Investigating the Sources of Linguistic Alignment in Conversation

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

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

Abstract

In conversation, speakers tend to "accommodate" or "align" to their partners, changing the style and substance of their communications to be more similar to their partners' utterances. We focus here on "linguistic alignment," changes in word choice based on others' choices. Although linguistic alignment is observed across many different contexts and its degree correlates with important social factors such as power and likability, its sources are still uncertain. We build on a recent probabilistic model of alignment, using it to separate out alignment attributable to words versus word categories. We model alignment in two contexts: telephone conversations and microblog replies. Our results show evidence of alignment, but it is primarily lexical rather than categorical. Furthermore, we find that discourse acts modulate alignment substantially. This evidence supports the view that alignment is shaped by strategic communicative processes related to the ongoing discourse.

1 Introduction

In conversation, people tend to adapt to one another across a broad range of behaviors. This adaptation behavior is collectively known as "communication accommodation" (Giles et al., 1991). Linguistic alignment, the use of similar words to a conversational partner, is one prominent form of accommodation. Alignment is found robustly across many settings, including in-person, computer-mediated, and web-based conversation (Danescu-Niculescu-Mizil et al., 2012; Giles et al., 1979; Niederhoffer and Pennebaker, 2002). In addition, the strength of alignment to conversational partners varies with relevant sociological factors, such as the power of the partners, their social network centrality, and their likability. Potentially, this alignment could be used to infer these factors in situations where they are difficult to observe directly.

Although linguistic alignment appears to reflect important social dynamics, the mechanisms underlying alignment are still not well-understood. One particular question is whether alignment is supported by relatively automatic priming mechanisms, or higher-level, discourse and communicative strategies. The Interactive Alignment Model proposes that conversational partners prime each other, causing alignment via the primed reuse of structures ranging from individual lexical items to syntactic abstractions (Pickering and Garrod, 2004). In contrast, Accommodation Theory emphasizes the relatively more communicative and strategic nature of alignment (Giles et al., 1991).

Relative to this theoretical landscape, a number of questions have emerged. First, does alignment occur at structural levels? If alignment is driven by interactive priming of structures, effects of alignment should be expected not only at the lexical level but also for structural elements or categories as well. In contrast, if alignment is primarily communicative, then alignment strength might differ and be greater for specific words that serve particular conversational or discourse functions in a particular situation.

Second, does alignment vary with conversational goals? If alignment is driven primarily by priming, it should be relatively consistent across different aspects of a discourse. In contrast, from a strategic or communicative perspective, alignment – in which preceding words and concepts are reused – must be balanced against a need to move the conversation forward by introducing new words and concepts. Thus, on a communica-

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 526–536,
Berlin, Germany, August 7-12, 2016.

tive account, alignment should be modulated by the speaker's discourse act, reflecting whether the balance of the concern is convergence on a current focus or conveyal of new information.

Our goal in the current work is to investigate these questions. We make use of a recent probabilistic model of linguistic alignment, modifying it to operate robustly over corpora with highly varying distributional structures and to consider both lexical and category-based alignment. We use two corpora of spontaneous conversations, the Switchboard Corpus and a corpus of Twitter conversations, to perform two experiments. First, in both datasets we measure alignment across different levels of representation and find very limited evidence for category-level alignment. Second, we make use of annotations in Switchboard to measure alignment across different discourse acts, finding that the level of alignment depends on the discourse actions that are included in the analysis. Taken together, these findings are consistent with the idea that alignment arises from discourse-level, strategic processes that operate primarily over lexical items.

2 Previous Work

2.1 Why does alignment matter?

Linguistic alignment, like other kinds of accommodation, can be a critical part of achieving social goals. Performance in cooperative decision-making tasks is positively related to the participants' linguistic convergence (Fusaroli et al., 2012; Kacewicz et al., 2013). Romantically, match-making in speed dating and stability in established relationships have both been linked to increased alignment (Ireland et al., 2011). Alignment can also improve perceived persuasiveness, encouraging listeners to follow good health practices (Kline and Ceropski, 1984) or to leave larger tips (van Baaren et al., 2003).

Alignment is also important as an indicator of implicit sociological variables. Less powerful conversants generally accommodate to more to powerful conversants. Prominent examples include interviews and jury trials (Willemyns et al., 1997; Gnisci, 2005; Danescu-Niculescu-Mizil et al., 2012). A similar effect is found for network structure: speakers align more to more network-central speakers (Noble and Fernández, 2015). Additionally, factors such as gender, likability, respect, and attraction all interact with the magnitude of accommodation (Bilous and Krauss, 1988; Natale, 1975).

