Behavior Research Methods
https://doi.org/10.3758/s13428-023-02267-2

ORIGINAL MANUSCRIPT

BERTs of a feather: Studying inter- and intra-group communication via information theory and language models

Zachary P Rosen[1] · Rick Dale[2]

Accepted: 5 October 2023

Abstract

When communicating, individuals alter their language to fulfill a myriad of social functions. In particular, linguistic convergence and divergence are fundamental in establishing and maintaining group identity. Quantitatively characterizing linguistic convergence is important when testing hypotheses surrounding language, including interpersonal and group communication. We provide a quantitative interpretation of linguistic convergence grounded in information theory. We then construct a computational model, built on top of a neural network model of language, that can be deployed to measure and test hypotheses about linguistic convergence in "big data." We demonstrate the utility of our convergence measurement in two case studies: (1) showing that our measurement is indeed sensitive to linguistic convergence across turns in dyadic conversation, and (2) showing that our convergence measurement is sensitive to social factors that mediate convergence in Internet-based communities (specifically, r/MensRights and r/MensLib). Our measurement also captures differences in which social factors influence web-based communities. We conclude by discussing methodological and theoretical implications of this semantic convergence analysis.

Keywords Alignment · Convergence · Information theory · Communication accommodation theory · Language models

Introduction

A surfer is not born with an innate knowledge of the appropriate usage of the word "dude" as an emphatic discourse marker. Nor would you expect an infant to understand the nuanced meaning of the word "slay" in "I don't play, I slay" in Todrick Hall's song "Nails, Hair, Hips, Heels." These are nevertheless acquired through engagement with the social environment, and indeed language is replete with instances of group-specific lexical patterns. These not only function to express particular meanings among individuals with shared knowledge, but they also signal commonalities amongst a community of speakers. To understand the word "dude" as an exclamation point made vocal is to understand that its speaker and their close associates likely know where the best local surf break is.

Much like the examples above, human beings rely on language for many purposes not limited just to the transfer of information. One of those functions is to signal their group identity to interlocutors, a process known as social identity signaling (SIS: Tajfel, 1979; Tajfel, Billig, Bundy and Flament 1971; Smaldino, 2019). Work in this area has found that shifts in how individuals communicate can convey subtle information about the social identities of speakers (Zhang et al., 2019; Doyle, Goldberg, Srivastava & Frank, 2017). These changes range in subtlety from being easily observable in speakers' utterances, to being "covert" depending on the social setting and the advantages conferred to the speaker (Smaldino, Flamson & McElreath, 2018). The use of language in this way serves an important function in both group formation and identity maintenance (Soliz, Giles & Gasiorek, 2021).

Communication accommodation theory (CAT) operationalizes SIS and its utility in group identity management as part of "convergence" – the process by which individuals tend to converge on a similar way of speaking with other members of the same group. This can be considered an instance

✉ Zachary P Rosen
zrosen@saddleback.edu

Rick Dale
rdale@ucla.edu

[1] Communication Studies Saddleback Community College, Mission Viejo, CA, USA

[2] Department of Communication UCLA, Los Angeles, CA, USA

Published online: 29 November 2023

of a general phenomenon referred to as "accommodation," the process by which individuals adapt their communication habits given input from other interlocutors (Soliz, Giles & Gasiorek, 2021). Linguistic convergence is observed in the ways individuals form coalitions by converging on similar words and meanings when discussing a topic (Dragojevic & Giles, 2014; Bradac, Mulac & House, 1988; Doyle, Goldberg, Srivastava & Frank, 2017; Pickering & Garrod, 2004). As a well-formulated conceptual framework, convergence has been used to study communication and identity management within a wide variety of intergroup scenarios ranging from political communication (Nganga, 2020), gender communication and power dynamics (Adams et al., 2018; Bamman, Eisenstein & Schnoebelen, 2014), race (Bailey, 2000), and a host of other social phenomena (Giles et al., 2007; Pérez-Sabater & Maguelouk, 2019; Bailey, 2000; Shin & Doyle, 2018; Mange, Lepastourel & Georget, 2009; Ba & Zhao, 2021; Bormann, 1982; Paxton, Dale & Richardson, 2016). Convergence thus provides an empirical framework to describe both how and why we tend to sound like our friend group.

Convergence is not limited to any particular class of lexical units, because individuals can (and do) converge stylistically across myriad linguistic dimensions (Tausczik & Pennebaker, 2010; Branigan et al., 2000; Garrod & Anderson, 1987; Pickering & Garrod, 2004; Paxton, Dale & Richardson, 2016). Such adaptations can occur at the level of the content and conceptual makeup of utterances (Pickering & Garrod, 2004), or surface level linguistic features (Branigan et al., 2000; Tausczik & Pennebaker, 2010).

From a slightly different perspective, convergence means that upon hearing one group member talk about a particular topic an interlocutor should be able to predict what a second group member might say when discussing the same topic. In this way, there is high *mutual information* in the way that group members collectively talk. Because of this dynamic quality of communication, convergence requires interlocutors to pull bits and pieces of lexical patterns from a history of interactions with one another. As Pickering and Garrod (2004) describe, linguistic convergence could operate through a low-level priming mechanism in which each utterance from one partner serves to "prime" or activate its use in the other.

Priming, however, should not yield precise copying or imitation, because interlocutors must construct novel utterances serving to advance an ongoing dialog while simultaneously signaling their social proximity to their fellow discursive partners. Because of these two competing constraints, it would be strange if interlocutors merely parroted one another at each turn – such behavior would not advance dialog in any meaningful way. In order to meaningfully signal their relationship to one another via convergence, interlocutors construct each new utterance from the bits and pieces of lexical patterns that they have previously heard used by their fellow discursive partners. This allows interlocutors to collaboratively build on their shared representation of the topic while advancing the conversation. The longer or more frequently interlocutors interact with one another, the larger their linguistic repertoire for convergence.

Linguistic convergence has repercussions for how members conceptualize various topics. This observation has been borne out in at least two distinct areas of inquiry – interactive alignment (IA: Pickering and Garrod, 2004; Branigan et al.,2000) and conceptual pact (CP: Brennan, Galati and Kuhlen, 2010; Brennan and Clark, 1996) theories. In both theoretical positions, when two people align in their words, meanings, and grammatical patterns, it may indicate an alignment in mental representations of the world. Despite this similarity, these theoretical positions propose distinct underlying mechanisms to be most important to this process. CP posits that interlocutors directly align their core conceptual understanding of objects and events in order to better coordinate joint action (Brennan, Galati & Kuhlen, 2010). In contrast, IA posits that alignment in interlocutors' mental representations is primed through the repetition of lower level lexical and grammatical structures (Pickering & Garrod, 2004).

