Coordinating Chaos: A Structured Review of Linguistic Coordination Methodologies

Benjamin Litterer
School of Information
University of Michigan
blitt@umich.edu

David Jurgens
School of Information and
Computer Science & Engineering
University of Michigan
jurgens@umich.edu

Dallas Card
School of Information
University of Michigan
dalc@umich.edu

Abstract

Linguistic coordination—a phenomenon where conversation partners end up having similar patterns of language use—has been established across a variety of contexts and for multiple linguistic features. However, the study of language coordination has been accompanied by a diverse and inconsistently applied set of measures and theoretical perspectives. This diversity has significant consequences, as replication studies have highlighted the brittleness of certain measures and called influential findings into question. While prior work has addressed specific modeling decisions and model types, linguistic coordination research has yet to fully examine, synthesize, and critique the space of modeling choices available. In this work, we present a framework to organize the linguistic coordination literature. Using this schema, we provide a high-level overview of the choices involved in the measurement process and synthesize relevant critiques. Based on both gaps and limitations surfaced from this review, we suggest directions for further exploration and evaluation. In doing so, we provide the clarity required for linguistic coordination research to arrive at interpretable and sound conclusions.

1 Introduction

Linguistic coordination is a general term used to describe two or more interlocutors adapting their language to be more similar to one another.[1] This phenomenon has been widely studied, with evidence linking coordination to group cohesion (Gonzales et al., 2010), performance on joint tasks (Fusaroli et al., 2012), differences in power (Danescu-Niculescu-Mizil et al., 2012), and relationship stability (Ireland et al., 2011). Given the variety of coordination results across domains and its association with prosocial outcomes, this behavior appears to be important for successfully navigating interactions in daily life.

While there is consensus that coordination is worthy of investigation, there is no consensus on the best way to characterize it; to do so, the researcher must make a number of consequential choices to define, operationalize, and measure this phenomenon. For example, consider the work of Danescu-Niculescu-Mizil et al. (2012), who investigate the link between language coordination and power. To do so, they identify and quantify linguistic features called function words and feed them into a mathematical model called SUBTRACTIVE CONDITIONAL PROBABILITY (SCP) to capture a form of coordination called Linguistic Style Matching (Niederhoffer and Pennebaker, 2002). While this methodological pipeline may seem straightforward, there are many variations on this approach that can be found in the literature on this topic, and could have been employed instead. Their work uses a particular definition of coordination, a choice that might depend on factors such as disciplinary context and theoretical assumptions. This definition is then operationalized and estimated through a mathematical model. Here, additional choices arise related to the model input, estimation strategy, and refinement or validation. It is only after these choices have been made that Danescu-Niculescu-Mizil et al. (2012) conclude that coordination is positively related to power.

Given these degrees of freedom, it is unsurprising that evidence for coordination effects is still mixed. Null results have disputed the link between coordination and interaction quality (Niederhoffer and Pennebaker, 2002), leadership status (Huffaker et al., 2006), team performance (Heuer et al., 2020), and negotiation success (Ireland and Henderson,

[1] There are multiple terms that are often used to refer to versions of this phenomenon, including *coordination*, *accommodation*, *alignment*, and *entrainment*. In linguistics, "coordination" also refers to syntactic structures that link together multiple elements. Unfortunately there is no single term that is not overloaded. Here, we follow Danescu-Niculescu-Mizil (2012), Ben-Haim and Tsur (2021), and others in calling this "linguistic coordination", with reference to other terms as appropriate.

2014). Furthermore, replication studies adopting different modeling decisions have reversed prior findings linking coordination to power (Gao et al., 2015; Xu et al., 2018). In addition to contradictory results, a lack of methodological agreement makes it difficult to compare findings and disentangle measurement decisions from empirical results. Given the numerous schools of thought with little cross-over, there is no unified framework for understanding the methodological choices in play and their connection to theoretical assumptions.

In this work, we systematize linguistic coordination research to address the lack of agreement and replicability in this domain. To do so, we first outline prior work (§2), the scope of our review (§3), and relevant theoretical background (§4). We then introduce a framework used to disentangle the modeling and estimation choices available to researchers (§5), and synthesize critiques of the linguistic coordination literature (§6).

By systematizing the space of potential choices, we make three principal contributions. First, we provide the background necessary for scholars who are new to this area to successfully choose and apply a method to a corpus of interest. Second, we articulate key shortcomings of this literature that enable readers to critically evaluate research in the linguistic coordination domain. Finally, we suggest promising areas for future research based on gaps in the current literature and recent advances in NLP.

2 Prior Work

Xu and Reitter (2015) consider three measures of linguistic alignment and evaluate them across criteria such as sensitivity to alignment, normality of distribution, and consistency across lexical and syntactic alignment types. Arnet et al. (2024), Xu et al. (2018), and Gao et al. (2015) focus on linguistic style matching (LSM; see §4), providing critiques surrounding bias and confounding resulting from a failure to properly model conversation turn-length. Doyle et al. (2016) cut across the LSM and linguistic alignment spaces and provide simulation tests for four commonly used measures.

While informative, this prior work has typically provided brief reviews as background material in the context of introducing a new method. This has limited the discussion to a small number of modeling details rather than a high-level overview summarizing the choices available at each step of the measurement process. Furthermore, the critiques offered by this literature are disjointed and have yet to be fully synthesized. To address this narrow scope and lack of standardization, we integrate existing measures and critiques under a broad framework that encapsulates far more variation in modeling approach. Such an approach allows researchers to easily parse extant literature for the purposes of applying, critiquing, or extending entrainment methods.

3 Scope of Review

In order to produce a coherent, parsimonious, and self-contained review, we bound our space of interest significantly. First, we focus on papers that explicitly introduce coordination methods or critique existing methods, as our primary goal is to cover the diverse space of modeling choices and their implications. In the same vein, we select measures on the basis of their methodological innovation, even if they have not witnessed wide-scale adoption in a particular research domain. Finally, we focus our attention towards the field of natural language processing, conducting a thorough review of methods in this space through a comprehensive search of the Association for Computational Linguistics (ACL) anthology (see Appendix A for more details).

In addition to these inclusion criteria, we further bound our review by excluding certain research domains. We first exclude studies of coordination with respect to auditory features (e.g. pitch, intonation) and movements (e.g. gestures, posture, gaze), opting to focus solely on models using text as input. There has been extensive work regarding these modalities of coordination, requiring theoretical and pragmatic considerations beyond the scope of this review (Levitan and Hirschberg, 2011; Pardo, 2006; Mousset et al., 1996; Chartrand and Bargh, 1999). In addition, we exclude any measure that is not designed for language freely produced between two or more speakers. This criterion eliminates the large body of experimental work on syntactic priming employing techniques such as sentence completion tasks (Branigan et al., 1999) and picture description tasks (Gries, 2005). For a similar reason, studies focusing on human-agent interaction are also excluded.

4 Theoretical Background

Computational Research regarding linguistic coordination is rooted in theoretical frameworks that guide the hypotheses and assumptions of interest.

In this section, we give a brief overview of these frameworks in order to properly contextualize our focus on modeling choices.