2.2 Sources of linguistic alignment

Despite the important outcomes associated with alignment, its sources are not clear. The most prominent strand of work on alignment has focused on the level of word categories, looking at how interlocutors change their frequency of using, for instance, pronouns or quantitative words (Danescu-Niculescu-Mizil et al., 2012; Ireland et al., 2011). These results show alignment effects at the category level, but it is in principle possible that these effects arose purely from alignment on individual words (and that conclusion would not be inconsistent with the interpretation of that work).

Syntactic alignment is one area in which theoretical predictions have been tested, though results have been somewhat equivocal. The Interactive Alignment Model has generally been taken to suggest that there should be cross-person priming of syntactic categories and structures (Pickering and Garrod, 2004). But while some studies have found support for syntactic priming (Gries, 2005; Dubey et al., 2005), others have found negative or null alignment (Healey et al., 2014; Reitter et al., 2006). In one particularly thorough study, Healey et al. (2014) found across two corpora that speakers syntactically *diverged* from their interlocutors once lexical alignment was accounted for.

Furthermore, positive alignment is generally regarded as a good conversational tactic, but there is clearly a limit to its virtues, at least when it comes to content words. Alignment is inherently backward-looking, while the general goal of a conversation is to exchange information that is not already known by both parties, an inherently forward-looking goal. Perhaps because of this, some recent work finding positive alignment has limited itself to "non-topical" word categories, which are less contentful (Danescu-Niculescu-Mizil et al., 2011; Doyle et al., 2016). And suggestively, alignment within a task-relevant syntactic category was a better predictor of decision-making performance than overall lexical alignment (Fusaroli et al., 2012).

In sum, although individual studies do bear on the sources of alignment, the picture is still not clear. Because most work on alignment has been done either on categories of words or aggregating

across the lexicon, we do not have a good sense of whether there are systematic differences in alignment at different levels of representation. A further complication is that there is no standard measure of alignment; we turn to this issue next.

2.3 Measures of alignment

The metrics used in previous work fall into two basic categories: distributional and conditional. Distributional methods such as Linguistic Style Matching (LSM) (Niederhoffer and Pennebaker, 2002; Ireland et al., 2011) or the Zelig Quotient (Jones et al., 2014) calculate the similarity between the conversation participants over their frequencies of word or word category use in all utterances within the conversation. In contrast, conditional metrics, such as Local Linguistic Alignment (LLA) (Fusaroli et al., 2012; Wang et al., 2014) and the metric used by Danescu-Nicolescu-Mizil et al. (2011), look at how a message conditions its reply, with alignment indicated by elevated word use in the reply when that word was in the preceding message.

While distributional methods have been popular, a major weakness of such methods is that they do not necessarily show true alignment, only similarity. A high level of distributional similarity does not imply that two conversational partners have aligned to one another, because they might instead have been similar to begin with. In contrast, conditional measures allow for stronger inferences about the temporal sequence of alignment (even though they cannot guarantee any causal interpretation). Thus, we focus here on conditional measures exclusively.

By-message conditional methods Several existing conditional methods have started from the simplified representation that messages either do or do not contain particular words ("markers"), irrespective of message length or marker count. (Danescu-Niculescu-Mizil et al., 2012; Doyle et al., 2016). We refer to these as "by-message" methods. Consider the following example of conditional alignment, using pronouns as the marker: Bob aligns to Alice if his replies are more likely to contain a pronoun when in response to a message from Alice that contains a pronoun.

	Bob's reply	
Alice's message	has pronoun	no pronoun
has pronoun	8	2
no pronoun	5	5

Here, Alice sends 10 messages that contain at least one pronoun, and 8 of Bob's replies contain at least one pronoun. But Alice also sends 10 messages that don't contain any pronouns, and only 5 of Bob's replies to these contain pronouns. This increased likelihood of a pronoun-containing reply to a pronoun-containing message is the conditional alignment.

Different models quantify this conditional alignment slightly differently. Danescu-Niculescu-Mizil et al. (2011) proposed a subtractive conditional probability model, where alignment is the difference between the likelihood of a pronoun-containing reply B to a pronoun-containing message A and the probability of a pronoun-containing reply to any message:

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

Doyle et al. (2016) showed that this measure can be affected by the overall frequency of the category being aligned on, though. To correct this issue, they proposed a Hierarchical Alignment Model (HAM), which defines alignment as a linear effect on the log-odds of a reply containing the relevant marker (e.g., a pronoun), similar to a linear predictor in a logistic regression.[1]

$$align_{HAM} \approx logit^{-1}(p(B|A)) - logit^{-1}(p(B|\neg A)) \quad (2)$$

These binary conditional methods depend on the assumption that all messages have similar, and small, numbers of words, however. The probability that a message contains at least one of any marker of interest is dependent on the message's length, so if messages vary substantially in their length, these alignment values can be at least noisy, if not biased. They are also not robust as messages increase in length, since the likelihood that a message contains any marker approaches 1 as message length increases.