It is not within the scope of the current paper to fully address these differences in detail. But empirical studies stemming from both IA and CP indicate that investigating alignment in communication practice offers a window into interlocutors' shared mental processes. New measurements along these lines may help mitigate these theories. For example, CAT does not outwardly or directly contradict either of these theories in terms of observed group behavior – i.e., convergence. However, CAT does propose a powerful social mechanism for why interlocutors might align in the first place. When considering all three theories, we see the importance of measuring and describing the verbal patterns of intragroup communication. It can elucidate what the most common mental models are amongst members of a group – what modes of conceptualizing and interacting with the world mark their group membership. It can, in turn, inform these theories and clarify their explanatory value.

Convergence likely plays a role in the formation of many identities, including identities associated with political and moral domains. Research on indoctrination and online extremism have shown that individuals accommodate over time to the lexical patterns used by existing members of extremist groups (Bäck, Bäck, Sendén & Sikström, 2018). With respect to cult indoctrination, particular deference is paid to convergence in the BITE model outlined by Hassan (2017) as a mechanism for inductees to signal their identity as part of the cult, and as a mechanism to communicatively isolate members from outside relationships (Hassan & Shah, 2019; Hassan, 2017).

Because of the central importance of convergence in group communication, new means of measuring and studying it are critical in an era when group formation is taking place rapidly online, sometimes in large and highly dynamic communities. The labile nature of online groups, and the massive data they generate online, pose research challenges. The current study provides a quantitative definition of linguistic convergence grounded in information-theoretic terms. This definition yields two major contributions for the study of identity signaling and linguistic convergence. First, it provides a structured definition that can be used to test hypotheses about group behavior in a quantitative framework. Second, this definition facilitates extending the study of linguistic convergence to data sets and corpora that would be prohibitively large to study under other circumstances. Importantly, our quantitative definition is descriptive, not prescriptive. We propose it to assist in ongoing research into group dynamics in conjunction with existing frameworks focusing on qualitative aspects of linguistic convergence.

In the next section, we outline our quantitative definition of convergence, as well as discuss the natural language processing (NLP) tools we use to implement this definition. In the following section, we provide a computational model that leverages advances in artificial intelligence and language modeling to deploy the quantitative definition on corpora of varying sizes. After that we present two case studies. The first shows that our model of convergence can quantify interpersonal interactions from a well-known corpus of conversation. The second case study explores the internal dynamics of far-right misogynist rhetoric compared to less extremist leaning groups on Reddit. We then conclude with a discussion of the model's drawbacks, how it might fit into a larger ethnographic framework, and potential extensions of the model.

Language models and group differences in linguistic convergence

Part of the problem with capturing subtle differences in lexical patterns between individuals (and groups) stems from the way that researchers have historically been limited to representing semantic meaning. Most research looking at linguistic convergence has traditionally relied on methods like linguistic inquiry and word count (LIWC: Tausczik and Pennebaker, 2010) or other methodologies leveraging hand-curated lists of lexical items to search for in group members' utterances. We instead propose leveraging a computational semantics approach based on advances in natural language understanding (NLU). In conjunction with our entropy-based definitions of convergence, we can capture stylistic differences in lexical choices between individuals and groups without needing to start from a hand-picked dictionary.

NLU researchers have used pre-trained language models (LMs) to capture lexical meaning for decades now, especially in psychology and cognitive science (Dumais et al., 1996; Lund & Burgess, 1996; Jones & Mewhort, 2007; Johns, 2021), and applied them to a variety of domains (e.g., Landauer, Foltz and Laham, 1998; Landauer, McNamara, Dennis and Kintsch, 2013). LMs take words/tokens and project them into a high-dimensional vector space, where any word/token's *word vector* in that vector space will be closer to word vectors for other words/tokens that share a similar meaning. With the advent of transformer models these pre-trained LMs become sensitive to the way that *context* modulates lexical meaning across utterances/sentences in a corpus. Consider the following examples:

Polysemy

(1) Adazee went to the *bank* to deposit a hefty check.
(2) I sat on the *bank* of Lake Aheme, my toes in the cool sand.

Intergroup variation

(3) It is unfortunate, that *slaying* the dragon became Lancelot's endless labor.
(4) Omg you're *slaying* today. Your makeup is perfection!

Examples 1 and 2 are examples of a linguistic phenomenon called polysemy, where the same word form can have different unrelated meanings. Previous LMs like GloVe or Word2Vec (Pennington, Socher and Manning Pennington, Socher and Manning and Mikolov et al. 2013, respectively) would represent the meaning of the word "bank" in 1 and 2 using the same word vector, effectively conflating the two separate meanings. This can cause problems when contextual differences fundamentally alter the intended meaning of a word. In contrast, transformer models like BERT are quite good at capturing differences in word senses like those shown in examples 1–4 Wiedemann, Remus, Chawla and Biemann (2019); Yenicelik, Schmidt and Kilcher (2020); Soler and Apidianaki (2021). This is because transformer language models compose a word vector for each word in a sentence by processing a series of weighted sums of the adjacent word vectors in a large, deep neural network (note: BERT repeats this weighted summation across 12 hidden layers in a large, deep neural network; Devlin, Chang, Lee and Toutanova (2019)). By doing this, words get assigned a new word vector that is entirely context dependent, circumnavigating the conflation of meaning problem.

This feature of BERT also enables researchers to use these word vectors to study phenomena like those described in examples 3 and 4. In these two examples, the meaning of the word “slaying” is different depending on its context, and that context varies by populations of speakers. Both hearken back to the image of “slaying” some foe, but the difference in the way that “slaying” is used in 4 is consistently contextualized in similar utterances, and based entirely on group-level habits in lexical pattern usage. When members of the same social group as the original speaker read or hear example 4, these group-level contextual differences enhance the semantics of “slaying” in a group-specific way. With respect to convergence, if we know that lexical patterns vary across groups as exemplified in 3 and 4, then transformer-based word vectors can grant us access to studying these subtle patterns in lexical usage.

An important question is if such word vectors correlate with human semantic processing in any meaningful way. While this question is being actively studied, there is some strong evidence that word vectors do indeed correlate well with human semantic processing, both behaviorally and neurologically. Previous research has noted correlations between word vector representations and patterns of activation in cortical tissue (Utsumi, 2020). Recent work indicates that transformer models in particular share computational principles with human semantic processing (Goldstein et al., 2022). Researchers have also shown that word vectors are particularly useful in encoding social and more abstract conceptual information embedded in language (Nishida, Blanc, Maeda, Kado & Nishimoto, 2021; Johns, 2021; Jones & Mewhort, 2007). While we do not make the claim that the models that generate word vectors represent lexical meaning in precisely the same way that the human mind does, it is safe to assume that these models do capture pertinent information about real human linguistic behavior.

Defining some recurring terms

Before describing our proposed framework, we define some terms that will be repeated throughout the remainder of this paper. We do this to clarify a number of key ideas, especially for readers who are unfamiliar with this kind of analysis.