4.1 Communication Accommodation Theory

Communication Accommodation Theory (CAT) seeks to explain how social factors influence the propensity of individuals to adapt (i.e. accommodate) their behaviors to one another (Giles et al., 1991). Originally formulated exclusively for speech (SAT), this theory grew to encompass accommodation with respect to other factors such as posture (Condon and Ogston, 1967), facial expression (Hale and Burgoon, 1984), and laughter (Bilous and Krauss, 1988). Accommodative behaviors are broadly thought to result from a desire for social approval, though factors such as cultural norms (White, 1989) and instrumental goals (van den Berg, 1985; Taylor et al., 1978; Danescu-Niculescu-Mizil et al., 2012) are also considered. Dis-accommodation, on the other hand, is said to result from identification with a desirable out-group (Giles et al., 1991), a desire which may be heightened in contexts where group identity is threatened (Bourhis, 1979).

Closely related to CAT is the phenomenon of Linguistic Style Matching (LSM) (Niederhoffer and Pennebaker, 2002). This conceptual framework adopts the framing of CAT (among other theories; Byrne, 1997), but provides a specific operationalization of coordination in language. In particular, LSM identifies style as the most relevant linguistic construct of interest and captures this construct using function words. This is largely motivated by research demonstrating that function word usage is largely subconscious and is associated with numerous psychological correlates (Chung and Pennebaker, 2007). If function word usage is coordinated between speakers at the turn or conversation level, LSM is said to occur. As part of our review, we further interrogate how the theoretical concept of LSM is measured in practice.

4.2 Interactive Alignment Model

The Interactive Alignment Model (IAM) was proposed by Pickering and Garrod (2004), and seeks to explain the mechanism(s) through which mutual understanding is developed in conversation. In particular, the IAM offers an automatic priming-based explanation for coordination in dialogue such that usage of a linguistic construct by one speaker increases the likelihood of this construct being used by the other speaker. Unlike LSM, the IAM considers many types of priming, including low and high-level constructs such as lexical, syntactic or semantic representations. Unlike CAT, this theory largely ignores social factors in favor of a more cognitive approach to coordination. This theory has been operationalized through a number of measures, and our review seeks to clarify how modeling choices impact what is actually being measured.

5 Categorizing Measures of Linguistic Coordination

From the literature, we identify 25 distinct methods for measuring linguistic coordination. To summarize these measures compactly and distinguish between key similarities and differences, we group these methods according to several criteria that align with important choices in the modeling pipeline. We now provide an explanation of each criterion and survey the space of possible modeling decisions, culminating in a handful of approaches that generalize from the many specific examples.

5.1 Measurement Target

Although there is broad agreement that coordination concerns similarity of language between speakers, more precise language is needed to develop a concrete measurement approach. To do so, we adopt the framework of Levitan and Hirschberg (2011), who introduce *proximity*, *convergence*, and *synchrony* as specific types of coordination.

Proximity occurs when speakers are close with respect to a specific linguistic feature over an entire conversation or series of turns. This could result from a decision or subconscious reflex to adapt one's language before the first utterance has occurred, perhaps due to social factors such as social distance, status, or power (Danescu-Niculescu-Mizil et al., 2012; Soliz et al., 2021). *Convergence* refers to a setting where proximity increases over a series of turns or an entire conversation. This can occur as a result of speakers coordinating their language, such as when developing a common lexicon to achieve a shared task (Pickering and Garrod, 2004; Fusaroli et al., 2012). Finally, *synchrony* is localized to the conversation-turn level, and relies not on the raw distance between speaker quantities but on their relative changes. If, as a general pattern, one speaker's increase in a feature relative to their baseline is associated with an increase in another speaker's feature relative to their baseline,

Name	Measurement Target	Gen. Framework	Estimator Type	Form	Input Features	Publication
LSM Correlation	Synch., Prox.	Metric	Model output	Sim.	fw count	Niederhoffer and Pennebaker (2002)
Word Mover's Dist.	Prox.	Metric	Model output	Sim.	neural embeddings	Nasir et al. (2019)
LSM Canberra$_1$	Prox.	Metric	Model output	Sim.	fw rate	Gonzales et al. (2010)
LSM Canberra$_2$	Prox.	Metric	Model output	Sim.	fw rate	Ireland and Pennebaker (2010)
LSM Canberra$_3$	Prox.	Metric	Model output	Sim.	fw rate	Müller-Frommeyer et al. (2019)
LSM Canberra$_4$	Prox.	Metric	Model output	Sim.	fw rate	Arnet et al. (2024)
LSM Canberra$_5$	Prox.	Metric	Model output	Sim.	fw rate	Müller-Frommeyer and Kauffeld (2022)
Ling. Accom.	Prox.	Metric	Model output	Sim.	fw rate	Jones et al. (2014)
Entrainment$_1$	Prox.	Metric	Model output	Sim.	high freq. word rate	Nenkova et al. (2008)
Entrainment$_2$	Prox.	Metric	Model output	Sim.	high freq. word count	Nenkova et al. (2008)
Embedding Sim$_1$	Prox., Synch., Conv.	Metric	Model output	Sim.	neural embeddings	Kejriwal and Beňuš (2023)
Embedding Sim$_2$	Prox., Conv.	Metric	Model output	Sim.	neural embeddings	Xu (2021)
Perplexity	Prox.	Metric	Model output	Sim	words	Weise and Levitan (2018)
KL Divergence	Prox.	Metric	Model output	Sim.	word counts	Weise and Levitan (2018)
Graph Similarity	Prox., Conv.	Metric	Model output	Sim.	word graphs	Mehler et al. (2010)
Latent Sem. Sim.	Prox.	Metric	Model output	Sim.	word counts	Babcock et al. (2013)
LA Fusaroli	Prox.	Discriminative	Model output	CP	word presence	Fusaroli et al. (2012)
LA Wang	Prox.	Discriminative	Model output	CP	word, syntax presence	Wang et al. (2014)
SCP	Synch.	Discriminative	Model output	CP	fw presence	Danescu-Niculescu-Mizil et al. (2012)
SCI	Synch.	Discriminative	Model output	CP	fw presence	Gao et al. (2015)
Reg. Xu	Synch.	Discriminative	Model parameter	RCP	fw presence; fw count	Xu et al. (2018)
Reg. Reitter	Synch.	Discriminative	Model parameter	RCP	syntax presence	Reitter et al. (2006)
HAM	Synch.	Generative	Model parameter	BCP	high freq word presence	Doyle et al. (2016)
WHAM	Synch.	Generative	Model parameter	BCP	high freq. word presence	Doyle and Frank (2016)
Hawkes Process	Prox.	Generative	Model parameter	BCP	word counts	Guo et al. (2015)

Table 1: Each of the reviewed methods is categorized according to the factors discussed in Section 5. Measurement target describes the concept that the measure is targeting. Generative framework dictates what assumptions are made about the data generating process. Estimator type describes whether the estimator is a parameter within a model or the output of one. Functional form categorizes measures into abstract mathematical forms through which the model achieves a measurement outcome. Input Features articulates the inputs into the model such as function words (fw) or syntactic structures (syntax).

synchrony is said to be present. The definition and construction of each speaker's personal baselines are discussed in Appendix B.

As shown in Table 1, most measures quantify the *proximity* of speakers, typically by implementing a simple similarity or distance function. *Synchrony* is less common than *proximity*, and is often quantified by adding some control or baseline term that measures whether speakers are the same distance from their baseline at the same point of a conversation. Notably, *convergence* is quite uncommon as a measurement target and was studied by only three approaches in our review.