By-word conditional methods A solution to the problem of variable message lengths is simply to shift from binarized data to count data. Instead of counting how many times Bob's replies contain at least one pronoun, we can count what proportion of his replies' word tokens are pronouns.

[1] Because the HAM estimated this quantity via Bayesian inference, the inferred alignment value depends on the prior and number of messages observed, so unlike the other measures, this equality is only approximate.

Some existing measures use a related quantity, the proportion of the preceding message that appears in its reply, to estimate alignment, notably Local Linguistic Alignment (LLA) (Fusaroli et al., 2012; Wang et al., 2014) and the lexical similarity (LS) measure of Healey et al. (2014). LLA is defined as the number of word tokens (w_i) that appear in both the message (M_a) and the reply (M_b), divided by the product of the total number of word tokens in the message and reply:

$$align_{LLA} = \frac{\sum_{w_i \in M_b} \delta(w_i \in M_a)}{length(M_a) length(M_b)} \quad (3)$$

These measures have an aspect of conditionality, as they only count words that appear in both the message and the reply. But they nevertheless fail to control for the baseline frequency of the initial marker, and hence may be biased in measurements across words or categories of different frequencies (Doyle et al., 2016). They also can be affected by reply length, as the maximum alignment estimate is only possible when the reply is shorter than the message.

All of these by-word conditional models treat the reply as a bag of words, without order information. The by-word models, including the WHAM model we propose, are agnostic about reply length effects, correcting for the artifactual length effects of by-message models, but assuming that all messages have similar alignment strengths independent of length. This is in contrast to models that explicitly model priming effects as decaying over time (Reitter et al., 2006; Reitter, 2008), which predict higher alignment in shorter replies. Future by-word alignment models could infer a discounting for words that occur later in the reply, similar to the beta value on the log-distance from the prime proposed in Reitter et al. (2006).

Our goal in this work is to create a model that combines the benefits of the existing by-message conditional models with the length-robustness of a by-word conditional method. We present WHAM, a modification of the HAM model that satisfies this goal.

3 The Word-Based Hierarchical Alignment Model (WHAM)

We propose the Word-Based Hierarchical Alignment Model (WHAM). Like HAM, WHAM assumes that word use in replies is shaped by whether the preceding message contained the marker of interest. But WHAM uses marker token frequencies within replies, so that a 40-word reply with two instances of the marker is represented differently from a 3-word reply containing one instance.

For each marker, WHAM treats each reply as a series of token-by-token independent draws from a binomial distribution. The binomial probability μ is dependent on whether the preceding message did (μ^{align}) or did not (μ^{base}) contain the marker, and the inferred alignment value is the difference between these probabilities in log-odds space (η^{align}). The graphical model is shown in Figure 1.

For a set of message-reply pairs between a speaker-replier dyad (a, b), we first separate the replies into two sets based on whether the preceding message contained the marker m (the "alignment" set) or not (the "baseline" set). All replies within a set are then aggregated in a single bag-of-words representation, with marker token counts $C^{align}_{m,a,b}$ and $C^{base}_{m,a,b}$, and total token counts $N^{base}_{m,a,b}$ and $N^{base}_{m,a,b}$, the observed variables on the far right of the model. Moving from right to left, these counts are assumed to come from binomial draws with probability $\mu^{align}_{m,a,b}$ or $\mu^{base}_{m,a,b}$. The μ values are generated from η values in log-odds space by an inverse-logit transform, similar to linear predictors in logistic regression.

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 then the change in log-odds of the replier using m above baseline usage, given that the initial message uses m.

The remainder of the model is a hierarchy of normal distributions that allow social and word category structure to be integrated into the analysis. In the present work, we have three levels in the hierarchy: category level, marker level,[2] and conversational dyad level. All of these normal distributions have identical standard deviations $\sigma^2 = .25$.[3] A $Cauchy(0, 2.5)$ distribution

[2] In the lexical and category-not-word alignment models, these markers are words within a category. The category alignment model does not include this level, since all words in a category are treated identically.

