of changes to the power dynamic: In U.S. Supreme Court transcripts, [25] showed that depending on the accommodation dimension, justices – who are more powerful by any intuitive assessment – may nevertheless accommodate more to lawyers, perhaps because the lawyers have the local power to answer justices' questions.

[1]This interest in function words over content words is effectively the mirror image of many document classification methods, such as topic models [2], which focus on content word co-occurrences and typically *exclude* function words.

3. DESIDERATA FOR MEASURES OF LINGUISTIC ALIGNMENT

Accommodation and related concepts have been approached from many different fields of study, leading to many different approaches to quantifying linguistic alignment. On one hand, this is a helpful proliferation. Alignment must be a very robust characteristic of human socialization if its effects appear using so many different estimation methods. On the other hand, measures used in different studies are difficult to compare, and some studies separate factors that other conflate. From the perspective of standardization and comparison, a single measure would allow further theoretical progress. But what measure to select?

In what follows, we propose a set of desiderata for linguistic alignment measures. At the highest level, the goal of a measure of linguistic alignment should be to quantify the amount that one person's language use is influenced by another's: we are interested in seeing how much a person changes when speaking to different people, and to what extent such changes increase the similarity between the speakers. Also, because linguistic alignment can, in principle, be measured on many different words and categories, we want a measure that can be compared across linguistic features with very different frequencies. Furthermore, because different features may align differently [1, 16], we want a measure that is flexible enough to account for these differences. Although previous desiderata have been proposed [47], these requirements focused on satisfying theoretical goals (e.g., consistency across different structural levels). Such goals are important, but presuppose the basic statistical properties mentioned above. Our concern here is with establishing these more basic statistical desiderata.

3.1 Conditionality and baselining

An alignment measure must provide a measure of directional linguistic influence, not just general similarity. In addition, many existing measures fail to account for the possibility that speakers may already be very similar before they start talking. One example of the importance of conditionality is [27], who show that speed daters with more similar word distributions are more likely to form a connection. But, because they do not control for the similarity of daters' language use independent of each other, they may actually be measuring similarity of daters' backgrounds rather than alignment.

The relevant theoretical distinction is between *accommodation* and *homophily*. If two people speak in a similar manner, it may be that they have *observed* each other's style of speech and have aligned to each other (accommodation) [12]. However, it also may be that these two people happen to have *inherently similar* speaking styles, perhaps because they have similar linguistic backgrounds or similar linguistic pressures (homophily). Accommodation and homophily are likely to

have similar effects on outcome measures such as comprehension, likability, and task success, based on similarity-attraction theories, but differ substantially in terms of their theoretical import [7, 43, 22].

Conditionality is perhaps the most critical desideratum for a measure of alignment: without it, any result may be due to mere homophily. To address this issue, Danescu-Niculescu-Mizil and colleagues [12, 13, 14] subtract off a speaker's average frequency of using a linguistic marker when calculating alignment, an important advance over other methods. The Hierarchical Alignment Model we present here extends their SCP measure to satisfy the full set of desiderata below.

3.2 Separability across markers

Accommodation is not a monolithic process; people may converge on some dimensions while diverging on others [1, 16]. In fact, similarly high levels of accommodation on multiple dimensions may even be counter-productive, giving the impression of mocking or condescending to the audience [22, 20]. Empirically, for specifically linguistic alignment, different markers may have distinctly different alignments [12, 27]. For instance, we may not expect second person pronoun (e.g., *you*) usage to align, since one speaker's *you* is the other speaker's me. These differences can have important implications for applications of alignment; [17] found that increased alignment on expressions of confidence improved group performance in a task, but across-the-board increases in alignment *reduced* group performance. Thus, we want a measure that can estimate different alignment values for different markers, with the possible option of aggregating over markers when needed.

3.3 Consistency across varying marker frequencies

Different words have radically different baseline frequencies: a few words are used very often, but the bulk of our vocabularies are rarely used. As discussed above, it is undesirable to aggregate alignment values across markers, but consistency in our alignment measure is important if we want to investigate how (or whether) baseline frequencies interact with alignment. As such, we want a measure for which alignment effect strengths are not significantly biased by baseline word frequencies. To assess measures on this desideratum, we will test potential alignment measures against simulated data with known alignment strengths and marker frequencies to ensure comparability across a wide range of marker frequencies.

3.4 Robustness to sparse data

Much work on communication accommodation and linguistic alignment, especially early work, focused on cases where a small set of people interact extensively, allowing accommodation effects to be estimated from a fairly large dataset [16, 24, 27]. In many applications, especially those on the Web, however, datasets have the opposite character: a large number of people interact briefly and data about any given interaction is sparse. For example, in the Twitter dataset we examine here, many of our interacting pairs exchange only two messages, containing a maximum of 280 characters. These passing interactions may be importantly different from repeated interactions with close friends. Our measures must be robust enough to extract accurate alignment values from these sparse interactions, so that they can be compared against estimates from more extensive interactions.

4. EXISTING MEASURES FOR LINGUISTIC ALIGNMENT

Rather than giving a comprehensive review, this section provides a sampling of some influential measurement methods, and discusses how they fit the desiderata described above.

4.1 Subtractive Conditional Probability (SCP)

Danescu-Niculescu-Mizil and colleagues [12] presented a subtractive conditional probability measure, capturing the increase in the conditional probability of using a marker, given that it has been used by a conversational partner. Consider a set of messages from speaker a that each gets a reply from speaker b. Let A indicate that a used the marker in a message, and B indicate that b used the marker in a reply. Then the subtractive conditional probability alignment score is:

$$SCP = p(B|A) - p(B) \tag{1}$$

This measure satisfies the conditionality/baselining condition because the alignment estimate takes into account how much more likely b is to use the marker in response to a than some baseline.[2] In addition, SCP is calculated independently for each marker, so it meets the marker separability criterion. It is the only existing measure we will look at that satisfies both of these desiderata.