5.2 Generative Framework

All coordination measures compare features from two speakers, but approaches differ concerning how they model the process assumed to be generating these features. To differentiate along this axis, we group models into the *metric*, *generative*, and *discriminative* categories. This is important because it determines the quantity used to measure coordination and the framework in which that quantity is understood. A *metric* approach does not model the data generating process (DGP), whereas *generative* approaches model the full DGP (i.e. the text from both interlocutors). Discriminative approaches sit in-between, modeling one speaker's features as a random variable and the other's as given.

Metrics are typically simple and provide interpretable descriptions of the data. While *generative* approaches are often more complicated, they offer three distinct advantages. First, one can sample from the fitted model and evaluate the plausibility of this generated data vis-a-vis the real data (i.e. a posterior predictive check). Second, one can integrate additional variables into this generative process to answer specific theoretical questions. Third, one can leverage existing estimation and inferential results to reason about the assumptions, efficiency, and unbiasedness of the estimator. Discriminative methods retain the latter two advantages of a fully generative approach, but do not support full posterior predictive checks, given that co-variates have no distribution from which to sample.

The coordination measures discussed in this review span across the generative frameworks outlined above. An example of a *metric* approach is the widely used LSM CANBERRA measure, which has several variants. The most basic form from Ireland and Pennebaker (2010) is given by the equation

$$s_c = 1 - \frac{|f_c^1 - f_c^2|}{(f_c^1 + f_c^2 + \alpha)}, \qquad (1)$$

where f_c^p is the proportion of words in speaker p's transcript that belong to a particular function word category c, α is a positive smoothing constant (preventing division by zero), and s_c is the similarity between speakers for category c. In contrast, the popular SCP model introduced by Danescu-Niculescu-Mizil et al. (2012) takes a *discriminative* approach of the form

$$\Delta_c = p(1_c[W_n^2] \mid 1_c[W_n^1]) - p(1_c[W_n^2]), \quad (2)$$

where $1_c[W_n^p]$ is an indicator for whether speaker p uses a function word in class c in their nth conversation turn, and $\Delta_c(n)$ is the difference between the speaker's probability of using such a term given that the other speaker did in the previous utterance, minus their baseline probability of doing so. Here, speaker 2's function word usage is estimated as a probabilistic quantity whereas speaker 1's usage is treated as given. Doyle et al. (2016) and Guo et al. (2015) both offer fully *generative*, Bayesian models that embed assumptions about the DGP and use existing estimation and inference strategies.

5.3 Estimator Type

A similar but non-overlapping distinction between measures is whether the estimator of interest is a *parameter* in a model or a quantity calculated as the *output* of a model or metric. This consideration is important because it constrains the types of models one is able to employ and how one does so. When estimating the *output* of a model or metric, we care about modeling details only insofar as they improve our ability to reliably produce an unbiased estimate efficiently. For example, Weise and Levitan (2018) present a coordination measure based on the perplexity of utterances estimated with a tri-gram language model. Because a *model-output* estimator is used, this model could easily be swapped for a more complex, predictive model while maintaining the same estimator of interest (i.e. perplexity). Broadly, these estimators allow us to leverage "black box" models for their predictive power while sacrificing some specificity about how the estimator should be interpreted.

These considerations are reversed when the estimator of interest is a *parameter* within a model. Although we lose the model-agnostic property, we typically gain a better understanding of how our estimator fits into the DGP. In the HIERARCHICAL ALIGNMENT MODEL (HAM), for example, coordination is estimated as a parameter representing the change in probability of speaker 2 using a function word given that speaker 1 uses a function word (Doyle et al., 2016). Because the DGP is modeled such that each parameter has a theoretical significance, there is far more clarity concerning the assumptions being made. This interpretable approach also allows Doyle et al. (2016) to integrate social groups into their model explicitly and test specific theoretical hypotheses.

5.4 Input Features

Models for linguistic coordination typically generate input features by extracting numerical quantities from strings of text. Importantly, this process often narrows one's coordination measure to a particular linguistic aspect of interest. For example, style is captured using the count or presence of words falling into different function word categories in the LSM literature. Likewise, form is captured by extracting parsed syntactic constructions, drawing on psycholinguistic work in the syntactic priming domain (Bock, 1986; Gries, 2005). Other methods operate over the entire set of words uttered, capturing general lexical coordination or a broader, non-specific aspect of language (e.g. LA WANG, LA FUSAROLI, PERPLEXITY, EMBEDDING SIMILARITY in Table 1).

In addition to capturing linguistic constructs of interest, feature extraction requires the researcher to decide whether they will work at the turn level or the conversation level. Conversation-level features are extracted from each speaker's entire conversation transcript. In contrast, turn-level features are extracted from only single utterances, though conversation-level estimates are often constructed from aggregating this turn-level information. Further implications of this important decision are discussed in Section 6.2.

5.5 Features, Notation, and Functional Form

Linguistic coordination measures are ultimately expressed through a mathematical function that takes one or more features as input. Though each measure captures coordination in some form, the specific choices made in doing so have significant impacts on the validity of one's measurement (Section 6). Prior work has yet to abstract away from specific functions and inputs to better generalize over this space of possible modeling choices. We now address this gap by introducing a set of four functional forms within which our reviewed methods are categorized.

To introduce our functional form categories, we first define the standardized notation used to express them. For a given Feature_n^p of interest, the superscript p indexes the speaker that this feature is associated with, and the subscript n indexes the conversation turn. For example, Feature_n^1 and Feature_n^2 refer to feature values for adjacent conversation turns from speakers 1 and 2. By convention, we label speakers in order of their first turn, so that speaker 1's nth turn precedes speaker 2's. To describe the precise details of diverse coordination methods, we introduce additional notation in Appendix E.

Similarity:

$$\frac{\mathrm{Sim}(\mathrm{Feature}^2_n, \mathrm{Feature}^1_n)}{\mathrm{Norm}(\mathrm{Feature}^2_n, \mathrm{Feature}^1_n)} \tag{3}$$

Conditional Probability:

$$p(\mathrm{Feature}^2_n|\mathrm{Feature}^1_n) - \mathrm{Baseline}(\mathrm{Feature}^2_n) \tag{4}$$

Regressive CP:

$$\begin{aligned}\mathrm{Feature}^2_n \sim \mathrm{Feature}^1_n + \mathrm{Control}^1_n + \\ \mathrm{Control}^2_n + \mathrm{Baseline}^1 + \mathrm{Baseline}^2\end{aligned} \tag{5}$$

Bayesian CP:

$$\begin{aligned}\mathrm{Feature}^2_n \sim \mathrm{Distribution}_1(\boldsymbol{\theta}, \mathrm{Feature}^1_n), \\ \mathrm{Feature}^1_n \sim \mathrm{Distribution}_2(\boldsymbol{\psi})\end{aligned} \tag{6}$$

Each measure presented in our review is captured by one of four abstract functional-forms. These forms describe the mathematical equation(s) used to quantify coordination. In addition, each form also coincides with a set of decisions regarding the researcher's generative framework (Section 5.2), and estimator type (Section 5.3). We now describe each form, identify its connection to the other choices in our framework, and provide a concrete example from Table 1. Additional modeling details and equations can be found in Appendices B, E, and F.