[3] This value was chosen as a good balance between rea-

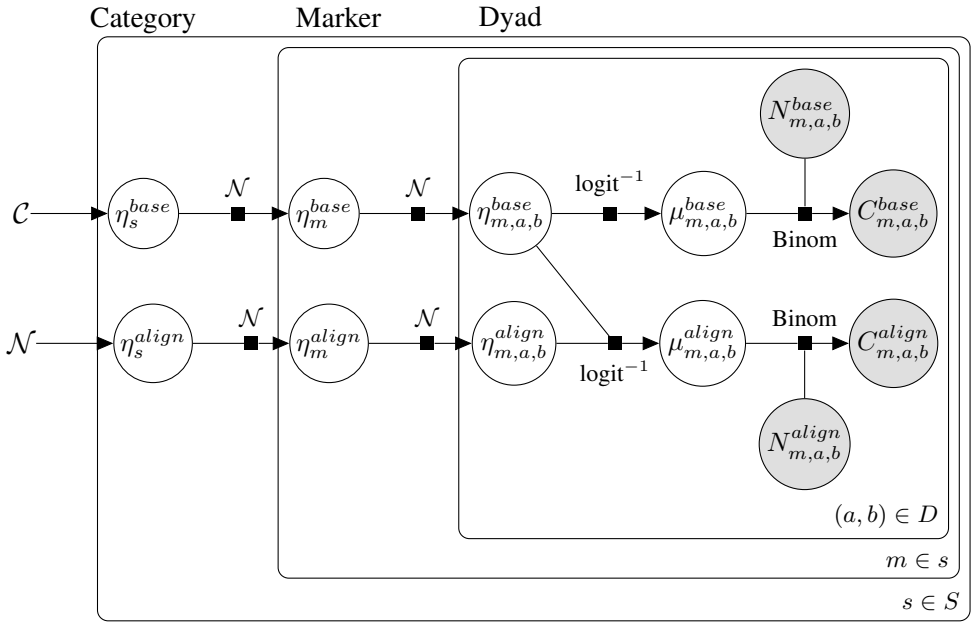

Figure 1: The Word-Based Hierarchical Alignment Model (WHAM). A chain of normal distributions generates a linear predictor η, which is converted into a probability μ for binomial draws of the words in each reply.

gives a relatively uninformative prior for the baseline marker frequency (Gelman et al., 2008). The alignment hierarchy is headed by a normal distribution centered at 0, biasing the model equally in favor of positive and negative alignments.

For our marker set, we adopt the Linguistic Inquiry and Word Count (LIWC) system to categorize words (Pennebaker et al., 2007). We use a set of 11 categories that have shown alignment effects in previous work (Danescu-Niculescu-Mizil et al., 2011). These can be loosely grouped into a set of five syntactic categories (articles, conjunctions, prepositions, pronouns, and quantifiers) and six conceptual categories (certainty, discrepancy, exclusion, inclusion, negation, and tentative). Categories and example elements are shown in Table 1. We manually lemmatized all words in each category. We implemented WHAM in RStan (Carpenter, 2015), with code available at `http://github.com/langcog/disc_align`.

Category	Examples	Size	Swbd Prob	Twit Prob
Article	*a, the*	2	.053	.047
Certainty	*always, never*	17	.014	.015
Conjunction	*but, and, though*	18	.077	.051
Discrepancy	*should, would*	21	.015	.019
Exclusive	*without, exclude*	77	.038	.028
Inclusive	*with, include*	57	.057	.028
Negation	*not, never*	12	.020	.023
Preposition	*to, in, by, from*	97	.097	.091
Pronoun	*it, you*	55	.17	.16
Quantifier	*few, many*	23	.028	.025
Tentative	*maybe, perhaps*	28	.033	.025

Table 1: Marker categories for linguistic alignment, with examples, number of distinct word lemmas, and token probability of in a reply in Switchboard and Twitter.

3.1 Validating WHAM

A major goal of our by-word alignment model, WHAM, is to fix the length issues discussed in Section 2.3. We test WHAM and the by-message HAM model on simulated data, using a method similar to Simulation 2 in Doyle et al. (2016), to see how robust they are to different reply lengths. We generate 500 speaker-replier dyads, each exchanging an average of 5 message pairs (drawn from a geometric distribution). Each message pair consists of a message whose length in words is drawn from a uniform distribution $[1, 25]$, and a reply of length L. Because our goal is to test the effect of length on the models' performances, we create separate simulated datasets for different values of L, and see whether the model correctly estimates the alignment value η^{align}. Three independent simulations were run for each alignment-length pair. We present data here for a simulated

sonable parameter convergence (improved by smaller σ^2) and good model log-probability (improved by larger σ^2).

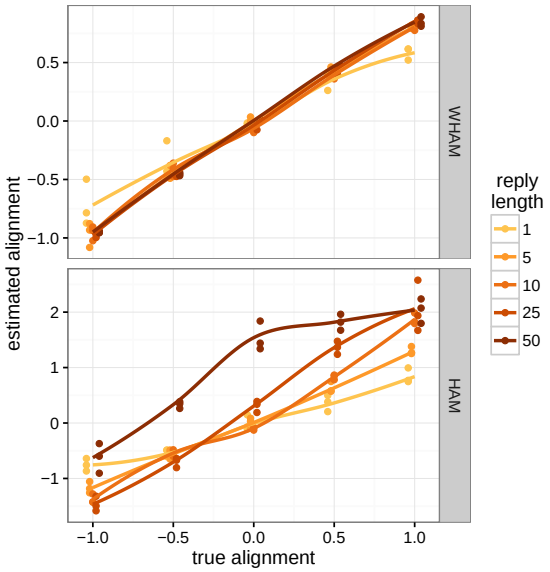

Figure 2: Actual versus estimated alignment on simulated data. Lines are loess-fit curves; colors represent the reply length in the simulation run. WHAM estimates alignment accurately regardless of reply length; HAM is highly affected by length.

word category with a baseline frequency of 0.1, around the middle of the attested category frequency range (see Table 1).