In this paper, we refer to some utterance whose semantic content we wish to analyze as x. Any utterance x is the locus of analysis – we are comparing other utterances to it. For example, we will describe an entropy measure of the utterance x based on some sample from another population. x can be thought of as a single, uninterrupted, discursive unit. In most cases, x is a single sentence though in some cases it may be comprised of multiple successive sentences in some discourse under analysis. For readers more acquainted with conversation analysis, we can think of x as a single turn in conversation.

An utterance x can be further deconstructed in the course of an analysis. The term i will refer to the i^{th} token in an utterance x.

We compare the utterance x to another utterance y. The utterance y is sampled from a corpus of utterances, $y \in Y$. Like x, the utterance y can also be deconstructed into a number of constituent tokens. Thus, when applicable, let the term j refer to the j^{th} token in an utterance y.

As we will discuss shortly, the framework we describe will rely in part on word vectors (also referred to as *embeddings*) generated by a contextually aware word vector model. The *set* of word vectors for every token in an either utterance x or y generated by a contextually aware word vector model, will be denoted via E. This means that E_x contains the word vectors for every token in the utterance x, and E_y contains the word vectors for every token in the sample y. The word vector for the i^{th} token in the utterance x is denoted as E_{xi}, and the word vector for the j^{th} token in the sample y is denoted E_{yj}.

A quantitative definition of convergence

By casting convergence as a sort of mutual information, it is possible to devise a formal framework to measure it. Imagine that you are in the midst of a conversation with somebody you know and you are listening to them speak. You may ask yourself, “do I and this particular interlocutor sound alike?” followed by, “do we talk about things the same way as one another? Do we think about things in similar ways?” These considerations may unfold quickly and naturally while you listen with rapt attention to the person you are conversing with, scanning what they just said for similar words, mutual turns of phrase, or even whole ideas you might share.

We could measure this convergence through the act of assessing the linguistic output of a speaker and giving a “yes” or “no” answer to the question of whether what they’ve said is conceptually similar between them and some other set of speakers (such as ourselves). This is because of the way that alignment/convergence shows up in the speech that people *produce* in general. Put simply, the longer our history with you and the greater the affinity we feel for each other, the greater the probability that whatever word you speak next will be pulled from some previous snippet of conversation we have engaged in Brennan, Galati and Kuhlen (2010).

This intuition motivates what could be described as a Bernoulli principle of convergence, and can be given expression in probability – a Bernoulli distribution. We will do this with another example. Imagine you are interested in measuring the degree to which some interlocutor who spoke some utterance x exhibits conceptual convergence with the ideas put forth in the utterances of some group g. We can measure this because we have both a record of the utterances $y \in Y$ made by the members of the group g, and the utterance x for reference. If convergence is marked by an increased predictability of the conceptual content in the utterance x based on what one has previously observed in the history Y from the group g, then we can define the following experiment to test the hypothesis that there is evidence of convergence. Here are three steps to test that hypothesis: (1) take the utterance x, (2) sample a number of individual utterances y from the total distribution of utterances Y produced by group g and then (3) ask a team of annotators to each read x, read one of the samples y, and report back to you if the concepts described in x showed up in the sample the annotator read. A particularly fastidious researcher may even formulate their instructions to the annotators in (3) as "for each token (i) in the sentence x, tell me if someone in the sample y used the same word, or a synonym for it, in the same way that it was used in x."

If we follow this logic, for each token i in x we can treat the question of "did this concept show up in this sample?" as a Bernoulli process with a rate of success p whose value we are estimating by taking a number of samples from the population. Bernoulli processes, and thus the Bernoulli distribution, is particularly useful in answering simple yes-no questions and understanding the rate at which a "yes" answer occurs in individual experimental trials. We calculate p using a binomial distribution by taking a number of samples from a population y (see step (2)). In the simplest case, we can assume that the probability of encountering any one token i's meaning from an utterance x within a single comparative sample is Bernoulli distributed[1]. For reference, the Probability Mass Function (PMF) for a Bernoulli distribution is as follows:

$$P(k) = p^k (1-p)^{1-k} \tag{1}$$

If we modify the PMF for a Bernoulli distribution to fit the conditions of the experiment we have described in the first paragraph of this section, the PMF could be updated to look like this:

$$P(xi \in y) = p^{\delta_{xi\in y}} (1-p)^{1-\delta_{xi\in y}} \tag{2}$$

where $\delta_{xi\in y}$ is the Dirac Delta function and returns a Boolean value $\{0, 1\}$ based on whether or not the condition $xi \in y$ was met in the given sample – that is, whether or not some i^{th} concept/token pairing in x was reportedly found in the sample y.

We can further eliminate the term for the reciprocal of the rate p. In the hypothetical experiment we've described, we only care about the rate that annotators find for positive evidence that a term i or a similar term with a similar meaning is used in y. Thus, let $(1-p)^{1-\delta_{xi\in y}}$ be subsumed into the proportional constant:

$$P(xi \in y) \propto p^{\delta_{xi\in y}} \tag{3}$$

From this vantage point, we can now go one step further and ask how much of x could you predict if you read the sample y first? If it turns out that you could indeed learn a lot about x by reading y, then this would be evidence that the author of x likely shares some characteristics of their mental model for the topic of conversation with the speaker who authored the utterance y. To answer this question, we take the Eq. 3 and use it to calculate the Shannon entropy for the distribution of each concept/token pairing in x when compared to the distribution of concepts in any sample y. This yields how easily one could predict x based on information known from y. The entropy of any communicative act is a simple though foundational measurement as defined by Shannon (1948). Shannon entropy decreases as "uncertainty" in a "code" decreases (i.e., the elements of the code become more predictable)[2]. Entropy increases when the opposite occurs – when the elements of the code become less predictable. A similarity measurement can be derived from Shannon entropy by calculating the predictability of a code relative to some expectation, such as a collection of words in a given context. For example, given a set of prior words used by a group of individuals, a member of that group would have greater linguistic similarity if the words of that individual could be predicted by those of the group – their word choices would have lower relative entropy with those of the rest of the group. The Shannon entropy of Eq. 3 can be expressed via the following formal equation:

$$H(x; y) = -\sum_i p^{\delta_{xi\in y}} \left(\delta_{xi\in y}\right) \log p \tag{4}$$

Ideolectical differences between group members in the exact meaning of a word renders a 1:1 matching of the word's usage

[1] However if we repeated this process for each word, the probability for each element of any utterance x based on a comparative sample can be thought of in terms of a multinomial distribution. For simplicity, we describe the sampling process in binary terms (yes/no between some x and some set y), and thus focus on the Bernoulli distribution for the rest of this paper.