The *Similarity* functional form describes methods that use a simple similarity (or distance) function to compare speakers and do not model the process generating these features (i.e. the DGP). Measures in this category are exclusively *metrics* that use the *output* of a model as their estimator. One widely used example of this functional form is LSM CANBERRA, which was introduced in §5.2. This approach uses the absolute difference in function word rates between speakers as the Sim() function from Eq. 3 and the sum of their function word rates as the Norm() function. A number of variants have also been proposed due to disagreement over the level at which to extract features (i.e. conversation, speaker-turn, sentence; Ireland and Pennebaker, 2010; Müller-Frommeyer et al., 2019; Arnet et al., 2024) and as extensions to more than two speakers (Gonzales et al., 2010; Müller-Frommeyer and Kauffeld, 2022).

Measures in the *Conditional Probability* category model the probability of speaker 2's feature(s) conditioned on speaker 1's feature(s), which are treated as fixed. Approaches with this functional form are therefore *discriminative* and use the model *output* as their estimator. The influential linguistic alignment model from Wang et al. (2014) belongs to this category and is estimated as the probability of speaker 2 using a given word or syntactic structure w in turn n conditioned on speaker 1 having done so, divided by the length of speaker 1's turn:

$$\textit{LA Wang} = \frac{p(w \in W^2_n \mid w \in W^1_n)}{|W^1_n|} \tag{7}$$

We use the *Regressive Conditional Probability* (*RCP*) category to describe regression approaches that estimate coordination as a *model parameter* and employ a *discriminative* probabilistic framework. Xu et al. (2018) provide a simple example of a model in this category. In their approach, coordination is estimated with a regression coefficient modeling the relationship between speaker 1's binary function word usage and speaker 2's probability of using a function word:

$$\textit{Reg. Xu:}\ \mathrm{logit}(1_c[W^2_n]) = \beta_0 + \beta_1 1_c[W^1_n] \tag{8}$$

Here $1_c[W^p_n]$ serves as an indicator function for whether speaker p used a function word in class c in their nth conversation turn. As shown in Eq. 5, the *RCP* functional form is flexible and allows for a number of additional covariates to be integrated into the model by the researcher.

Finally, *Bayesian Conditional Probability* describes a flexible set of approaches that estimate coordination as a parameter in a Bayesian model. These measures are thus *generative*, *model-parameter* estimators. The flexibility of this approach is demonstrated by Guo et al. (2015), who integrate Bayesian language modeling and the Hawkes process (Hawkes, 1971) to create a complex new model of mutual excitation in language. As demonstrated through this example, researchers using this functional form have the distinct opportunity to leverage longstanding Bayesian approaches to modeling text data (Blei et al., 2003; Teh, 2006).

As described above, our functional forms broadly align with specific choices of generative framework and parameter type. This list of choices is non-exhaustive, and a series of more subtle decisions determine how one normalizes to account for factors such as length or word frequency, and how one employs baselining to capture synchrony rather than proximity. These details are further elaborated in Appendix B. Ultimately, careful consideration must be given to choices across many levels of abstraction to ensure conceptually and mathematically valid estimation.

6 Critiques

The diversity and quantity of linguistic coordination measures has resulted in disagreement concerning measurement constructs and the most appropriate way to capture them. Having summarized important modeling choices, we now synthesize the main critiques and points of contention within the linguistic coordination literature, which we group into four main categories.

6.1 Conflicting Definitions of Coordination

One difficulty in organizing and discussing definitions of linguistic coordination is an abundance of theories, terms, and methods that are applied inconsistently (Doyle et al., 2016; Müller-Frommeyer et al., 2019; Paxton and Dale, 2013). For example, Niederhoffer and Pennebaker (2002) specify that they are not measuring synchrony, but define a measure that clearly captures synchrony under Levitan and Hirschberg (2011)'s framework.

In the case of LSM, synchrony and proximity measures have both been applied without recognition of the conflicting nature of these two measurement targets. As demonstrated by Doyle et al. (2016), this is highly problematic, as synchrony and proximity may be inversely related in some cases. Furthermore, the notion of an increased proximity throughout the conversation (convergence) is suggested by both the CAT (Giles et al., 1991; Soliz et al., 2021) and IAM (Pickering and Garrod, 2004) theories, yet only three measures for convergence were identified in our review. In future work, researchers can mitigate these concerns by precisely defining their measurement target and theoretical justification before introducing a mathematical operationalization.

6.2 Temporal Assumptions

Conversations evolve over time, and thus form time series created by each speaker's reliance on past utterances (Pickering and Garrod, 2004; Müller-Frommeyer et al., 2019). As described in Section 5.5, these complex temporal dependencies are typically simplified during feature extraction when the researcher decides to either 1) concatenate utterances to form speaker-level transcripts or 2) work at the utterance level. While concatenation may result in more precise estimates by pooling more data, it ignores all temporal dependency. This prevents the researcher from studying synchrony or convergence, obscuring key turn-level phenomena proposed by the IAM and CAT theories (Müller-Frommeyer et al., 2019). While turn-level measures can identify temporal effects such as priming (Reitter et al., 2006), estimation at this level may result in noisy estimates or bias due to short turns (see Section 6.3). Furthermore, these approaches typically assume that each utterance is independent after conditioning on the preceding utterance, which is likely insufficient to model the true complexity of conversation.

6.3 Bias

A significant hindrance to any measure's validity is bias, which results when an estimate's expected value is not equal to its true value. This means that even if one's theoretical assumptions and data collection processes are perfect, they are unable to recover the true value of the phenomenon they seek to measure, on average. Unfortunately, simulation data has suggested bias in linguistic coordination measures such as LSM CANBERRA, LA WANG, SCP and HAM. Furthermore, this bias is exacerbated for extreme word frequencies and ground-truth coordination values (Doyle et al., 2016). Additional evidence has investigated the potential for bias when applying the LSM CANBERRA measure to short, utterance-level data. Here, Arnet et al. (2024) find that utterances with approximately the same ground-truth coordination can vary by up to 0.15 in their estimate solely as a function of utterance length.[2] While this research has begun to describe instances of bias and their causes, more work is required to fully address these issues across the full spectrum of coordination measures.

6.4 Confounding

Even if a measure is unbiased for simulated data, there is no guarantee that coordination measurements represent truly causal relationships in real-world data. This issue arises when one would like to measure the relationship between two things (e.g. function word rates), but there exists a common cause behind both of these variables. This common cause is known as a confounder and can convince researchers of false-positive effects if not carefully accounted for.

[2] This applies to the LSM CANBERRA estimator, which has a range of $(0, 1)$.

One such confounder arises when raw function word counts are used as input features, rather than length-normalized word rates (Appendix Section B). In this setting, utterance length can act as a common cause of the word counts for both speakers and the researcher can falsely identify coordination of function words when in fact the true coordination occurs between utterance lengths (Gao et al., 2015; Xu et al., 2018). Confounds can also arise in observational data due to the context in which two speakers interact. For example, speaking partners with similar speaking patterns may be identified in data simply because they share an identity characteristic that makes them more likely to interact (Doyle et al., 2016). Rather than interaction causing coordinated dialogue, the identity characteristic acts as a common cause of both speaking patterns and interaction probability in this scenario.