Figure 2 plots the true alignment value in the simulations against the model-estimated alignment values. Different colors represent different reply lengths L, ranging from single-word replies (light yellow) to 50-word replies (dark orange). The WHAM model shows consistently accurate alignment estimates over the range of simulated alignment values and reply lengths. The HAM model estimates the alignment far less accurately, and the reply length biases its estimates.

4 Data

Moving on to real data, we use two corpora for our experiments. The first is a collection of Twitter conversations collected by Doyle & Frank (2015) to examine information density in conversation. This corpus focuses on conversations within a set of 14 mostly distinct sub-communities on Twitter, and 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 in reply to the first (according to the Twitter API), although this does not necessarily mean that the content of the reply is a response to the preceding tweet. Retweets (including explicit retweets and some common manual retweet methods) were removed automatically. This processing leaves us with 122,693 message pairs, spanning 2,815 users. The tweets were parsed into word tokens using the Twokenizer (Owoputi et al., 2013).

The second corpus is the SwDA version of the Switchboard corpus (Godfrey et al., 1992; Jurafsky et al., 1997).[4] This corpus is a collection of transcribed telephone conversations, with each utterance labeled with the discourse act it is performing (e.g., statement of opinion, signal of non-understanding). It contains 221,616 total utterances in 1,155 conversations. We combine consecutive utterances by the same speaker without interruption from the listener into a single message and treat consecutive pairs of messages from different speakers as conversation turns, resulting in 110,615 message pairs.

5 Experiment 1: Lexical- and Category-Level Alignment

Our first experiment examines how alignment differs across the lexical and categorical levels. We use the WHAM framework to infer alignment on word and category counts, and also introduce a measure to estimate the influence of one word in a category on other words in its category, "category-not-word" alignment. We include this last type of alignment because it is possible that the category alignment effects in previous work are the result of lexical alignment on the individual words in the category, without any influence across words in the category. If categorical alignment is a real effect over and above lexical alignment, as an interactive-priming source for alignment would suggest, then the presence of a word in a message should not only increase the chance of seeing that word in the reply, but also other words in its category.

5.1 Category-not-word-alignment model

Assessing the amount of alignment triggered across words in a category (which we call "category-not-word alignment" or CNW) is not trivial, as there are a variety of interactions between lexical items within a category that can cause the lexical alignment to actually be less than

[4] Available courtesy of Christopher Potts at `http://compprag.christopherpotts.net/swda.html`.

	Reply		
Message	$\emptyset$	*he*	*she*
$\emptyset$	25	25	25
he	20	50	10
she	20	10	50

Table 2: A theoretical case where lexical alignment surpasses categorical alignment due to negative CNW between the words.

the category alignment. Table 2 illustrates this with a theoretical distribution over the pronouns *he* and *she*; one use of the pronoun *he* makes another use more likely (*A: Did he like the movie? B: Yeah, he loved it.*) while also reducing the likelihood of *she*, since the topic of conversation is now a male, and vice versa for *she*. For both *he* and *she*, the lexical alignment is approximately $logit^{-1}(p(B|A) - p(B|\neg A)) = logit^{-1}(\frac{50}{80} - \frac{25}{75}) \approx 1.2$, but categorical alignment is approximately $logit^{-1}(\frac{120}{160} - \frac{50}{75}) \approx 0.4$. On the other hand, the pronouns *you* and *I* might trigger each other more than themselves (*A: Did you like the movie? B: Yeah, I loved it.*).

The differences between lexical, categorical, and CNW alignment are also relevant to discussions of "lexical boosts" in the syntactic priming literature, an increased priming effect at the categorical level when there is lexical repetition. Lexicalist residual activation accounts (Pickering and Branigan, 1998) predict such a boost, while implicit learning accounts do not (Bock and Griffin, 2000; Chang et al., 2006). In the context of this experiment, such a lexical boost could make lexical and categorical alignment appear elevated and closer together, but would not have a substantial effect on CNW alignment.[5]

To investigate CNW alignment, we look at a subset of the data: for each word w, exclude all messages that contain a word from that category (S) that is not w. This limits the category alignment influence on the reply to the single word w. Then, instead of looking at how often w appears in the reply, we look at how often *all other words* in category S appear in the reply. The model then infers the influence of w on the other words in the category independent of their lexical alignment. Within the WHAM model, we change the count variables $C^{\cdot}$ and $N^{\cdot}$ so that C^{align} is the number of tokens of $\{S - w\}$ in replies to messages containing w but not $\{S - w\}$. C^{base} is then the number in replies to messages not containing any words in S. Similarly, N^{align} is the total token counts over replies containing w but not any other words in S, and N^{base} the total token counts over replies containing no words in S.