It fails on the marker comparability criterion, however. First, the range of possible alignment values for SCP depends on the baseline $p(B)$, with the alignment estimate falling in the interval $[-p(B), 1-p(B)]$. In addition, $p(B) = p(B|A)p(A) + p(B|\neg A)p(\neg A)$, making the alignment range also dependent on $p(A)$:

$$\begin{aligned} SCP &= p(B|A) - (p(B|A)p(A) + p(B|\neg A)p(\neg A)) \\ &= (1 - p(A))(p(B|A) - p(B|\neg A)) \end{aligned}$$

This definition means that the range of the SCP alignment estimate for a given marker is the intersection of the intervals $[-p(B), 1-p(B)]$ and $[2(p(A)-1), 2(1-p(A))]$, making the direct comparison of alignment on markers with different baseline frequencies difficult; this point is illustrated through simulations in Section 6. Lastly, in previous work SCP has been applied only to conversations with at least 10 messages; we discuss its robustness to sparse data below.

4.2 Local Linguistic Alignment (LLA)

Local linguistic alignment (LLA) was originally proposed by [17]; we use the formalization from [45]. Suppose a sends message M_a to b, who replies with M_b. Then:

[2]The computation of this baseline could be questioned, however; [12] limit the calculation of $p(B)$ to the conversations between a and b, not all of b's conversations.

$$LLA = \frac{\sum_{w_i \in M_b} \delta(w_i \in M_a)}{length(M_a) length(M_b)} \quad (2)$$

Intuitively, LLA is the percentage of words in the reply that also appeared in the first message, divided by the length of the first message. This fulfills half of the conditionality/baselining desiderata; the numerator is a conditional distribution, only counting words that have been repeated in the reply. But no baselining is being done to separate homophily from alignment; if two speakers happen to have similar vocabulary distributions, they can end up with high LLA values without accommodating each other at all.

"Discriminate" LLA, where only words from a particular category are counted, meets the marker separability desideratum. "Indiscriminate" LLA, which counts all words, does not meet this desideratum, and [17] show opposite alignment effects on task performance depending on whether the disparate markers are treated separately or lumped together. For these reasons, we only test the discriminate LLA method in our simulations. Discriminate LLA is also not consistent across different message lengths, as the maximum value of the numerator is $length(M_b)$, meaning that the LLA value is bounded above by $1/length(M_a)$. Replies to short messages, then, have higher maximum LLA values than replies to long messages. We also find evidence of inconsistent behavior on markers with different frequencies in simulated data.

4.3 Linguistic Style Matching (LSM)

A great deal of the work on linguistic alignment from a psychological perspective comes from Pennebaker and colleagues [36, 24, 27], including the Linguistic Inquiry and Word Count system that we use below to establish our marker categories [39]. The alignment measure used in this work is Linguistic Style Matching (LSM). As with the SCP measure, suppose we have a set of messages exchanged between a and b, and A (or B) indicates that a (or b) has used the marker. Then LSM is defined as:

$$LSM = 1 - \frac{|p(A) - p(B)|}{p(A) + p(B)} \quad (3)$$

This measure does not meet the conditionality/baseline desideratum, as it may reflect homophily rather than alignment. Consider a pathological dyad, where the replier refuses to align. In this case, the replier b will use the marker only when the initiator a does not, but never use it when a does. The intuitive sense of alignment is that b is being divergent, and should have a large negative alignment score, while a is simply not aligning at all. If a uses the marker in about half of their messages, though, $p(A) \approx p(B) \approx 0.5$, and the LSM for this pair would be near a perfect 1. In addition, if a replier shows no actual alignment to their conversation partner, but the coincidentally have similar distributions, they could show a much higher LSM score than two people with very dissimilar inherent distributions who are making an effort to align.

LSM fits the marker separability desideratum, as it is independently calculated for each marker; however, we will show on simulated data that LSM behaves very differently on markers of different probabilities is affected by data sparsity.

5. HIERARCHICAL ALIGNMENT MODEL

5.1 Motivation

Building on the probabilistic intuition motivating the SCP measure, we propose a Hierarchical Alignment Model (HAM). Specifically, our goal is to use model-based estimation of conditional probabilities to create a measure that is consistent across different marker frequencies and robust to sparse data.

To handle large baseline differences in marker frequency, we change the form of our alignment estimate from an additive effect in probability space to an additive effect in log-odds space. Having alignment as an additive effect in probability space means that the range of possible alignment strengths moves with the baseline; a more frequent marker can not show as much convergence as a less frequent marker. If alignment is a linear effect in log-odds space, then alignment is defined in the range $(-\infty, +\infty)$ and is freed from the biasing influence of baseline probability, since even high probabilities can increase by large amounts in log-odds space.

To improve performance on sparse data we introduce a hierarchical prior on alignment values, parameterized by both marker and dyad (speaker/replier pair). In doing so, we introduce an assumption that, unless the data argues strongly otherwise, dyads are likely to have similar alignment behaviors on a given marker. This leads to more accurate estimates of marker frequency and alignment in sparse data dyads, since they are influenced by the alignment found in the dataset as a whole. This kind of hierarchical regularization has proven extremely valuable in a wide variety of applications [18].

5.2 Model

We begin by conceptualizing a conversation as a tree; each message, aside from the first, is in response to a particular preceding message, but a message may elicit multiple replies. Tweets already fit this structure, because Twitter replies explicitly include which tweet a reply is directed to. In settings without explicit reply structure, a message can be treated as a reply to all messages that came before it [45] or thread reconstruction can be used to find individual reply links [28].

Suppose we observe a conversation that starts with Alice, who says "hi." Bob replies to this message with "hello there," and Alice responds with "how are you?". In addition, Eve jumps into the conversation by replying to Bob ("hi Bob!"). This conversation provides three message pairs:

$$[(A, \text{hi}), (B : \text{hello there})]$$
$$[(B : \text{hello there}), (A : \text{how are you?})]$$
$$[(B : \text{hello there}), (E : \text{hi Bob!})]$$

Following [12], we treat messages as binary vectors over words, rather than probability distributions or counts over