[2] Entropy also decreases as two distributions become increasingly polar. If the probability of some term i is 1 in distribution a but the probability of i is 0 in distribution b the entropy of $P(i|a)$ and $P(i|b)$ is 0. This is because you can easily predict i in a by knowing that i in a is the opposite of what you have observed for i in b. Realistically, it is not clear how *this* condition could be met in language, however, and research shows that the language models that we will use to estimate entropy have a baseline similarity between any two randomly sampled word vectors $> 0.$ and < 1. Ethayarajh (2019) thus rendering this condition impossible with our specific method.

and meaning between group members impossible, however. In other words, you are almost guaranteed to never see a word i used in exactly the same way by two different speakers, even from the same group g. People do not parrot their conversational partners. They add to the discourse by building on similar concepts, but according to their understanding of them. To formalize this ideolectical difference, we add a noise parameter ϵ_y to $\delta_{xi \in y}$ which quantitatively implements the idea that there is no perfect "hit" where two usages of the same term (or a synonym) are exactly identical and that we expect noisiness between the two usages of a term (or a synonym for it) in intra-group communication. In simpler terms, ϵ represents a base level of difference between how members in a group use a particular set of terms, and thus the conceptual structures that those terms evoke. We thus get the following Eq. 5 for the entropy of a noisy Bernoulli distribution.

$$H(x; y) = -\sum_{i} p^{(\delta_{xi \in y} - \epsilon_y)} \left(\delta_{xi \in y} - \epsilon_y\right) \log p \qquad (5)$$

This measurement of entropy, thus, is a *Convergence Measurement*. We can use it as a continuous value representing how much convergence is observed between an utterance x and an utterance y. In the following section, we will outline how to analogously estimate $H(x; y)$ efficiently using tools developed in the field of NLP.

Computational model

The way we defined convergence in the last section is useful in that it allows us to theoretically craft a quantitative measurement. In practice, however, it would be difficult to recruit sufficient annotators and train them to accomplish the task as we've described it. Additionally, in order to automate analyses without the use of computational linguistics methods one would likely need to generate a substantial amount of normed data where words would need to be annotated for the various concepts that they might be used to invoke. This process would likely yield data scarcity issues, especially as the linguistic behaviors of groups change over time. Instead, by framing the question of how to measure convergence in terms of the entropy of a noisy Bernoulli distribution, we can efficiently and closely estimate its value using some existing tools from the world of NLP. More specifically, we can estimate the entropy value of the noisy Bernoulli using word vectors (also referred to as "embeddings") generated by Transformer Language Models like BERT or any of the Large Language Models that currently exist.

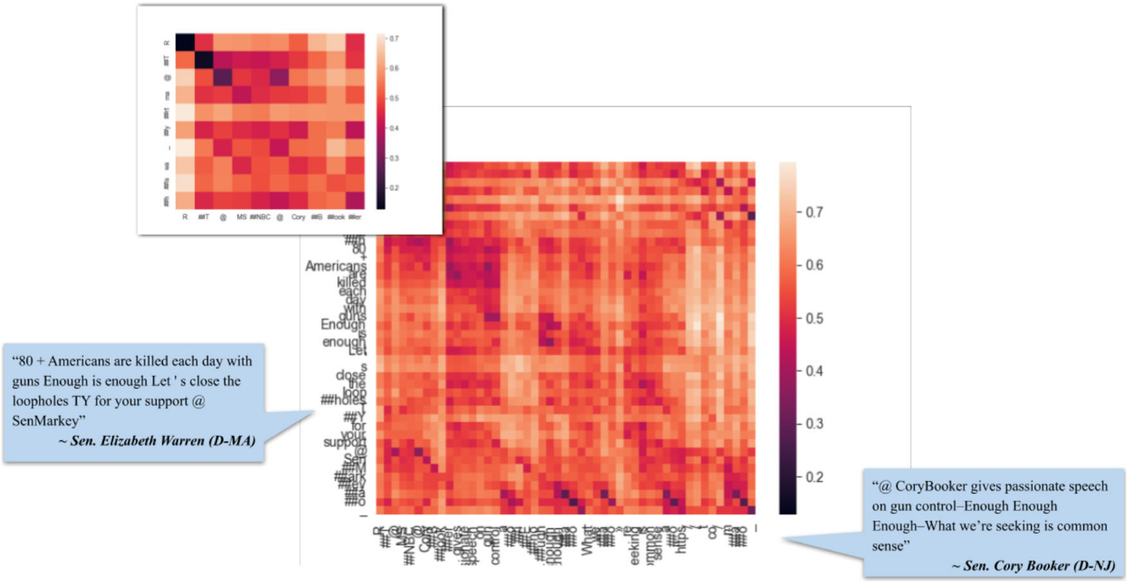
We start by converting all of the tokens in both x and y to BERT word vectors (Devlin, Chang, Lee & Toutanova, 2019). This will allow us to capture similarity between tokens that are semantically similar but are not a 1:1 mapping of the same word. Let E_{xi} be the set of BERT word vectors for each token i in a sentence x and E_{yj} be the set of BERT word vectors for each token j in a sample y of utterances from a group. The Eq. 6 below shows the process of converting tokens $i \in x$ to word vectors. Tokens $j \in y$ are converted to word vectors via the same process.

$$E_{xi} = BERT(i \in x) \qquad (6)$$

The utility of word vector models is that they represent the meaning of words in a geometric form and in ways that have been shown to accurately reflect human semantic cognition (e.g., Dumais et al., 1996; Jones and Mewhort, 2007) . Even in contextually uninformed models, words that are semantically similar to one another based on their word vectors cluster closer together in word vector space (Pennington, Socher & Manning, 2014; Mikolov et al., 2013; Devlin, Chang, Lee & Toutanova, 2019). In contextually aware models like BERT (and all subsequent transformer models) words that have similar word senses cluster separately from other word senses. This allows us to make fine-grained distinctions between the different meanings of polysemous words like the many meanings of "bank," but it also allows us to capture subtle community/group-specific differences in word usage like the differences in the use of the word "slay" in example 4 (Devlin, Chang, Lee & Toutanova, 2019). Put simply, if a word vector represents the meaning of a word as a point in space, words that are more semantically related to one another will be closer to one another. And if those word vectors are generated by a contextually aware model the closer *the word senses* of two words are to one another the closer two word vectors will be to one another in vector space. A popular way to measure the proximity of two word vectors to one another is to use cosine error (CoE), where a CoE value of 0 indicates that the word vectors for two words in high-dimensional space are in a superposition of one another, and 2 means that they are maximally divergent. For reference, a visual representation of this step is provided in Fig. 1(a) at the end of this section.

However, *proximity* in vector space is different from a probability, and CoE values are just a scaled measurement of proximity, not the probability that two vectors are the same or similar. An additional step is needed to render CoE values as probabilities that can be used as part of a statistical framework. While our method is markedly different from theirs in both its conceptualization of how to measure convergence, and the assumptions we make[3], we follow the same

[3] The method described in Rosen (2022) has two major differences from the current method. First, their measurement of convergence requires that the data contains samples from two groups with all individuals pre-



"80 + Americans are killed each day with guns Enough is enough Let ' s close the loopholes TY for your support @ SenMarkey"
~ Sen. Elizabeth Warren (D-MA)

"@ CoryBooker gives passionate speech on gun control–Enough Enough Enough–What we're seeking is common sense"
~ Sen. Cory Booker (D-NJ)

(a) Heatmap of Cosine Error (CoE) values between tokens for two separate tweets (indicated in blue text boxes along the axes) written by different authors.