Unfortunately, the absence of confounders in observational data cannot be empirically demonstrated and is always predicated on theoretical assumptions (Angrist and Pischke, 2009). Scholars can, however, take the approach of Doyle et al. (2016) and Gao et al. (2015) and reason about plausible confounders such as length and conversation context. Accounting for these confounders and others is the first step in ensuring the validity of one's causal claims outside of an experimental setting.

7 Discussion

As with all classification systems, our framework for linguistic coordination measures is inherently incomplete and subject to dispute and limitations. Due to the scope of our review, we were unable to consider measures for prosodic entrainment using audio features. Models in this domain have clear relevance to coordination and would have potentially increased the breadth of our framework. We were also unable to fully review every aspect of each measure due to limitations of parsimony and space. Aspects such as directionality, dyadic versus group communication, and feature extraction methods all present distinct challenges worthy of consideration. Furthermore, the methodological research under review frequently cited multiple overlapping and potentially conflicting theoretical bases for modeling decisions. For this reason our review was unable to fully disentangle the relationship between models such as the IAM and CAT, and the approaches used to implement them.

Our review revealed a number of methodological approaches that warrant further exploration. Trigram language models were employed by Weise and Levitan (2018), but advances in the predictive capabilities and contextual nuance captured by generative language models offer an opportunity to better model the perplexity of language. Future approaches may attempt to isolate or disentangle the type of coordination represented by these models to engage more directly with conventional approaches focused on syntactic and lexical priming.

As discussed in Section 6, another potential area for improvement involves designing models for complex temporal dependencies. One potential direction is the extension of classical time-series methods to this domain. For instance, vector autoregressive (VAR) models can be used to infer the relationship between lags across multiple time-series and have been widely applied in economics (Lütkepohl, 1991; Sims, 1980). Dynamic Bayesian Networks offer a similar approach but under a Bayesian inference framework (Dagum et al., 1992). In light of their ability to model complex temporal phenomena, neural approaches such as recurrent neural networks and transformer-based models could also be leveraged to estimate the influence between two or more speakers. Recent work has extended such models to infer interactions in dynamic graphs, but further inquiry is required to demonstrate the feasibility of these models in the context of linguistic coordination (Li et al., 2024; Wu et al., 2024; Wang et al., 2025).

Finally, our review found only three models for linguistic convergence—an increase or decrease in proximity throughout the course of a conversation. This suggests a methodological underdevelopment with respect to this phenomenon that warrants further inquiry and research.

While modeling choices require careful consideration of one's specific context, we now offer some general recommendations for those looking to apply linguistic coordination methods in their work. With respect to generative framework (Section 5.2), generative and discriminative approaches both have clear advantages over metrics. This is due to their (at least partial) modeling of the DGP, which allows them to communicate their assumptions clearly, leverage pre-existing modeling approaches, and sample data for posterior predictive tests.

For estimator types (Section 5.3), model output estimators allow for the usage of less interpretable yet highly predictive models; however, model parameter estimators are often more interpretable and allow one to reason about the entire DGP. While functional form is a secondary concern, one must be prudent that their model properly controls for extraneous linguistic factors and avoids common types of bias. A Hierarchical Bayesian model such as HAM meets many of these criteria, as it is generative, utilizes a parameter within an interpretable model, and is unbiased with respect to word frequency (Doyle et al., 2016). Likewise, the regression approach from Xu et al. (2018) provides a discriminative, model-parametric estimator that properly controls for length confounding. Despite these two recommendations, comprehensive testing is still required to analyze the robustness and unbiasedness of these models when measuring proximity, synchrony, and convergence. Furthermore, more work is needed to expand the space of linguistic convergence models as well as models that incorporate multiple types of linguistic coordination simultaneously.

8 Conclusion

In this review, we introduce a framework characterizing significant axes of variation in linguistic coordination measures. This schema considers measurement target, generative framework, estimator type, and functional form, discussing which choices are most common and how they impact the measurement approach. We then offer a set of critiques related to definitions of coordination, temporal assumptions, bias, and confounding, demonstrating many threats to validity in the linguistic coordination literature. Finally, we suggest directions for future work, including integrating insights from time-series modeling, and additional systematic comparison of methods using real or simulated data. We find that there is much room for improvement of linguistic coordination measures, and that attention to key modeling choices and common critiques is crucial in developing these new approaches.

Limitations

As discussed above, this review contains several limitations with respect to scope, coverage, theory, and implications. Although there are potentially fruitful links between linguistic coordination in audio and textual data, we have not included studies focused on audio data in this review, as it would have expanded the scope too much. Similarly, while any linguistic coordination method could be applied to experimental settings, including for human-computer interaction, we chose not to review this literature, as our focus was on models developed for studying naturalistic speech.

For similar reasons, we have not provided in-depth treatments of the theoretical background relevant to this work, or details of the substantive findings made using these methods. Although such information is relevant to users of these methods, the focus here is on the details of the methods themselves. An additional limitation results from the selected papers' focus on English as a language of study. Verification that each method behaves similarly when applied to other languages is critical to expand the scope and applicability of this work.

Finally, while the dimensions of variation we have identified have helped to surface the main types of approaches used to study linguistic coordination, there are other dimensions along which these methods could have been discussed, such as choices made around confounding factors. Depending on the applications, others might place emphasis on different aspects, but we hope that the details provided in the Appendix will still be helpful for identifying relevant work.

Acknowledgments

This work was supported in part by the National Science Foundation under Grant No. IIS-2143529. Thank you to Daniel Romero, Julie Boland, and anonymous reviewers for helpful comments on earlier drafts.

References

Joshua D. Angrist and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. Princeton University Press, Princeton.

Sandro Arnet, Anne Scherer, and Florian von Wangenheim. 2024. Reading between the lines: A refined methodology for measuring language style matching in conversations. *SSRN:4790188*.

Meghan Babcock, Vivian Ta-Johnson, and William Ickes. 2013. Latent Semantic Similarity and Language Style Matching in Initial Dyadic Interactions. *Journal of Language and Social Psychology*, 33.

Aviv Ben-Haim and Oren Tsur. 2021. Open-mindedness and style coordination in argumentative discussions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1876–1886, Online. Association for Computational Linguistics.

Frances R. Bilous and Robert M. Krauss. 1988. Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads. *Language & Communication*, 8(3):183–194.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

J Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive psychology*, 18(3):355–387.

Richard Y. Bourhis. 1979. Language in ethnic interactions: A social psychological approach. In Howard Giles and Bernard Saint-Jacques, editors, *Language and Ethnic Relations*, pages 117–141. Pergamon, Oxford.

Holly P. Branigan, Martin J. Pickering, and Alexandra A. Cleland. 1999. Syntactic priming in written production: Evidence for rapid decay. *Psychonomic Bulletin & Review*, 6(4):635–640.

Donn Byrne. 1997. An overview (and underview) of research and theory within the attraction paradigm. *Journal of Social and Personal Relationships*, 14(3):417–431.

Tanya L Chartrand and John A Bargh. 1999. The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893.

Cindy Chung and James Pennebaker. 2007. The Psychological Functions of Function Words. *Social communication*.

William S Condon and William D Ogston. 1967. A segmentation of behavior. *Journal of psychiatric research*, 5(3):221–235.