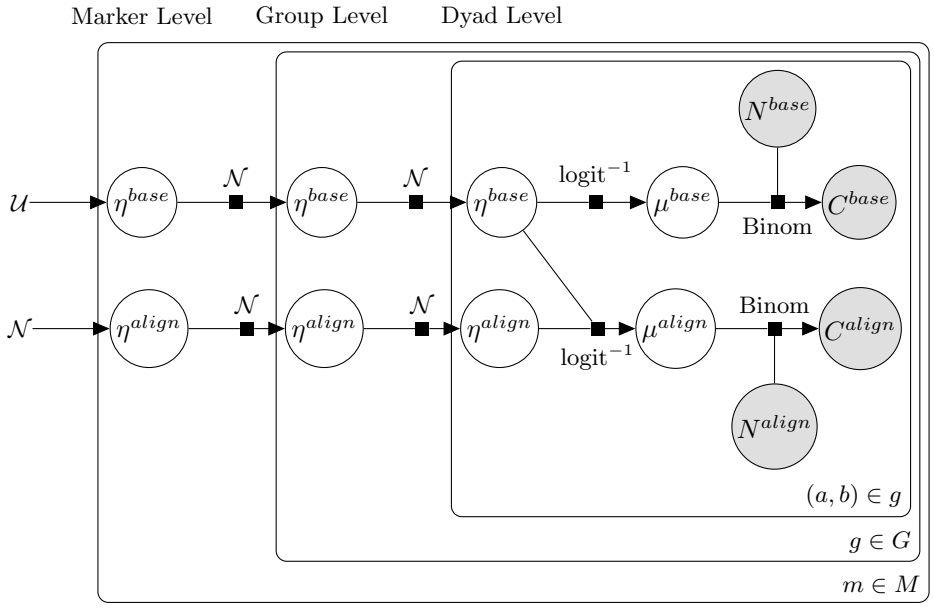

Figure 1: The Hierarchical Alignment Model (HAM). A chain of normal distributions generates a linear predictor η, which is converted into a probability μ for binomial draws of marker presence/absence.

words. Because these are conversational messages, they tend to be short – in the particular case of Twitter, messages contain on average approximately 3 to 6 markers—and thus this binarization instead of a count is not a severe simplification.[3] Alignment then is an increase in the probability of seeing a given marker (or marker category) in the second message of a pair given that it appeared in the first message of that pair.

Figure 1 gives our graphical model for alignment. We treat pairs where the first message contained the marker separately from those where the first message did not contain the marker. The number of message pairs between a dyad of speakers (a, b) is split into $N^{base}_{m,a,b}$, the number of pairs where a did not use the marker m, and $N^{align}_{m,a,b}$, the number of pairs where a did use m. We also calculate the count of message pairs where the replier b used the marker m when a did not ($C^{base}_{m,a,b}$) and when a did ($C^{align}_{m,a,b}$). These counts are assumed to come from binomial draws with probability μ^{base} or μ^{align}. These μ values are generated from the η values in log-odds space by an inverse-logit transform, similar to linear predictors in logistic regression.

We implement alignment on these η values; the η^{base} variables are representations of the baseline frequency of a marker in log-odds space, and μ^{base} is simply a conversion of η^{base} to probability space, the equivalent of an intercept term in a logistic regression. η^{align} is an additive value, with $\mu^{align} = \text{logit}^{-1}(\eta^{base} + \eta^{align})$, the equivalent of a binary feature coefficient in a logistic regression. Alignment is the change in log-odds of the replier using m above their baseline usage of the marker.

[3]In fact, the shortness of the messages could introduce substantial noise via normalization within tweets, reducing robustness to sparse data.

Going up the hierarchy, the inner plate η values are specific to each marker-dyad combination (m, a, b), and the μ values are calculated based on these. One level above this are η values for marker-group combinations (m, g). "Group" here is an intentionally vague classifier for groups of dyads. In the present work, we divide dyads into groups based on the power differential in the dyad. Group divisions based on gender, conversation role, or any other variables are also possible, and this layer may be omitted in cases where no group effects are being studied.

Lastly, there are η values at the marker level m, shared across all groups $g \in G$. For η^{align}, this is a normal distribution centered at 0, biasing the model equally in favor of positive and negative alignments. For η^{base}, the overall frequency of a marker, we set the prior on this highest level parameter to be an uninformative uniform distribution over $[-5, 5]$, as there is no strong reason to expect one particular marker frequency over another. This range was chosen for convenience; it translates to approximately $[.006, .993]$ in probability space, which is well beyond the $[0.1, 0.6]$ range of marker frequencies considered here.[4] Each layer of the η parameters is generated by a normal distribution with variance σ^2.

5.3 Model Fitting and Inference

Alignment measures can be extracted from multiple different levels of this model hierarchy; we focus on the $\eta^{align}_{m,s}$ parameter, which is a single value estimating the mean alignment by dyads within a subpopulation. This value represents the

[4]While in principle an unbounded distribution is appropriate, [19] discuss the $[-5, 5]$ interval for logistic regression coefficients as capturing likely coefficients; for modeling markers with very high or low probabilities, the Cauchy distribution recommended in that paper could replace the uniform distribution, allowing (rare) extreme η^{base} values.

change in the log-odds of using m when replying to someone who has already used it, which constitutes our operationalization of alignment.

We implemented this model in RStan [8], with code available at `http://github.com/langcog/alignment`. The model is fit with 200 iterations of the sampler (100 discarded as burn-in) for each dataset; judging from trace plots, this setting led to reliable convergence. We then extracted alignment estimates from each of the final 100 iterations of the model, and we report the 95% highest posterior density interval on the parameter values in these plots.

6. EXPERIMENT 1: SIMULATIONS

We begin by testing the HAM model for consistency across different marker frequencies and robustness to sparse data. We report the results of two simulations: the first generates alignment on a per-message basis (more like the SCP/HAM measures' assumptions), and the second generates alignment on a per-word basis (presumably more similar to actual production). We show greater consistency and more accurate alignment estimates for the HAM model over the existing measures in both cases.

6.1 Simulation 1: Per-message Alignment

Our first simulation uses a simple generative model that treats the presence or absence of a marker in a message as the relevant quantity. Instead of attempting to estimate word productions based on unigram probabilities and known message lengths, we simplify the process by generating messages with a given probability of containing the marker, and treat alignment as an adjustment to that per-message frequency.