5.2 Methods

We conducted three sets of simulations, fitting the model with marker categories, individual words, and with the CNW scheme described above. In each, the model was fit with two chains of 200 iterations of the sampler for each dataset. We then extracted alignment estimates from each of the final 100 samples, and we report 95% highest posterior density intervals on η_S^{align}.

5.3 Results

Figure 3 shows the alignment on each marker category in the Twitter and Switchboard corpora. There were substantial differences in the overall rate of alignment between the corpora: Mean category alignment on Twitter was .19, while Switchboard category alignment was $-.051$. These differences may reflect the nature of the two discourse contexts: Replies on Twitter are composed while looking at the preceding message, encouraging the replier to take more account of the other tweeter's words, and a replier can draft and edit their reply to make it better fit the conversation. Messages on Switchboard, on the other hand, are evanescent, so a replier must compose a reply without looking back at the message, without editing, and in real-time. Differences in the discourse structure of these corpora may also be contributing, an effect we will consider in Experiment 2.

Despite the difference in reply construction in the two corpora, the results across levels of alignment were similar. Alignment was found primarily at the lexical – rather than the category – level. Lexical and category alignment were not significantly different from each other, but the strength of lexical alignment was significantly larger than the CNW alignment, according to a t-test over categories (Twitter: $t(10) = .21, p < .001$; Swbd: $t(10) = .12, p = .003$). CNW alignment was significantly negative on Switchboard ($t(10) = -.11, p = .01$) and not significantly different from zero on Twitter ($t(10) = .009, p = .79$).

[5]The categories being investigated in our work contain mostly non-topical, closed-class words, which have not exhibited lexical boosts in past research (Bock, 1989; Pickering and Branigan, 1998; Hartsuiker et al., 2008), but such boosting may be detectable in estimates on topical categories.

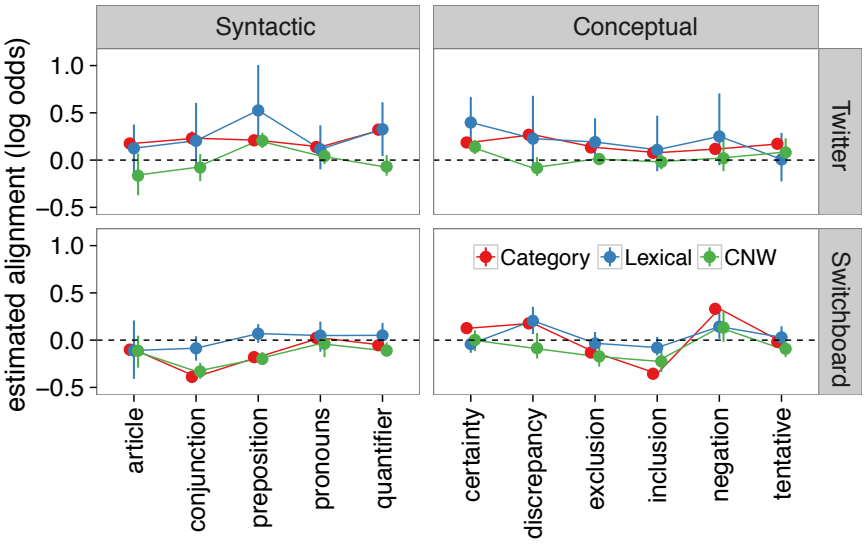

Figure 3: Categorical (red), lexical (blue), and CNW (green) alignments plotted by category, on the Twitter (left) and Switchboard (right) datasets. 95% HPD intervals from WHAM shown.

WHAM – unlike other previous measures – provides estimates of alignment that are unbiased by either marker frequency or message length, but we still observed modest alignment on Twitter, replicating previous work (Doyle et al., 2016; Danescu-Niculescu-Mizil et al., 2011). Alignment was smaller in Switchboard, and in both cases there were no category effects. Thus, the categorical alignment results may result primarily from lexical alignment, inconsistent with the predictions of interactive priming accounts of alignment.

6 Experiment 2: Discourse Acts and Alignment

Messages within a discourse can serve a very wide range of purposes. This variety has effects for both linguistic structure and the relationship to neighboring messages. For example, a simple yes/no question is likely to receive a short, constrained reply, while a statement of an opinion is more likely to yield a longer reply. In addition, different types of messages can either introduce new information to the conversation (e.g., statements, questions, offers) or look back at existing information (e.g., acknowledgments, reformulations, yes/no answers). We hypothesize that alignment will be substantially different depending on the discourse act, as speakers' conversational goals vary. Thus, our second experiment examines how alignment differs depending on discourse act.