Paul Dagum, Adam Galper, and Eric Horvitz. 1992. Dynamic network models for forecasting. In *Uncertainty in artificial intelligence*, pages 41–48. Elsevier.

Cristian Danescu-Niculescu-Mizil. 2012. A computational approach to linguistic coordination. Ph.D. thesis, Cornell University.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power diffe grences in social interaction. *arXiv:1112.3670*.

Gabriel Doyle and Michael C. Frank. 2016. Investigating the Sources of Linguistic Alignment in Conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 526–536, Berlin, Germany. Association for Computational Linguistics.

Gabriel Doyle, Dan Yurovsky, and Michael C. Frank. 2016. A robust framework for estimating linguistic alignment in Twitter conversations. In *Proceedings of the 25th International Conference on World Wide Web*, pages 637–648, Montréal, Québec, Canada.

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.

Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. 2015. Understanding Confounding Effects in Linguistic Coordination: An Information-Theoretic Approach. *PLOS One*, 10(6):e0130167.

Howard Giles, Nikolas Coupland, and Justine Coupland. 1991. Accommodation theory: Communication, context, and consequence. In *Contexts of accommodation: Developments in applied sociolinguistics*, pages 1–68. Paris, France.

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

Stefan Th. Gries. 2005. Syntactic Priming: A Corpus-based Approach. *Journal of Psycholinguistic Research*, 34(4):365–399.

Fangjian Guo, Charles Blundell, Hanna Wallach, and Katherine Heller. 2015. 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.

Jerold L Hale and Judee K Burgoon. 1984. Models of reactions to changes in nonverbal immediacy. *Journal of Nonverbal Behavior*, 8:287–314.

Alan G. Hawkes. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Katharina Heuer, Lena C Müller-Frommeyer, and Simone Kauffeld. 2020. Language matters: The double-edged role of linguistic style matching in work groups. *Small Group Research*, 51(2):208–228.

David Huffaker, Joseph Jorgensen, Francisco Iacobelli, Paul Tepper, and Justine Cassell. 2006. Computational measures for language similarity across time in online communities. In *Proceedings of the Analyzing Conversations in Text and Speech*, pages 15–22, New York City, New York. Association for Computational Linguistics.

Molly Ireland and Marlone Henderson. 2014. Language style matching, engagement, and impasse in negotiations. *Negotiation and Conflict Management Research*, 7:1–16.

Molly Ireland and James Pennebaker. 2010. Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of personality and social psychology*, 99:549–71.

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

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: Technical Papers*, pages 455–465, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Jay Kejriwal and Štefan Beňuš. 2023. Relationship between auditory and semantic entrainment using deep neural networks (DNN). In *Interspeech 2023*, pages 2623–2627. ISCA.

Rivka Levitan and Julia Hirschberg. 2011. Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Interspeech 2011*, pages 3081–3084. ISCA.

Dongyuan Li, Shiyin Tan, Ying Zhang, Ming Jin, Shirui Pan, Manabu Okumura, and Renhe Jiang. 2024. Dygmamba: Continuous state space modeling on dynamic graphs.

Helmut Lütkepohl. 1991. *Introduction to multiple time series analysis*. Springer-Verlag, Berlin; New York.

Alexander Mehler, Andy Lücking, and Petra Weiß. 2010. A Network Model of Interpersonal Alignment in Dialog. *Entropy*, 12(6):1440–1483.

E. Mousset, W.A. Ainsworth, and J.A.R. Fonollosa. 1996. A comparison of several recent methods of fundamental frequency and voicing decision estimation. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 2, pages 1273–1276 vol.2.

Lena C. Müller-Frommeyer, Niels A. M. Frommeyer, and Simone Kauffeld. 2019. Introducing rLSM: An integrated metric assessing temporal reciprocity in language style matching. *Behavior Research Methods*, 51(3):1343–1359.

Lena C. Müller-Frommeyer and Simone Kauffeld. 2022. Capturing the temporal dynamics of language style matching in groups and teams. *Small Group Research*, 53(4):503–531.

Md Nasir, Sandeep Nallan Chakravarthula, Brian R. Baucom, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2019. Modeling interpersonal linguistic coordination in conversations using word mover's distance. In *Interspeech 2019*, pages 1423–1427.

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Short Papers - HLT '08*, page 169, Columbus, Ohio. Association for Computational Linguistics.

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

Jennifer S Pardo. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393.

Alexandra Paxton and Rick Dale. 2013. Frame-differencing methods for measuring bodily synchrony in conversation. *Behavior research methods*, 45:329–343.

Martin J. Pickering and Simon Garrod. 2004. The interactive-alignment model: Developments and refinements. *The Behavioral and Brain Sciences*, 27(2):212–225.

David Reitter, Frank Keller, and Johanna D. Moore. 2006. Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 121–124, New York City, USA. Association for Computational Linguistics.

Christopher A. Sims. 1980. Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Jordan Soliz, Howard Giles, and Jessica Gasiorek. 2021. Communication accommodation theory. In *Engaging Theories in Interpersonal Communication*, 3 edition, pages 130–142. Routledge, New York.

D. M. Taylor, L. Simard, and D. Papineau. 1978. Perceptions of cultural differences and language use: A field study in a bilingual environment. *Canadian Journal of Behavioural Science*, 10:181–191.

Yee Whye Teh. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992.

M. E. van den Berg. 1985. *Language Planning and Language Use in Taiwan*. ICG Printing, Dordrecht.

Yafei Wang, David Reitter, and John Yen. 2014. Linguistic adaptation in conversation threads: Analyzing alignment in online health communities. In *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics*, pages 55–62, Baltimore, Maryland, USA. Association for Computational Linguistics.

Zhiqiang Wang, Baijing Hu, Kaixuan Yao, and Jiye Liang. 2025. Hawkes point process-enhanced dynamic graph neural network. In *Proceedings of the Eighteenth ACM International Conference on Web*

Search and Data Mining, WSDM '25, page 401–409, New York, NY, USA. Association for Computing Machinery.

Andreas Weise and Rivka Levitan. 2018. Looking for structure in lexical and acoustic-prosodic entrainment behaviors. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 297–302, New Orleans, Louisiana. Association for Computational Linguistics.

Sheida White. 1989. Backchannels across cultures: A study of americans and japanese. *Language in Society*, 18(1):59–76.

Yuxia Wu, Yuan Fang, and Lizi Liao. 2024. On the feasibility of simple transformer for dynamic graph modeling. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 870–880, New York, NY, USA. Association for Computing Machinery.

Yang Xu. 2021. Global divergence and local convergence of utterance semantic representations in dialogue. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 116–124, Online. Association for Computational Linguistics.

Yang Xu, Jeremy Cole, and David Reitter. 2018. Not that much power: Linguistic alignment is influenced more by low-level linguistic features rather than social power. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 601–610, Melbourne, Australia. Association for Computational Linguistics.

Yang Xu and David Reitter. 2015. An evaluation and comparison of linguistic alignment measures. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–67, Denver, Colorado. Association for Computational Linguistics.

Appendix

A Searching the ACL Anthology

To ensure coverage of linguistic coordination measures within the natural language processing discipline, we conducted a search of the ACL Anthology. This search was comprised of keywords "Linguistic Style Matching", "Linguistic Alignment", "Entrainment", "Accommodation", and "Linguistic Coordination". We then compiled the papers returned by this search and selected a reading list according to methodological innovation, citation count, and relevance to the criteria listed in Section 3. Our initial list was comprised of 75 papers, and our deeper reading list was then narrowed to roughly 30 papers, with rolling additions and deletions upon closer reading and review.