Token	$\min_j\,(CoE(E_{xi}, E_{yj}))$	$P(E_{xi}\|E_y)$	$H(x_i; y)$
...	...	...	...
80	.42	.95	.05
+	.45	.92	.07
Americans	.37	.98	.01
are	.38	.98	.05
killed	.42	.94	.10
each	.48	.89	.10
day	.47	.90	.09
with	.39	.91	.09
guns	.39	.96	.03
...	...	...	...

(b) Operation of moving from minimum cosine error (CoE) value for each token in the first column to an entropy value in the final column, following the steps outlined in equations 7-9.

Fig. 1 Visualization of steps for our computational model as outlined in "Computational model"

principle described in Rosen (2022) to convert CoE values to probabilities. To convert CoE to a probability, we leverage a half-Gaussian distribution, continuous on an interval of $[0, \infty)$, with two parameters: (1) a location parameter $\mu = 0.0$ such that as the CoE value for the comparison of two word vectors approaches 0 we have maximum confidence that the two words mean the same thing, and (2) a scale parameter σ that sets a penalty weight for CoE values farther away from 0[4].

$$P(E_{xi}|E_{yj}) = P_{\mathcal{N}_{[0,\infty]}}\left(CoE(E_{xi}, E_{yj})\middle|\mu = 0., \sigma\right) \qquad (7)$$

However, we almost never have a reason to compare any one vector from a sentence i to every single vector fromanother

labeled according to their group status in order to show that rhetoric is internally consistent within groups and inconsistent outside of them. Our current method does not require the presence of multiple groups in order to measure convergence. Second, their method requires the preselection of some set of key terms for analysis. Our method treats every word in an utterance as a unique experiment, and thus does not require any predefined lexicon in order to capture convergence.

[4] By no means is the use of a Gaussian Distribution the only way of converting a Cosine value to a probability. For example, one could go so far as to use $\frac{1+CoS(E_{xi},E_{yj})}{2}$ to convert scalar Cosine Similarity (or CoS: which is the reciprocal of CoE) values to a ratio in terms of maximum similarity. We prefer the use of a Gaussian distribution here as a means of increasing the burden of proof required to claim two words mean the same thing based on the proximity of their word vectors, because of the way that the scale parameter σ can be used to increase penalties on word vectors that are dissimilar to one another.

sentence/distribution, j. After all, the question we are trying to answer as described in the previous section is "for each token (i) in the sentence x, I want you to tell me if someone in the sample y used the same word, or a synonym for it, in the same way that it was used in x." The question is not if every token j in y is similar to the token i, it's whether *any* j is similar to i. Thus, it's both a better match to our question and more computationally efficient to compare the token i to only the token j in y that is the most similar to i. To do this, we take the probability of a token i from the sentence x and the token j from y that has the lowest CoE with i. This effectively replicates the hypothetical study participant in the example given in the previous section whose job is to decide whether each word/token i in the utterance x has a semantic correlate in some word/token j in the utterance y. Furthermore, if nothing in the distribution y is semantically similar, nor embedded in a similar context as i is in x, then the minimum CoE value will be high and thus indicates that the token i doesn't have anything approximating a similar term or usage in y. We thus rewrite Eq. 7 as follows:

$$P(E_{xi}|E_y) = P_{\mathcal{N}_{[0,\infty]}}\left(\min_j \left(CoE(E_{xi}, E_y)\right)\middle|\mu = 0., \sigma\right) \tag{8}$$

From the perspective of a transformer language model like BERT, the only way that the function $\min_j \left(CoE(E_{xi}, E_y)\right)$ can approach 0 is if there exists some overlapping, similar context surrounding some token in x and another token in y. Thus, in most cases there are three potential phenomena that increase the probability of $P(E_{xi}|E_y)$. Either (1) there exists a number of lexical items in $j \in y$ that tend to be semantically similar to i based on their context, such that any sample from the distribution y will likely contain items that maximize $P(E_{xi}|E_y)$ (i.e., things that are semantically similar to i are common in the distribution y), (2) the distribution y influenced the construction of i in x or vice-versa (i.e., x can be found in sample y which should be avoided at all costs), or (3) the sample y is large enough that something semantically approximating i eventually shows up in the data by sheer chance (which necessitates careful selection of an appropriate sample size). Because group members actively seek to increase similarity between each other's ideolects in intragroup communication (CAT: Gallois, Gasiorek, Giles and Soliz 2016) (1) and (2) are more likely than (3) as long as sample sizes are constrained[5].

Using this probability calculation, we obtain the entropy for the entirety of an utterance x, by comparing the vectors for words/tokens i (i.e., all $i \in x$ or i) and the distribution y.

$$H(x; y) = -\sum_i P(E_{xi}|E_y) \log P(E_{xi}|E_y) \tag{9}$$

Equation 9 efficiently estimates the convergence measurement described in Eq. 4. A visual representation of the steps described in Eqs. 8 and 9 is provided in Fig. 1(b).

This process is related to operating over a similarity matrix, such as a recurrence matrix in recurrence quantification analysis (RQA, Angus et al., 2012; Dale and Spivey, 2005; Dale, Duran and Coco, 2018). This entropy calculation compares lexical patterns in a way that is similar to RQA, token by token across utterances, and can similarly uncover relative dynamics, such as that words with similar temporal contexts have more in common with one another. The more words with similar contexts there are, the greater the similarity between two texts. This is similar to several metrics in RQA that describe how speakers show sequential similarity in time. For example, we might test how entropy is minimized is when the maximum values of $P(E_{xi}|E_{yj}) \log P(E_{xi}|E_{yj})$ form longer sequential relationships between speakers (known as "maximum line length" in RQA: Dale, Duran and Coco 2018). We revisit these relationships across methods in the conclusion.

The current model, however, differs from RQA on two important axes: (1) it formalizes similarity between texts in a way that is not bound to the same central focus on replication of lexical patterns in exactly the same order of elements. In other words, it is possible for entropy to be minimized when lexical items with similar semantic meaning are spread across many spans or even possibly in an entirely different order when comparing utterances to one another, rather than needing to be locally adjacent to one another. (2) Secondly, by using BERT word vectors, the current model allows for fuzzy matching between words based on semantic similarity across whole spans rather than in just local, adjacent contexts. Despite these differences, this comparison is useful to consider, and the current model is a close methodological cousin to RQA (cf. Angus et al. 2012).

A complete Python package implementation of this model has been made publicly available in a GitHub repository and can be accessed through the following OSF Repository: https://bit.ly/bertsofafeather. Included on the first page of the GitHub repository is a short guide titled "example.ipynb"

[5] In truly unconstrained cases where one is comparing utterances to one another irrespective of interest in any one lexical item – i.e., comparing all sentences that invoke a specific phrase like "forced birth" – one should look for smaller sample sizes but a greater number of random samples taken in order to characterize the possible diversity of utterances in the data.

which is designed to get researchers started with using the measurement described here.