We start by generating a set number of dyads, in this case 500, each a pair of people who are observed talking to each other. For each dyad, we draw a number of message pairs from a geometric distribution with mean 5. This is a sparser dataset compared to our Twitter dataset, which contains 16864 dyads with a mean of 9.94, but is generally representative of the Web-based setting as it has a large number of dyads with a small number of messages. By testing on this sparse dataset, we can detect a lack of robustness to sparse data as well.

For each message pair, we perform a Bernoulli draw with probability p that the first message will contain the marker.[5] If the first message does not contain the marker, there is no alignment and the probability that they reply contains the marker is p. If the first message does contain the marker, we change the probability of the reply containing it by adding an alignment strength α in log-odds space, as this keeps the probabilities inside the $[0, 1]$ range. $\alpha = 0$ implies no alignment; positive α indicates linguistic convergence, negative α divergence. We test over a range of marker baseline probabilities that cover the range of the marker frequencies seen in the Twitter data. We do not evaluate LSM and LLA in this simulation because they require by-word calculations; see Simulation 2 for results with these measures.

[5]Conditionalized, baselined alignment can not be estimated unless the first member of the dyad has at least one message that contains the marker and one message that does not contain the marker. Dyads that do not meet the criterion were thus re-drawn.

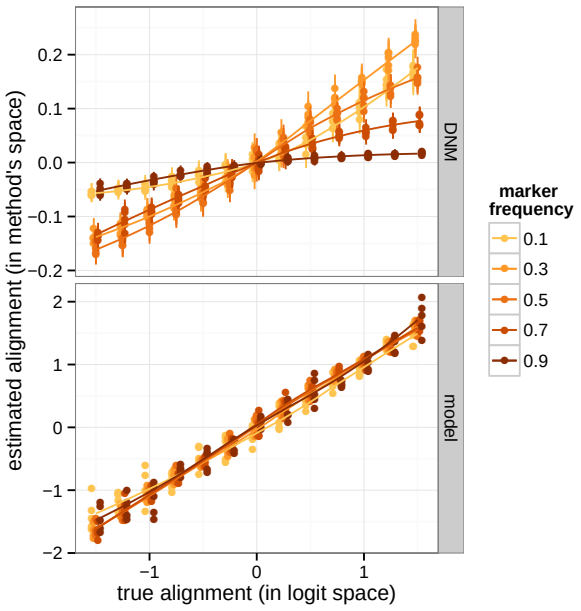

Figure 2: Results from simulation 1. Plot shows the actual alignment values from the simulations against the model-inferred values of the alignment. Lines are loess-fit curves. Dots show multiple independent simulation runs. HAM shows more consistent behavior across marker frequencies than SCP.

Figure 2 shows alignment values recovered by SCP and HAM measures, based on five simulated datasets for each combination of alignment and marker frequency. SCP shows a decreasing slope as the marker baseline frequency increases. Additionally, within a given baseline frequency, the relationship between true and estimated alignment is non-linear, especially at high or low marker frequencies and alignments. Both of these indicate substantial bias in the SCP measure based on frequency and alignment. In contrast, HAM shows a linear relationship between true and estimated alignment, and the slope of the true-estimated alignment relationship is consistent across different marker frequencies. Thus, HAM (but not SCP) satisfies the desideratum of being consistent across differing marker frequencies and is robust in its estimates from sparse data.

6.2 Simulation 2: By-word Generation

Our second simulation uses a slightly more complex generative process, first generating a length for a message and then filling in words within the message. This process moves closer to the true generative process underlying person-to-person conversation; because it generates full messages it also allows testing of the LLA and LSM measures.

We again start by generating 500 interacting dyads, exchanging a mean of 5 message pairs. The number of words in each message is drawn from a uniform distribution on the interval $[1, 25]$, approximating the 140-character limit of

Twitter. We specify a unigram probability p for the marker. The first message in each pair is generated with this unigram probability. If the initial message does not contain the marker, the reply is also generated with unigram marker probability p. If the initial message does contain the marker, the reply marker unigram probability is changed by α in log-odds space. We vary the marker frequencies over a range representative of common words (.005 $\approx$ *by*, .01 $\approx$ *that*, .05 $\approx$ *the* in [31]) or word categories (.1 $\approx$ personal pronouns, .2 $\approx$ all pronouns, in [29]), appropriate comparisons for the marker categories used in the Twitter experiments.

Figure 3 shows the relationship between true and estimated alignments over the four measures, again based on five runs for each alignment and marker frequency combination. As in the earlier simulations, the SCP measure is positively correlated with true alignment, but the relationship is somewhat non-linear and dependent on the marker frequency. With conditionality but no baselining, LLA is able to detect changes in (positive) alignment strength, but is greatly affected by a marker's baseline frequency, as expected. In contrast, LSM fails to correctly capture alignment in this simulation, detecting the greatest "alignment" when there is *no* simulated alignment. Since speakers in this simulated dataset all have the same baseline marker frequencies, if they speak independently of each other, their rate of marker usage will be approximately the same. If the replier conditions their marker use on the initial speaker, the replier's rate of marker usage will move *away* from that of the speaker, reducing LSM. This simulation thus shows that LSM actually quantifies the homophily of a dyad, rather than the alignment—providing evidence that the conditionality/baseline desideratum is critical for separating out alignment from homophily. HAM performs best, showing consistency across markers and robustness in getting accurate estimates from sparse data.

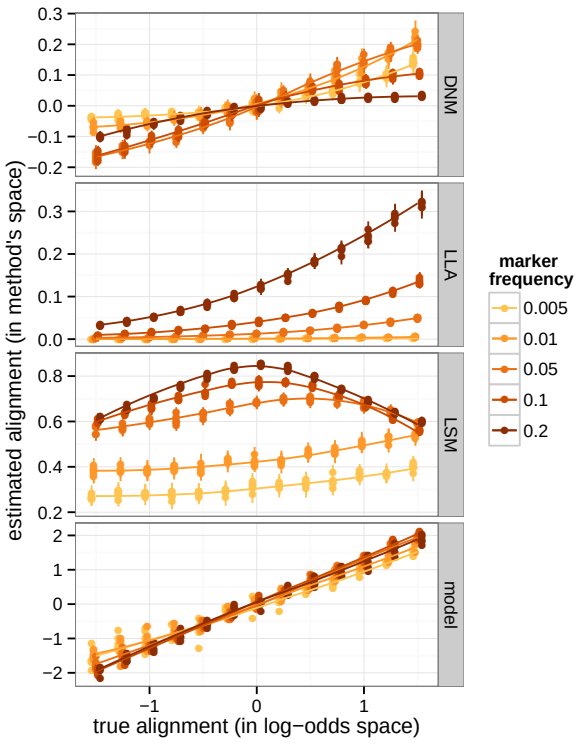

Figure 3: Results from simulation 2. Plot shows the actual alignment values from the simulations against the model-inferred values of the alignment. Lines are loess-fit curves. Dots show multiple independent simulation runs. HAM shows more consistent behavior across marker frequencies than any other measure.