We focus on a particular kind of discourse act, the backchannel (Yngve, 1970). Backchannels are extremely common in Switchboard, accounting for almost 20% of utterances, and include utterances such as single words signaling understanding or misunderstanding (*yeah, uh-huh, no*) or simple messages expressing empathy without trying to take a full conversational turn (*It must have been tough*). Backchannels are a particularly interesting case because their short and constrained nature makes it difficult to align on some categories (e.g., backchannels rarely contain quantifiers or prepositions), while the purpose of giving feedback to the speaker makes it important to align on others (e.g., matching the positive/negative tone or certainty of a speaker). In addition, backchannels are primarily restricted to spoken corpora. Twitter conversations contain far fewer backchannels than Switchboard, which may account for some of their alignment differences—especially as the results of this experiment suggest that backchannels reduce overall alignment.

6.1 Methods

We use the discourse-annotated Switchboard corpus to compare alignment in conversations containing backchannels with those whose backchannels have been removed. We make this comparison by creating a second corpus, removing every utterance classified as a backchannel from the corpus prior to parsing the utterances into conversation turns as before.

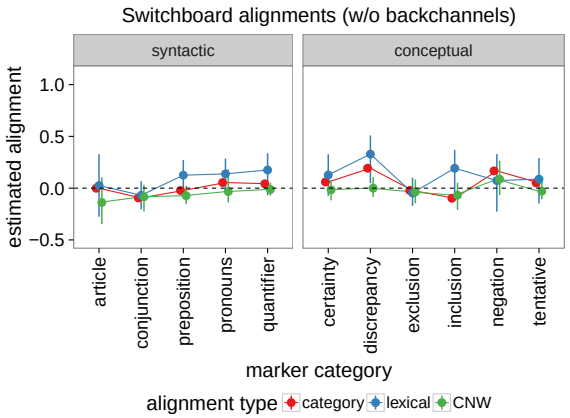

Figure 4: Categorical (red), lexical (blue), and CNW (green) alignments on the Switchboard dataset with backchannels removed. 95% HPD intervals from WHAM shown.

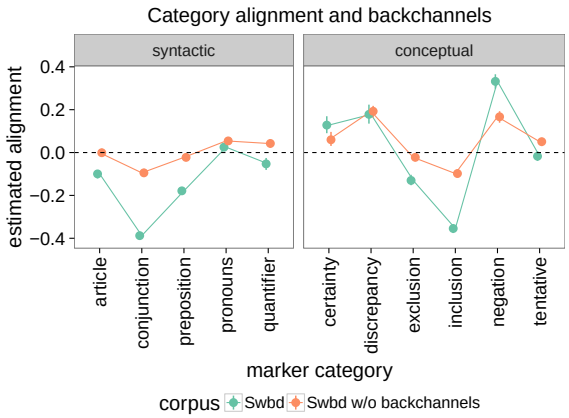

Figure 5: Comparing categorical alignment on the Switchboard dataset with and without backchannels. 95% HPD intervals from WHAM shown.

6.2 Results

Alignment values for the Switchboard corpus without backchannels are shown in Figure 4. As expected, alignment is on average higher without the backchannels ($p = .09$ for category, $p < .05$ for lexical and CNW), reflecting the constrained nature of backchannels. Lexical alignment is significantly higher than category alignment ($t(10) = -.08, p = .03$), consistent with the findings of Experiment 1. The mean category alignment without backchannels is .029.

Figure 5 compares the category alignments for the full Switchboard corpus (green) and Switchboard without backchannels (orange). Alignment on the full corpus is lower for all but two categories, exhibiting the reduced opportunity for alignment provided by backchannels. Syntactic category alignment is especially affected by backchannels, whose constrained forms provide very little ability to align syntactically.

Interestingly, the two categories that do show greater alignment when backchannels are included are certainty and negation. These categories are both important for backchannels; a negative backchannel is generally inappropriate in reply to a non-negative message, and similarly a confident backchannel would often be out of place in reply to an uncertain message. These influences of discourse acts on alignment are more consistent with a discourse-strategic origin for alignment than a priming-based account.

7 Discussion

Linguistic alignment is a prominent type of communicative accommodation, but its sources are unclear. We presented WHAM, a length-robust extension of a probabilistic alignment model. Using this model, we find evidence that linguistic alignment is primarily lexical, and that it is strongly affected by at least some aspects of the discourse goal of a message.