B Further Model Considerations

Here, we introduce additional modeling considerations surrounding normalization and baselines. These two considerations are incorporated differently across different measures, but ultimately ensure that a mathematical equation is better measuring the construct it purports to.

B.1 Normalizing for Length and Frequency

Normalization terms are often employed to account for utterance length and word-frequency. Utterance length must be considered to avoid estimating turn length in lieu of word frequency. For example, a function word *w* will likely appear more in a turn with 100 words than 10. If one counts *w* without accounting for the length of the utterance(s) in which it occurred (e.g., Niederhoffer and Pennebaker, 2002), there is a clear confound between the count of *w* and the length of the turns it's used in. This issue can be accounted for by employing rates as an input feature rather than raw counts (e.g., Ireland and Pennebaker, 2010), adding a divisor to an equation using raw counts (e.g., Xu et al., 2018), or by incorporating length as a covariate in a Regression approach (Fusaroli et al., 2012; Wang et al., 2014). Although length normalization is important in reducing confounding (Section 6.4; Appendix Section D), it must be applied carefully to avoid producing a biased estimator, especially when using short conversation turns (Section B).

Normalization terms are also applied to account for the the overall frequency of a given linguistic construct. Continuing with the example from above, speaker 1 may have a rate for *w* that is 0.2 higher than speaker 2; however, this difference in rates is more meaningful if the overall frequency of *w* for speaker 1 and speaker 2 is 0.01 rather than 0.10 (i.e. *w* is a low vs. high frequency word). By normalizing for function word rate, one achieves "frequency sensitivity" (Müller-Frommeyer et al., 2019), a property allowing one to compare coordination measures across linguistic constructs with varying frequencies (Doyle et al., 2016).

B.2 Targeting Synchrony through Baselining

Whereas proximity refers to simply the distance between features for speakers 1 and 2, synchrony captures a more subtle phenomenon of repeated, parallel movement in adjacent conversation turns (Section 5.1). In modeling terms, synchrony is typically captured as a proximity measure with an additional baseline term. In this formulation, synchrony is defined as similarity in features for speaker 1 and speaker 2, relative to their personal baseline across the entire conversation. Even when two speakers do not have similar feature values, if their features are similarly offset from their baselines in adjacent turns, then synchrony has occurred within this measurement framework.

The clearest example of this operationalization is Danescu-Niculescu-Mizil et al. (2012)'s SCP measure. Here, speaker p's baseline probability of using function word *w* is subtracted from their probability of using *w* conditioned on speaker p' (where $p \neq p'$) having done so. Xu et al. (2018) also employ a synchrony measure in their replication of this work, in their case using a regression model where speaker p's baseline usage of *w* is controlled for as a covariate.

While Levitan and Hirschberg (2011) introduce synchrony as a "relative coordination" between speaking partners, they do not define what this coordination is relative to. In practice, the synchrony measures in our review all employ a constant, conversation-level baseline for each speaker as their reference point; however, the assumption that this baseline is constant over time may or may not reflect the underlying structure of the data. If, for example, speakers have baseline rates that change throughout a conversation, then it is no longer appropriate to deploy static baselines against which to measure synchrony. At present, it appears that this consideration is lacking from the coordination literature and is a potential area for further investigation. The use of baselines also raises the question of how to estimate them, which itself presents additional issues of conceptualization and measurement.

C Bias in Coordination Measures

Coordination Strength and Marker Frequency Doyle et al. (2016) provide evidence of bias for varying combinations of marker frequency and alignment strength. To do so, they create simulation data with a predetermined degree of coordination, and attempt to recover this ground-truth value with the LSM CANBERRA, LA WANG, SCP, and HAM estimators. These simulations reveal that, with the exception of HAM, coordination measures are not robust to changes in marker frequency. As noted by Doyle and Frank (2016), this likely results from a failure to appropriately normalize for this factor, an issue that has also been raised by Müller-Frommeyer et al. (2019) and Ireland and Pennebaker (2010). For further discussion on this type of normalization, see Appendix B.

Doyle and Frank (2016)'s simulation data also reveal bias with respect to a range of ground-truth coordination values for the SCP, LA WANG, and LSM CANBERRA measures. In these cases, a linear change in the amount of ground-truth coordination may result in a highly non-linear change in the amount of coordination reported by these measures—resulting in drastic over or under-estimations.

Rather than an insurmountable instance of bias, however, this finding primarily reflects disagreement surrounding the scale on which to measure linguistic coordination (Doyle and Frank, 2016). Indeed, there is still widespread disagreement surrounding the appropriate "unit" with which to measure coordination, with various authors using, for example, probability (Danescu-Niculescu-Mizil et al., 2012), information content (Gao et al., 2015), or correlation (Niederhoffer and Pennebaker, 2002). Without standardizing this decision, any simulation is bound to identify bias when testing a method that measures coordination with respect to a different scale.

Length Arnet et al. (2024) demonstrate bias of the LSM CANBERRA measures when applied to short utterances. This implies that for the same coordination strength, estimates can vary drastically depending on the length of text used to estimate them. Using random pairs of equal-length utterances, Arnet et al. (2024) find that LSM CANBERRA increased by over 0.15 when increasing turn-length from 21 to 320 words. This is further evidenced by Müller-Frommeyer and Kauffeld (2022)'s finding that for the same pair of speakers, LSM CANBERRA is much higher when calculated on the concatenated text from two speakers (i.e. Table 1: LSM Canberra$_{1,2}$) versus an average of their turn-level LSM scores (i.e. Table 1: LSM Canberra$_{3,4,5}$). As discussed by Arnet et al. (2024), this problem results from discrete approximations of function word rates combined with an estimator that is non-linear with respect to these inputs. For short messages, estimated function word rates can only take on a very limited set of discrete values, which (in most cases) either under or overestimate the true function word rate of interest. When fed into a non-linear function, under- and over-estimates of the same magnitude are no longer equivalent, so estimates do not "average out" to the true coordination value. For longer turns, function word rates are more faithfully approximated, which minimizes this problem.

D Confounding

Length The first source of confounding in the coordination literature is turn-length, which has been discussed as a critique of Danescu-Niculescu-Mizil et al. (2012)'s study that established a link between function-word synchrony and status. In a replication study, both Gao et al. (2015) and Xu and Reitter (2015) argue that Danescu-Niculescu-Mizil et al. (2012) failed to properly control for utterance length. As a result, what was primarily a coordination in utterance lengths was falsely attributed to coordination in function-word counts. After controlling for utterance length, function word synchrony greatly decreased and its relationship with status was no longer present.

Although length has only been discussed as a confounder within the context of Danescu-Niculescu-Mizil et al. (2012)'s work, this critique can be reasonably applied to a number of other methods. Approaches taken by Niederhoffer and Pennebaker (2002); Nenkova et al. (2008) and Doyle and Frank (2016) operate over raw word counts, leaving them vulnerable to problems of length-confounding. Luckily, this issue can be addressed by working with function word rates (e.g., Ireland and Pennebaker, 2010; Gonzales et al., 2010; Arnet et al., 2024) or holding utterance lengths constant when estimating the relationship between counts (e.g., Xu and Reitter, 2015; Gao et al., 2015).