7. EXPERIMENT 2: TWITTER DATA

We next turn to an examination of alignment on Twitter. Because of its size, the diversity of its userbase, and the public accessibility of its data, Twitter is an important source of data about naturalistic linguistic interaction that overcomes many of the limitations of other more restricted sources. On the other hand, as noted above, the brevity of tweets and the sparsity of individual interactions pose substantial challenges to investigators.

Alignment on Twitter was initially investigated by [12], who found overall positive alignment on all 14 tested marker categories, but no significant effects of social power/status on alignment. This null finding was contrary to findings of power/status effects in many other settings [23], including some Web-based settings [14, 37]. While it is possible that social media does not display power/status-based differences in alignment – perhaps because status differences are less obvious in the social media setting – it is also possible that differences in that earlier study were masked by precisely the problems we highlight above, namely sparse data and widely varying baseline marker frequencies. Using the HAM model, we provide evidence for this latter possibility.

7.1 Corpus

We use a collection of Twitter conversations collected by [15] to examine information density in conversation. This corpus focuses on conversations within a set of 14 mostly distinct subcommunities on Twitter. These subcommunities contain all the messages exchanged between Twitter users who sent at least one message to a "seed user" with a reasonably large number of followers. This corpus contains 63,673 conversation threads, covering 228,923 total tweets. We divide these conversations into message pairs, also called conversational turns, which are two consecutive tweets within a conversation thread. The second tweet is always an explicit reply to the first, and the two tweeters in the pair must be distinct users (i.e., no self replies are included).

One additional piece of processing was done: while formal retweets (sharing another tweeter's message to one's own timeline) are removed from the data automatically, there are also informal retweeting methods that are not marked as retweets by the Twitter API. We therefore removed all pairs where the reply tweet contained all of the words of its preceding tweet and additionally had either the bigram *RT @username:* (indicative of a "manual retweet") or Unicode curly quote characters (indicative of a type of quoting that some Twitter apps use). Including such retweets would artificially inflate the alignment scores, as the entirety of the previous message was included in the reply, but not as the

Table 1: Marker categories for linguistic alignment, with examples, number of distinct word types, and probability of appearing in a tweet.

Category	Examples	Size	$p(A)$
Article	*a, an, the*	3	.44
Certainty	*always, never*	83	.18
Conjunction	*but, and, though*	28	.39
Discrepancy	*should, would*	76	.20
Exclusive	*without, exclude*	17	.27
Inclusive	*with, include*	18	.30
Indefinite pronoun	*it, those*	46	.39
Negation	*not, never*	57	.21
Preposition	*to, in, by, from*	60	.58
Quantifier	*few, many*	89	.26
Tentative	*maybe, perhaps*	155	.23
1st person singular	*I, me, mine*	12	.57
1st person plural	*we, us, ours*	12	.14
2nd person pronoun	*you, yourself*	20	.25

replier's own words. This processing leaves us with 122,693 message pairs, spanning 2,815 users.

The tweets were parsed into word tokens using the Twokenizer [38], with usernames and URLs removed. We then calculated linguistic alignment on the fourteen marker categories used by [12] in their study of Twitter messages. These categories come from the Linguistic Inquiry and Word Count (LIWC) system [39]; category names and sample words are shown in Table 1. These categories were chosen (from the complete set of 74 LIWC categories) as "strictly non-topical style dimensions" as they had limited to no content words in them, and were not focused on specific topics. These can be roughly divided into four pronoun categories (indefinite, 1st singular, 1st plural, 2nd), four other syntactic categories (article, conjunction, preposition, quantifier), and six conceptual categories (certainty, discrepancy, exclusive, inclusive, negation, tentative).

A tweet is counted as containing a given category if it contains at least one word from that category. Alignment is calculated based on category rather than on specific word types; thus if the first tweet contains *an* and its reply contains *the*, this pair is counted as an example of positive alignment.

7.2 Overall Alignment

[12] found significant positive alignment on all fourteen marker categories on Twitter, but did not detect an effect of power on alignment. We start by replicating the overall positive alignment result before moving on to the effects of power. Figure 4 shows the alignment values for this Twitter dataset using the measure from [12] on the left, and our model-based alignment measure on the right. The SCP measure is plotted with 95% confidence intervals from a 1000-sample bootstrap over dyads. The HAM measure is plotted with 95% highest posterior density intervals on the inferred parameter values $\eta^{align}_{m,g}$. The alignment values for the measures have a correlation of 0.40. There is no obvious relationship between the strength of the alignment values and the conceptual/syntactic/pronominal division. As expected, both measures find consistent and significant convergent alignment on all fourteen marker categories.

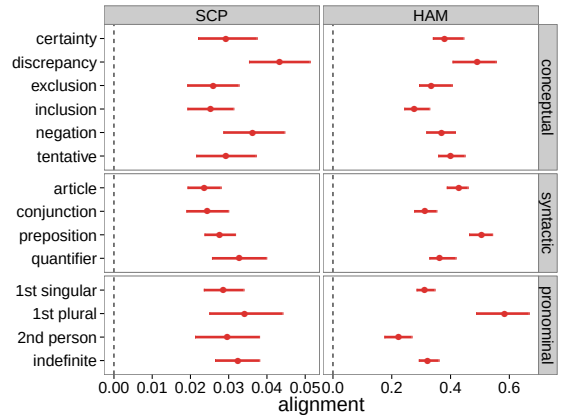

Figure 4: Overall alignment per marker category for SCP and HAM alignment. All 14 marker categories show significant convergent alignment. 95% confidence intervals shown, based on by-dyad bootstrapping for SCP and by-iteration parameter estimates for HAM.