This combination of a primarily-lexical origin for linguistic alignment and its variation by word category and discourse act suggest that alignment is primarily a higher-level discourse strategy rather than a low-level priming-based mechanism. This set of results is consistent with both Accommodation Theory and the set of findings, reviewed above, that sociological factors affect the level of observed alignment. The effect of discourse acts on alignment further suggests that alignment is not a completely automatic process but rather one of many discourse strategies that speakers use to achieve their conversational goals.

Acknowledgments

We wish to thank Dan Yurovsky, Aaron Chuey, and Jake Prasad for their work on and discussion of earlier versions of the model, Herb Clark for discussions of potential effects of message length, and, of course, the reviewers. The authors were funded by NSF BCS 1528526, NSF BCS 1456077, and a grant from the Stanford Data Science Initiative.

References

Frances R. Bilous and Robert M. Krauss. 1988. Dominance and accommodation in the conversational behaviours of same-and mixed-gender dyads. *Language & Communication*.

Kay Bock and Zenzi M. Griffin. 2000. The persistence of structural priming: Transient activation or implicit learning. *Journal of Experimental Psychology: General*, 129:177–192.

Kay Bock. 1989. Closed-class immanence in sentence production. *Cognition*, 31:163–186.

Bob Carpenter. 2015. Stan: A Probabilistic Programming Language. *Journal of Statistical Software*.

Franklin Chang, Gary S. Dell, and Kay Bock. 2006. Becoming syntactic. *Psychological Review*, 113:234–272.

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 - WWW '11*, page 745, New York, New York, USA. ACM Press.

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 - WWW '12*, page 699.

Gabriel Doyle and Michael C. Frank. 2015. Audience size and contextual effects on information density in Twitter conversations. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

Gabriel Doyle, Dan Yurovsky, and Michael C. Frank. 2016. A robust framework for estimating linguistic alignment in Twitter conversations. In *WWW 2016*.

Amit Dubey, Patrick Sturt, and Frank Keller. 2005. Parallelism in coordination as an instance of syntactic priming: Evidence from corpus-based modeling. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 827–834. Association for Computational Linguistics.

Riccardo Fusaroli, Bahador Bahrami, Karsten Olsen, Andreas Roepstorff, Geraint Rees, Chris Frith, and Kristian Tylén. 2012. Coming to Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological Science*, 23(8):931–939.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*.

Howard Giles, Klaus R. Scherer, and Donald M. Taylor. 1979. Speech markers in social interaction. In Klaus R. Scherer and Howard Giles, editors, *Social markers in speech*, pages 343–81. Cambridge University Press, Cambridge.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequences. In Howard Giles, Justine Coupland, and Nikolas Coupland, editors, *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge University Press, Cambridge.

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

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing.*, volume 1, pages 517–520. IEEE.

Stefan Th Gries. 2005. Syntactic priming: A corpus-based approach. *Journal of psycholinguistic research*, 34(4):365–399.

Robert J. Hartsuiker, Sarah Bernolet, Sofie Schoonbaert, Sara Speybroeck, and Dieter Vanderelst. 2008. Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, 58:214–238.

Patrick G. T. Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PloS one*, 9(6):e98598.

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*, 22:39–44.

Simon Jones, Rachel Cotterill, Nigel Dewdney, Kate Muir, and Adam Joinson. 2014. Finding Zelig in text: A measure for normalising linguistic accommodation. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 455–465.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*, pages 97–102.

Ewa Kacewicz, James W. Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C. Graesser. 2013. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143.

Susan L. Kline and Janet M. Ceropski. 1984. Person-centered communication in medical practice. In *Human Decision-Making*, pages 120–141. SIU Press, Carbondale.

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

Kate G. Niederhoffer and James W. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Bill Noble and Raquel Fernández. 2015. Centre Stage: How Social Network Position Shapes Linguistic Coordination. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah Smith. 2013. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–391.

James W. Pennebaker, Roger J. Booth, and Martha E. Francis. 2007. Linguistic Inquiry and Word Count: LIWC.

Martin J. Pickering and H. P. Branigan. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, 39:633–651.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

David Reitter, Johanna D. Moore, and Frank Keller. 2006. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

David Reitter. 2008. *Context Effects in Language Production: Models of Syntactic Priming in Dialogue Corpora*. Ph.D. thesis, U. of Edinburgh.

Rick B. van Baaren, Rob W. Holland, Bregje Steenaert, and Ad van Knippenberg. 2003. Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental Social Psychology*, 39(4):393–398.

Yafei Wang, David Reitter, and John Yen. 2014. Linguistic Adaptation in Conversation Threads: Analyzing Alignment in Online Health Communities. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Michael Willemyns, Cynthia Gallois, Victor Callan, and Jeffrey Pittam. 1997. Accent accommodation in the employment interview. *Journal of Language and Social Psychology*, 15(1):3–22.

Victor Yngve. 1970. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistics Society*, pages 567–577.