Case study 1: Semantic convergence in interpersonal dynamics

In a first demonstration, we illustrate that this method can capture semantic convergence between two or more conversation partners. Prior research on interactive language suggests that alignment between words and phrases used by interlocutors can index distinct characteristics of their dialogue. Such alignment is often measured as the similarity between conversants in the form of words, concepts or syntactic patterns used in dialogue (Duran, Paxton & Fusaroli, 2019). This alignment can index interaction style and quality in language learning (Dale & Spivey, 2005) and language use Reitter and Moore (2014), and can offer insights into particular interactive contexts such as in problem-solving teams Tolston, Riley, Mancuso, Finomore and Funke (2019) and healthcare interactions (Angus, Watson, Smith, Gallois & Wiles, 2012). Our goal in this analysis is to use a publicly available corpus of conversation and show that conceptual convergence characterizes these interactions. We also model convergence based on who is talking and at what point in time. Results show that who is speaking and when have strong effects on semantic convergence.

Data

The CABNC corpus is a transcribed segment from the British National Corpus that contains natural conversations (Albert, de Ruiter & De Ruiter, 2015). The CABNC contains over 1,400 conversations, which vary in length and topic.[6]

We filtered the CABNC corpus to conversations within a given length range so that we analyze a relatively tractable transcript for each, but also because we can conduct the same analysis across conversations to build a baseline ("virtual pairs"). After filtering, only conversations between 100 and 200 turns were chosen so as to ensure that the baselines would represent approximate beginning and end of conversations when aligned. If a baseline transcript was shorter than a referent transcript, the referent transcript was shortened to the length of the baseline. A total of 225 transcripts satisfied the length requirements.

Analysis

First, we conducted convergence analysis as described above using the CABNC conversations as the "observed" data. We then performed convergence analysis between each observed transcript and a separate transcript from another conversation. This offers an additional statistical "virtual pair" baseline.

Table 1 LME coefficients, t values, and p values for convergence

Var	Coefs.	Stat	p
(Intercept)	.2637	2.265e+02	$< 10^{-9}$
Self	-.0107	-1.881e+01	$< 10^{-9}$
Baseline	.02831	5.592e+01	$< 10^{-9}$
Distance (k)	2.51e-03	4.175e+01	$< 10^{-9}$
Self x Baseline	.01069	1.272e+01	$< 10^{-9}$
Self x k	4.767e-04	5.005	$< 10^{-9}$
Baseline x k	-2.576e-04	-2.962e+01	$< 10^{-9}$
Three-way interaction	4.376e-04	-3.096	.00098

We analyzed three properties of observed and baseline convergence measures. First, we ascertained the impact of temporal remoteness of the convergence comparison (how far apart, k, conversational turns were). We assessed convergence when the speaker was the same ("self") for both conversational turns (i.e., t and $t + k$ were the same speaker). And then we assessed the impact of the baseline – convergence measures when x and y comparisons cut across different conversations. We look across up to $k = 10$ conversational turns for convergence.

Results

To test the contribution of these variables, we used a linear mixed effects model to predict convergence (H) from k, self (whether it is the same person speaking both compared utterances), and baseline with a random effects structure that included transcript and the identity of the two speakers. Note that it is possible for "self" to still be true for the baseline comparison, because as convergence analysis is conducted across lags k, it remains possible that the same speaker can be compared.

Because very short conversational turns can yield unstable estimates of convergence, we restricted the analysis only when both turns had at least five tokens or more. This resulted in 341,461 comparisons across 225 conversations, with 184,036 in the observed case, and 157,425 in the baseline comparisons (Table 1).

When we look at the effect of k, we see that close to $k = 0$, H seems to anomalously rise, suggesting that there may be a drive to avoid "repetition" near local conversational turns. To test for this, we used the stringsim function in R's stringdist library. This is akin to Levenshtein distance, but it handles strings of different length by recycling symbols and recovering a value between 0 and 1. Indeed, this measure over

[6] https://github.com/saulalbert/CABNC

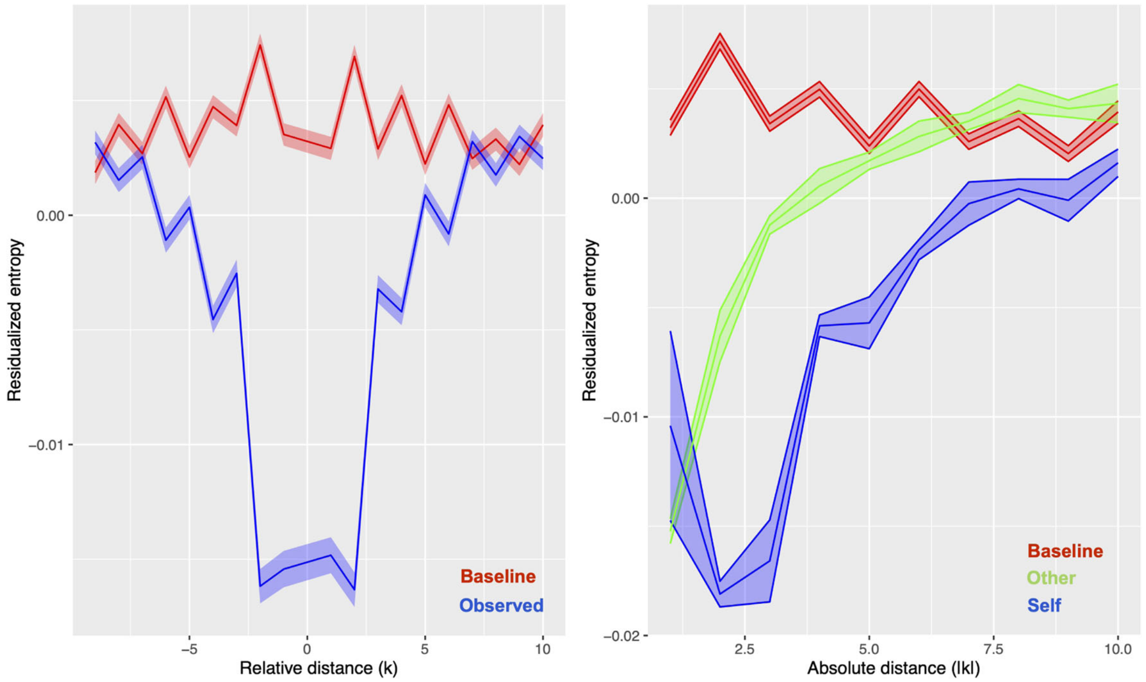


Fig. 2 Left: Average residualized entropy between utterances in dialogue controlling for string distance for baseline and observed; Right: average residualized entropy by $|k|$ showing the effect of "Self" vs. "Other," that "Self" exhibits the lowest entropy before all gradually rise (remoteness in time increases). Ribbon width is SE

our turn comparisons t and $t + k$ is significantly correlated with H: $r = -.34$, $p < .0001$.