7.3 Alignment and Power

We assess a user's power on Twitter in two ways. First, we assess power internal to the Twitter network based on the number of other users following our user of interest. Users with more followers have their tweets read by more users, get more retweets and favorites, and so on, giving them power within the network. In addition, if they retweet or reply to another user, it can substantially increase that user's status and follower count, leading low-follower users to attempt to catch their eye.

Second, we take advantage of Twitter's user verification process as an external measure of power. Twitter verifies important users to show that their accounts are not impostors or parodies. Verified accounts range from heads of state (`@POTUS, @MedvedevRussiaE`) to famous athletes (`@KingJames, @Shaq`) to Youtube stars (`@camerondallas, @pewdiepie`). Twitter only verifies users who they consider to be significant, generally for accomplishments outside of Twitter (though the service does not provide a public standard for verification). We expect these to measures of power to behave similarly; both verified users and users with high follower counts will be aligned to more than unverified users and users with small follower counts.

Intuitively, alignment to power captures the idea that we show deference to important people that we do not show to our peers. While increased alignment to the power has been observed in many non-Web settings (e.g., [25]), it is possible that the different dynamics of social media would remove or even reverse this effect. Trolls, cranks, and a variety of other non-cooperative conversationalists may fill people's Twitter timelines (some celebrities and athletes leave or avoid Twitter for this reason). We predict that social media is not so different from other settings, however, and that a sufficiently sensitive measure should find an effect of power. We also predict that the pronouns may pattern differently from the other syntactic and conceptual markers, as pronoun usage has previously been shown to differ depending on one's

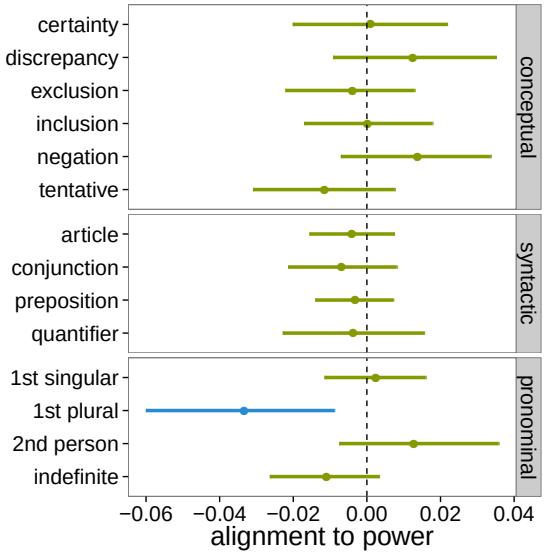

Figure 5: Difference in SCP-estimated alignment on the 14 marker categories depending on follower ratio. Positive values indicate greater alignment to high-follower users; negative indicates greater alignment to low- and equal-follower users. One category, first-person plural pronouns, shows a significant *negative* effect of power on alignment, while the rest show no effect.

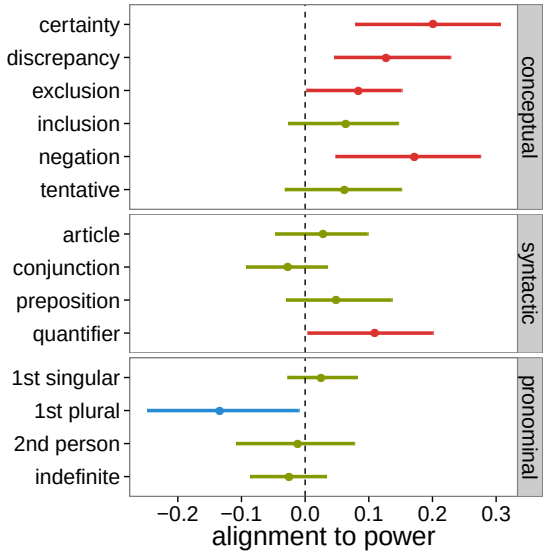

Figure 6: Difference in HAM-estimated alignment on the 14 marker categories depending on follower ratio. Five markers show significantly more alignment to power, while only one shows less alignment.

social power [10, 29].

7.3.1 *Follower ratio as a measure of power*

For each pair of users, we calculate their *follower ratio* by dividing the first tweeter's follower count by the sum of their follower count and their replier's follower count. Numbers above 0.5 indicate that the first tweeter has more followers than the replier. We use 100/101 as our cutoff, meaning that the first tweeter has at least 100 times as many followers as the replier, and thus has a substantially larger audience. 38% of our pairs have this property.[6] Under the hypothesis of increased alignment to power, we expect to find increased alignment when the follower ratio is high. We do not distinguish dyads with an even number of followers from those where the replier has substantially more followers because high-follower users very rarely reply to low-follower users.

Figure 5 shows the difference in alignment to power derived from follower counts, using the SCP measure. The intervals in this plot are 95% confidence intervals estimated by a 1000-sample bootstrap over the dyads. Only one category shows a significant effect of power on alignment, and it shows reduced alignment to powerful users, contrary to our power expectations. However, this outlier is the first person plural pronominal category (e.g., *we, us*), which is known to have different usage patterns for those with and without power. The rest of the categories show no significant effects, meaning that this result essentially replicates the lack of significant alignment to power found by [12].

Figure 6 shows the alignment differences based on power assessed by the follower count ratio, using our model-based measure. The intervals in this plot are 95% highest posterior densities based on samples from the posterior distribution. Here we find five marker categories showing significantly more alignment to powerful users, with only one (first person plural pronouns again) showing significantly less. If we exclude the pronouns, half of our markers show significantly greater alignment to power. Thus it appears that Twitter users do show a similar increase in alignment to power as had been seen in other settings.

7.3.2 *Verification as power*

We next examine tweet pairs with a mismatch in verification to look at alignment to externally-derived power. Because there are relatively few verified users, there were few message pairs with a verified replier in our dataset, and especially few tweets from verified users to other verified users. We thus limit our analysis to how unverified repliers adjust their alignment depending on the verification status of the user they are replying to. We predict that tweets sent from an unverified user to a verified user will show greater alignment than tweets sent from an unverified user to a fellow unverified user.