Contextual Confounding In addition to confounding due to length, additional confounders may arise due to the context in which speakers interact.[3] One such case is homophily, where speakers' outcomes (e.g., conversation enjoyment, task success, choice of interlocutor) and speaking patterns share a common cause, such as a demographic characteristic. Doyle and Frank (2016) discuss one such example where an association was drawn between speed dating success and coordinated language (Ireland et al., 2011). Without accounting for potential confounds, it is unclear whether success is linked to coordination or whether both variables are caused by a factor such as common background.

In a similar vein, Gao et al. (2015) design a test to identify contextual confounding in Danescu-Niculescu-Mizil et al. (2012)'s work. By randomizing conversation turns, they argue that all remaining coordination must be due to contextual factors rather than turn-level influence (synchrony in this case). Under this assumption, they find that contextual confounds were present in one of Danescu-Niculescu-Mizil et al. (2012)'s two datasets.

E Unified Notation

To describe many methods with a cohesive framework, we re-use and extend notation from Guo et al. (2015). Each speaker is indexed by $p \in \{1, ..., P\}$ and has a set of N^p utterances $W^p = \{W_1^p, ..., W_{N^p}^p\}$. Each utterance W_n^p is composed of a set of L_n^p words, such that the nth utterance from speaker p is represented as $W_n^p = \{w_{0,n}^p ... w_{L_n^p,n}^p\}$. Speakers are indexed by p, utterances are indexed by n, and words are indexed l. This can be expressed compactly as follows:

$$\mathcal{W} = \{\{\{w_{l,n}^p\}_{l=1}^{L_n^p}\}_{n=1}^{N^p}\}_{p=1}^{P}$$

Methods also frequently operate over the entire collection of words from a particular utterance or speaker concatenated together. To represent these cases, we use W_n^p to signify the concatenated string of words for speaker p in their nth utterance, and W^p to represent the concatenation of all words for the pth speaker. The cardinality of the words in these sets is written as $|W_n^p|$ and $|W^p|$ respectively.

A number of methods rely on function word classes, $c_k \in \{c_k\}_{k=1}^K$. We represent the number of words in a particular utterance W_n^p belonging to a function word class as $|W_n^p|_{c_k} = |\{w_{l,n}^p : w_{l,n}^p \in c_k\}|$. The percentage of function words in utterance W_n^p belonging to a function word class is represented as $\%_{c_k} W_n^p = |W_n^p|_{c_k} / |W_n^p|$. Speaker-Level analogs are represented as $|W^p|_{c_k}$ and $\%_{c_k} W^p$ respectively.

F Coordination-Measure Equations

Here, we present some of the methods summarized in Table 1 in more detail, using the unifying notation from Appendix E.

LSM Canberra is a similarity measure that takes in function word rates from two speakers and reports a normalized similarity between them (Ireland and Pennebaker, 2010). Below, we show this measure at the turn-level with inputs W_n^1 and W_n^2, with smoothing constant α; however, this measure has also been proposed as a sentence and conversation-level approach.

$$\textit{LSM Canberra} = 1 - \frac{|\%_{c_k} W_n^2 - \%_{c_k} W_n^1|}{\%_{c_k} W_n^2 + \%_{c_k} W_n^1 + \alpha} \tag{9}$$

LA (Fusaroli) is a metric approach that calculates the probability that an arbitrary word *w* in speaker 1's n'th turn is also used in speaker 2's n'th turn. This is calculated as the difference between the sets of words in speaker 2 and speaker 1's n'th turn divided the length of speaker 2's turn.

$$\begin{aligned} \textit{LA (Fusaroli)} &= p(w \in W_n^1 \mid w \in W_n^2) \\ &= \frac{\sum_{w_l \in W_n^1} 1_{W_n^2}(w_l)}{|W_n^2|} \end{aligned} \tag{10}$$

Where $1_{W_n^2}(w_l)$ is an indicator function for whether speaker 2's n'th utterance contains the word w_l from speaker 1's n'th turn W_n^1, and turns have length greater than zero, by definition.

LA (Wang) is similar to LA (Fusaroli) but additionally divides by the length of speaker 1's turn:

$$\begin{aligned} \textit{LA (Wang)} &= \frac{p(w \in W_n^2 \mid w \in W_n^1)}{|W_n^1|} \\ &= \frac{\sum_{w_l \in W_n^1} 1_{W_n^2}(w_l)}{|W_n^1| * |W_n^2|} \end{aligned} \tag{11}$$

Subtractive Conditional Probability (SCP) is an approach that estimates the increase or decrease

[3] Notably, one of CAT's earliest impacts was to distinguish the role of social context from that of interpersonal influence (Giles et al., 1991).

in the probability of a speaker *p* using a function word in class c_k when another speaker *p'* has also done so, relative to speaker *p*'s baseline.

$$\mathit{SCP} = p(1_{c_k}(W_n^2) \mid 1_{c_k}(W_n^1)) - p(1_{c_k}(W_n^2)) \tag{12}$$

where $1_{c_k}(W_n^{s^p})$ is an indicator function for whether speaker *p*'s n'th utterance contains a word w_l^n belonging to function word class c_k:

$$1_{c_k}(W_n^p) := \exists\ w_l^n \in\ W_n^p \text{ s.t. } w_l^n \in c_k \tag{13}$$

Regression approaches encompass a wide variety of statistical estimation techniques. In our classification system, we use this term to describe discriminative models where a model parameter serves as the estimator of interest. Xu et al. (2018) propose one such model that includes a speaker-level baseline (Appendix B) to capture synchrony:

$$\textit{Reg. Xu:}\ \text{logit}(m) = \beta_0 + \beta_1 1_{c_k}(W_n^{s^p}) \tag{14}$$

The **Perplexity** approach from Weise and Levitan (2018) creates a language model using speaker 1's text, and computes the perplexity of this model's predictions on speaker 2's text. For a given ground-truth probability distribution and corresponding estimator of this probability distribution, perplexity is a mathematical equation expressing how well the estimator predicts samples from the ground-truth distribution. High perplexity indicates high uncertainty of the model with respect to the data. Thus, a high perplexity in Weise and Levitan (2018)'s setting indicates that a model learned from speaker 1's text does not provide a good estimate for speaker 2's text.

The **Hawkes Process** is a self-exciting model where the probability of q events at time t depends on rate parameter λ, which itself depends on the number of events in each of the previous time steps. Guo et al. (2015) build a model of *mutual* excitation between speakers such that the distribution of words used by speaker *p* at time *t'* depends on the distribution of words used by other speakers at steps $t < t'$. The parameters of interest in this model lie within a matrix ρ^{qp} estimating the degree of excitation between speakers q and p. By including appropriate priors, the model becomes fully generative and is estimated using a Gibbs sampling approach. For the sake of brevity, we refer the reader to Guo et al. (2015)'s work for further detail.

The **Hierarchical Alignment Model**, or *HAM*, is a generative model that measures linguistic coordination as the increased probability of speaker 2 using a function word *w* given that speaker 1 has done the same. This model includes a chain of prior distributions allowing the researcher to model differences in this conditional probability for different group statuses (e.g. high vs. low power) and types of linguistic markers.