So we residualized our convergence analysis by factoring out this string distance, and returning a residual variable H_r. Though there are limitations to such residualization (Wurm & Fisicaro, 2014), in general we can interpret low and high values of H_r as high conceptual convergence vs. low conceptual convergence, respectively. This is because the string comparison function, related to surface similarity of strings t and $t + k$, has been factored out. Indeed, the result yields a still reliable linear mixed effects model as described above:

Table 2 LME coefficients, t values, and p values for residualized convergence

Var	Coefs.	Stat	p
(Intercept)	-.01412	-1.259e+01	$< 10^{-9}$
Self	-8.644e-03	-1.599e+01	$< 10^{-9}$
Baseline	.01593	3.297e+01	$< 10^{-9}$
Distance (k)	2.121e-03	3.698e+01	$< 10^{-9}$
Self x Baseline	.01377	1.718e+01	$< 10^{-9}$
Self x k	3.933e-04	4.328	.00001
Baseline x k	-2.336e-03	-2.816e+01	$< 10^{-9}$
Three-way interaction	-6.212e-04	-4.607	$< 10^{-9}$

Discussion

Our results show that the convergence measurement described in this paper captures semantic convergence in dialogue, replicating prior work. Specifically, we provide evidence that convergence is dynamic, with its effects being strongest between turns in close proximity to one another. This is observable in the shape of the graphs in Fig. 2a and b. Note that entropy is higher for turns farther away from the current turn (turn 0) in both the past and in subsequent turns. Results from our LME model for the variable "distance" (k) (indicating the distance between utterances as measured in k-turns) also indicate that entropy between semantic structures sharply increases when comparing two turns in a conversation that are temporally more distant from one another. This is particularly salient after residualizing for string similarity further shows that the semantic convergence as captured here is robust especially when controlling for speakers' aversion to repeat themselves and one another.

As expected these effects disappear when cutting across conversations, and convergence is stronger the closer two turns are to one another in the same conversation. This is verified in the results from our LME model as shown in Table 2. Note there that the variable "baseline" (i.e., when comparing the semantic content of a turn from one conversation to that of a turn from a different conversation) is predictive of higher entropy when two turns are from different conversations. It is well established that individuals engaged in conversation should show conceptual convergence with respect to their current context (Angus et al., 2012).

There are a few additional variables and interactions that were significant predictors of entropy scores worth noting. First is the variable "Self" (whether two utterances were written by the same person) in Table 2. This was associated with lower entropy in the same conversation, and higher entropy in different conversations ("Self x Baseline"). One would expect this to be the case as individuals are more likely to repeat similar conceptual structures (namely, when compared to themselves). Turns by the same person farther away from the current turn still predict higher entropy ("Self x *k*"), which again is expected if convergence is dynamic and stronger across local turns.

The current case study demonstrates the utility of our convergence measurement in dyadic conversation but leaves out any effects arising from social or para-linguistic factors. In the following case study, we demonstrate that our convergence measurement can be used to capture the impact of these additional factors on the convergence behavior of interlocutors in more complex ensembles of speakers.

Case study 2: Convergence in internet communities

Having shown that our model captures convergence in dyadic communication, we now turn our attention to group communication practices in Internet-based communities. Previous research in CAT has shown that semantic convergence is influenced by a number of social factors (Soliz, Giles & Gasiorek, 2021), including social status (Bradac, Mulac & House, 1988; Danescu-Niculescu-Mizil, Gamon & Dumais, 2011), demographic differences (Hilte, 2023; Adams et al., 2018; MacIntyre, 2019), perceived popularity of ideas presented during ongoing discourse (Soliz, Giles & Gasiorek, 2021; MacIntyre, 2019; Gallois, Ogay & Giles, 2005; Keblusek, Giles & Maass, 2017), and sub-networks formed within groups (Dougherty, Mobley & Smith, 2010; Dougherty, Kramer, Klatzke & Rogers, 2009). Reddit is divided into a number of smaller communities called "subreddits". Subreddits have several features that render them a useful microcosm for group communications research: Reddit data is richly annotated for things like the intragroup popularity of posts and comments, and it naturally captures the back and forth between users due to the way users respond to each other's comments. In addition to these factors, Previous research has shown that subreddit communities exhibit differences in user demographics (Shatz 2017 and implied by Roozenbeek and Salvador Palau 2017), and lexical usage patterns (Hamilton, Clark, Leskovec & Jurafsky, 2016; LaViolette & Hogan, 2019; Krendel et al., 2021; Park & Conway, 2018; Johns, 2021) and topical focus (Stine & Agarwal, 2020; Barker & Rohde, 2019). Each subreddit on Reddit operates like a sub-culture with varying pressures on communication strategies used by its members.

Our objective in this case-study is to show how our measurement can be used to test hypotheses about social factors that affect the degree of convergence between individuals in these groups. Based on data availability, we test how influential the following social factors are in enhancing convergence: whether or not two comments are embedded within the same conversation (indicated via "same-post", and mirrors the baseline condition in "Case study 1: semantic convergence in interpersonal dynamics" , the salient popularity of the ideas expressed in a given comment (measured in a comment's "upvotes" or "comment-ups"), the identity of the individual who wrote the comment *x* (indicated via "x – user"), the identity of the individual who wrote the comment *y* (indicated via "y – user"), and the absolute time difference between when comments are posted to a submission/post ("t-delta-abs"). Table 3 lays out our hypotheses about each of these variables in detail.

In this case study, we focus on two subreddit communities: r/MensLib and r/MensRights. Both are embedded in what researchers have come to call the "manosphere" – a network of online communities interested in men's gender ideologies (Ribeiro et al., 2021). A sizable portion of the manosphere openly expresses misogynistic ideologies (Ribeiro et al., 2021; Krendel et al., 2021; Khan, 2019; LaViolette and Hogan, 2019 though r/MensLib may be an exception). To date, some of these groups have acted as incubators for acts of domestic terrorism (Male Supremacy, 2021). There are two a priori reasons to focus on these groups. First, previous research has indicated that r/MensLib and r/MensRights represent opposing views of men in society: r/MensLib allegedly applies a feminist framework to issues of social pressures on men. In opposition to this, r/MensRights adopts a "male supremacist" ideology (Krendel et al., 2021; LaViolette & Hogan, 2019; DiBranco, 2020) stemming from its roots in men's rights activism (MRA).

Additionally, we include data from the feminist subreddit r/Feminism to evaluate prior claims that r/MensLib is more closely aligned with other feminist communities than it is with other subreddits from the manosphere (LaViolette & Hogan, 2019). Second, prior research has focused on gross level differences in the rhetoric espoused by these groups, but no work to date has focused on understanding the social factors that influence the uptake of rhetorical structures used by group members. Thus, our case study of convergence within r/MensLib and r/MensRights can have immediate impact on research into these communities and others like them by showing the ways in which the rhetorical similarity between members is influenced by various social forces that may be at work.