Figure 7 shows the mean difference in per-category alignments when unverified users reply to verified versus unverified users, according to the SCP measure. A positive value indicates increased alignment to the verified, powerful tweeters. However, none of the marker categories show significant increases in alignment to verified users, based on the 95% confidence intervals from a 1000-sample bootstrap. This replicates the result in [12], using their measure.

[6]We repeated these analyses on a range of cutoffs, down to 10 times as many followers, but did not see substantially different results.

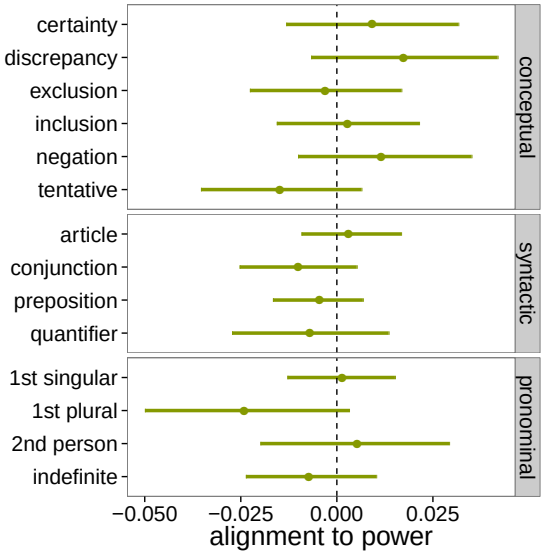

Figure 7: Difference in SCP-estimated alignment on the 14 marker categories when speaking to verified or unverified users. Positive values indicate greater alignment to verified users; negative indicates greater alignment to unverified users. No markers show significant effects of power on alignment, based on 95% bootstrap CIs.

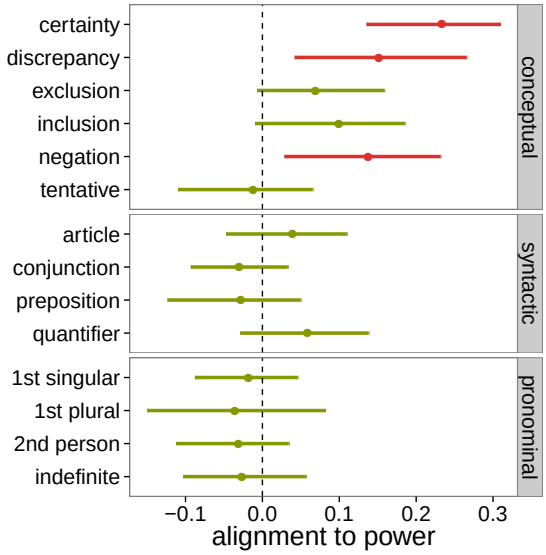

Figure 8: Difference in HAM-estimated alignment on the 14 marker categories when speaking to verified or unverified users. Positive values indicate greater alignment to verified users; negative indicates greater alignment to unverified users. Three markers show significantly more alignment to power, based on 95% HPD; none show significantly less.

In contrast, Figure 8 shows the difference between alignment to verified and unverified tweeters estimated with HAM; intervals are again the 95% highest posterior densities. Here we see positive differences, indicating higher alignment to verified tweeters, in three categories, with no significantly negative differences.

In these results, we see substantial evidence for a general increase in alignment to a user with more social power, even if that social power is extrinsic to the social network. Interestingly, we also find compelling evidence for the non-monolithic nature of linguistic alignment. All of the marker categories showed an overall positive alignment effect (Figure 4), but when we compare differences in how people align to with or without power, the markers show idiosyncratic effects. One class of markers, the pronouns, have previously been shown to interact with power [29], and these consistently show less alignment to power than most other markers. Another class of markers, conceptual markers, generally showed the greatest alignment to power. We speculate that these categories represent different framings of a topic under discussion, and that the choice of how to conceptualize a topic is largely made by the more powerful person in the conversation. Detecting such patterns is a key reason behind the marker separability desideratum.

8. CONCLUSIONS

Accommodation to one's conversational partners is a deeply ingrained human characteristic, as such is a useful tool for assessing the nature of conversation. Linguistic alignment is an important measure of accommodation, and has been a focus for much previous research. Despite this attention, measures of alignment have varied from study to study and field to field, leading to a large body of incommensurable results. In addition, as we show here, many widely-used measures either fail to distinguish alignment from homophily, or else suffer from bias across different baseline probabilities and different levels of accommodation.

We assessed measures on four desiderata: conditionality/baselining, marker separability, consistency across marker frequencies, and robustness to sparse data. We also introduced a hierarchical, model-based alignment measure (HAM) and showed that it outperformed previous measures in simulations. In an analysis of Twitter data, we also showed that this measure is able to detect differential alignment based on differences in social status that were undetectable using previous measures.

Making theoretical progress on the psychological mechanisms underlying linguistic alignment, as well as communication accommodation more generally, will require a consistent and robust set of empirical measurements. We believe our work here takes a step towards developing the kind of method that will facilitate this kind of consistency. Our measures are relatively straightforward to fit, and they result in measurements with a natural scale that can be compared across different settings (i.e., log-odds change). We hope that future work will adopt this probabilistic standard, facilitating a more coherent body of investigations of this set of phenomena.

9. REFERENCES

[1] F. Bilous and R. Krauss. Dominance and accommodation in the conversational behaviours of same-and mixed-gender dyads. *Language & Communication*, 1988.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, mar 2003.

[3] R. Y. Bourhis and H. Giles. The language of intergroup distinctiveness. In H. Giles, editor, *Language, ethnicity, and intergroup relations*, pages 119–135. Academic Press, London, 1977.

[4] R. L. Boyd and J. W. Pennebaker. Did shakespeare write double falsehood? identifying individuals by creating psychological signatures with text analysis. *Psychological science*, page 0956797614566658, 2015.

[5] H. Branigan, M. Pickering, J. Pearson, and J. McLean. Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355–2368, 2010.

[6] B. Burleson and D. Fennelly. The effects of persuasive appeal form and cognitive complexity on children's sharing behavior. *Child Study Journal*, 11:75–90, 1981.

[7] D. Byrne. Attitudes and attraction. *Advances in Experimental Social Psychology*, 4:35–89, 1969.

[8] B. Carpenter. Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 2015.

[9] T. Chartrand and R. B. van Baaren. Human Mimicry. *Advances in Experimental Social Psychology*, 41:219–274, 2009.

[10] C. Chung and J. Pennebaker. The psychological functions of function words. In K. Fiedler, editor, *Social communication*, chapter 12, pages 343–359. Psychology Press, New York, 2007.

[11] W. Condon and W. Ogston. A segmentation of behavior. *Journal of psychiatric research*, 1967.

[12] C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web - WWW '11*, page 745, New York, New York, USA, mar 2011. ACM Press.

[13] C. Danescu-Niculescu-Mizil and L. Lee. Chameleons in imagined conversations: a new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87. Association for Computational Linguistics, jun 2011.

[14] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg. Echoes of power: Language effects and power differences in social interaction. *Proceedings of the 21st international conference on World Wide Web - WWW '12*, page 699, 2012.

[15] G. Doyle and M. C. Frank. Audience size and contextual effects on information density in Twitter conversations. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 2015.

[16] K. Ferrara. Accommodation in therapy. In H. Giles, J. Coupland, and N. Coupland, editors, *Contexts of accommodation: developments in applied sociolinguistics.* Cambridge University Press, Cambridge, 1991.

[17] R. Fusaroli, B. Bahrami, K. Olsen, A. Roepstorff, G. Rees, C. Frith, and K. Tylen. Coming to Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological Science*, 23(8):931–939, 2012.

[18] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

[19] A. Gelman, A. Jakulin, M. Pittau, and Y. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2008.

[20] H. Giles, N. Coupland, and J. Coupland. Accommodation theory: Communication, context, and consequences. In H. Giles, J. Coupland, and N. Coupland, editors, *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press, Cambridge, 1991.

[21] H. Giles, K. R. Scherer, and D. M. Taylor. Speech markers in social interaction. In K. R. Scherer and H. Giles, editors, *Social markers in speech*, pages 343–81. Cambridge University Press, Cambridge, 1979.

[22] H. Giles and P. Smith. Accommodation theory: Optimal levels of convergence. In H. Giles and R. St. Clair, editors, *Language and Social Psychology*, pages 45–65. Blackwell, Oxford, 1979.

[23] A. Gnisci. Sequential strategies of accommodation: A new method in courtroom. *British Journal of Social Psychology*, 44(4):621–643, 2005.

[24] A. L. Gonzales, J. T. Hancock, and J. W. Pennebaker. Language Style Matching as a Predictor of Social Dynamics in Small Groups. *Communication Research*, 37(1):3–19, 2010.

[25] F. Guo, C. Blundell, H. Wallach, K. Heller, and U. Gatsby Unit. The bayesian echo chamber: Modeling social influence via linguistic accommodation. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2015.

[26] J. Hale and J. Burgoon. Models of reactions to changes in nonverbal immediacy. *Journal of Nonverbal Behavior*, 1984.

[27] M. E. Ireland, R. B. Slatcher, P. W. Eastwick, L. E. Scissors, E. J. Finkel, and J. W. Pennebaker. Language style matching predicts relationship initiation and stability. *Psychological Science*, 22:39–44, 2011.

[28] S. Jones, R. Cotterill, N. Dewdney, K. Muir, and A. Joinson. Finding Zelig in text: A measure for normalising linguistic accommodation, aug 2014.

[29] E. Kacewicz, J. W. Pennebaker, M. Davis, M. Jeon, and C. Arthur. Pronoun use reflects standings in social hierarchies. *Journal of Language . . .*, 33(2):125–143, 2013.

[30] S. Kline and J. Ceropski. Person-centered communication in medical practice. In J. T. Wood and G. M. Phillips, editors, *Human Decision-Making*, pages 120–141. SIU Press, Carbondale, 1984.

[31] H. Kucera and W. N. Francis. *Computational Analysis of Present-Day {A}merican {E}nglish*. Brown University Press, Providence, RI, 1967.

[32] W. Levelt and S. Kelter. Surface form and memory in question answering. *Cognitive psychology*, 1982.

[33] P. Lieberman. Intonation, perception, and language. *MIT Research Monograph*, 1967.

[34] C. Nass and K. Lee. Does computer-generated speech manifest personality? An experimental test of similarity-attraction. *Proceedings of the SIGCHI conference on Human ...*, 2000.

[35] M. Natale. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32(5):790–804, 1975.

[36] K. Niederhoffer and J. Pennebaker. Linguistic style matching in social interaction. *Journal of Language and Social ...*, 2002.

[37] B. Noble and R. Fernández. Centre Stage: How Social Network Position Shapes Linguistic Coordination. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 2015.

[38] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. Smith. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters, 2013.

[39] J. Pennebaker, R. Booth, and M. Francis. Linguistic inquiry and word count: LIWC. *Austin, TX: liwc. net*, 2007.

[40] J. Pennebaker and L. King. Linguistic styles: language use as an individual difference. *Journal of personality and social ...*, 1999.

[41] M. J. Pickering and S. Garrod. Toward a mechanistic psychology of dialogue. *The Behavioral and brain sciences*, 27(2):169–190; discussion 190–226, 2004.

[42] J. Thakerar, H. Giles, and J. Cheshire. Psychological and linguistic parameters of speech accommodation theory. *Advances in the social psychology of language*, 1982.

[43] H. C. Triandis. Cognitive similarity and communication in a dyad. *Human Relations*, 13:175–183, 1960.

[44] R. B. van Baaren, R. W. Holland, B. Steenaert, and A. van Knippenberg. Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental ...*, 39(4):393–398, 2003.

[45] Y. Wang, D. Reitter, and J. Yen. Linguistic Adaptation in Conversation Threads : Analyzing Alignment in Online Health Communities. *Acl2014*, 2014.

[46] M. Willemyns, C. Gallois, V. Callan, and J. Pittam. Accent accommodation in the employment interview. *Journal of Language and Social Psychology*, 15(1):3–22, 1997.

[47] Y. Xu and D. Reitter. An Evaluation and Comparison of Linguistic Alignment Measures. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 2015.