In the remainder of this section, we will (1) go through the data used to complete this study, (2) the specific tests we ran to test our hypotheses, (3) report on the results of our tests, and (4) discuss briefly some implications of those results.

Data

We use the Python PRAW Reddit package to index all comments posted to r/MensLib, r/MensRights, and r/Feminism from the top 3 posts that mention the term "women" for the month of April (2023). We specifically focused on posts that invoked the word "women" as opposed to including other gender terms based on observations in previous work that indicate that r/MensRights uses the terms "women", "woman", and "girl(s)" map to different usage patterns in corpora (Krendel et al., 2021). Thus, in order to provide a more direct comparison across groups, we focused on the single term "women". We analyzed every comment made to each post starting from the day the post was first written.

A summary of the total number of comments and the total number of tokens pulled from each subreddit is included in Table 4.

Analyses

We first converted all comments collected to word vectors using the "bert-base-uncased" model freely available in the HuggingFace Python library (Wolf et al., 2020). We then measured convergence between every comment collected to all other comments – excluding comments written by the same author – using our convergence measurement implemented in PyTorch. We set the scale parameter $\sigma = .3$. Because comments can vary greatly in length, we controlled for the effect of length by averaging the entropy returned by our convergence measurement for all the tokens in the comment x.

Table 4 Comments and total tokens for each subreddit represented in our web-scraped corpus

Subreddit	Comments	Total tokens
r/MensLib	573	47696
r/MensRights	1196	70420
r/Feminism	409	32436

Table 3 Description of all hypotheses and their associated variables for social factors that may influence convergence within r/MensLib and r/MensRights

Variable	Hypothesis
"same-post[T.True]"	*Whether comments x and y are in same post (Boolean):* We predict that comments made within the same post will have lower entropy with one another in the same way that one would expect comments made in the same conversation to have lower entropy with one another.
"x-comment-ups"	*The total number of up-votes for the comment x (Integer):* As a salient marker of how popular a particular way of framing an idea is, we predict that the greater the number of up-votes a comment x receives, the lower its entropy will be with other comments in general. This is because individuals tend to converge towards concepts and ideas that are marked as more popular by other group members (Bradac, Mulac & House, 1988; Danescu-Niculescu-Mizil, Gamon & Dumais, 2011)
"y-comment-ups"	*The total number of up-votes for the comment y (Integer):* As a salient marker of how popular a particular way of framing an idea is, we predict that the greater the number of up-votes a comment y receives, the lower its entropy will be with the comment x. This is because we expect the author of the comment x to tend to converge towards concepts and ideas that are marked as more popular by other group members (Bradac, Mulac & House, 1988; Danescu-Niculescu-Mizil, Gamon & Dumais, 2011).
"t-delta-abs"	*The absolute difference in time between the comments x and y (Integer, Unix Time):* We predict that comments that are written farther apart from one another in time will exhibit lower convergence than those comments that are closer to one another . . .
"1 | x-user"	*The identifier for the user who wrote the comment x (Categorical, Random Effect):* We predict that individual users will have different rates of convergence with other users based on prior empirical research in CAT (Jones et al., 2014; Xu & Reitter, 2015).
"1 | y-user"	*The identifier for the user who wrote the comment y* (Categorical, Random Effect): We predict that individual authors of any comment x will show variable rates of convergence with other individuals (the author of the comment y) based on either the presence of additional sub-networks, or factions, within groups or differences in esteem granted to other speakers from within the group. This has been found to be the case in other studies of intragroup communication (MacIntyre, 2019; Soliz, Giles & Gasiorek, 2021; LaFree et al., 2016) that share similarities with MRA. In this latter case, differences arise from competition between group members (LaFree et al., 2016; Velásquez et al., 2021)

Table 5 LME coefficients, t-values, and p-values for r/MensLib

Var	Coefs.	Stat	p
Intercept	.04236	6.094e+01	$< 10^{-9}$
same-post[T.True]	-6.296e-04	-8.294	$< 10^{-9}$
x-comment-ups	1.418e-05	2.055	.03984
x-comment-ups:same-post[T.True]	3.446e-06	2.242	.02496
y-comment-ups	-1.261e-05	-1.011e+01	$< 10^{-9}$
y-comment-ups:same-post[T.True]	-8.154e-06	-5.51	$< 10^{-5}$
x-comment-ups:y-comment-ups	-2.605e-08	-.7719	.4402
x-comment-ups:y-comment-ups:same-post[T.True]	3.909e-08	1.060	.289
t-delta-abs	2.070e-10	1.933	.05324
same-post[T.True]:t-delta-abs	1.11e-09	4.875	$< 10^{-5}$
x-comment-ups:t-delta-abs	5.865e-12	2.330	.01979
x-comment-ups:same-post[T.True]:t-delta-abs	-1.225e-11	-2.498	.01249
y-comment-ups:t-delta-abs	-1.153e-11	-4.726	$< 10^{-5}$
y-comment-ups:same-post[T.True]:t-delta-abs	1.962e-11	4.026	$< 10^{-3}$
x-comment-ups:y-comment-ups:t-delta-abs	-1.711e-14	-.2202	.8257
x-comment-ups:y-comment-ups:same-post[T.True]:t-delta-abs	-2.287e-12	-4.079	$< 10^{-3}$
1 | x-user	-3.006e-07	-.2792	.7801
1 | y-user	5.399e-07	9.728	$< 10^{-9}$
Group Var	.6634	1.655e+01	$< 10^{-9}$

To calculate the time difference ("t-delta-abs") between comments, we take the absolute difference between the time for the comment x and the time for the comment y.

We tested the impact of each social factor/variable by attempting to predict the average convergence measurement conditioned on the social factors/variable being testing. We used a linear mixed effects model implemented in Python in the statsmodels package. We grouped comments according to the comment ID ("comment-id") for the comment x. We performed separate analyses for both r/MensLib and r/MensRights. Our model script is below.

$$avgH \sim same\text{-}post * x\text{-}comment\text{-}ups * y\text{-}comment\text{-}ups \\ * t\text{-}delta\text{-}abs + \ (1|x\text{-}user) + (1|y\text{-}user)$$

An exploratory analysis was performed within each group to identify lexical differences between comments that exhibit higher convergence and those that do not. First, we defined comments as being either "convergent" or "not-convergent" based on whether a comment met both of the following conditions: (1) Within the same post, a comment x has lowest entropy with either the comment immediately before it or immediately after it, and (2) The comment x has lowest entropy with comments from the same post (i.e., entropy is lowest in the same post, as opposed to the baseline condition). We then used TF-IDF to extract the top five terms that were associated with convergent comments.

We test the claim that r/MensLib is more aligned with/converges more closely with other feminist ideological communities using a simple t test procedure. The inputs to this test were generate by (1) calculating the pairwise convergence measurement for all comments in r/MensLib and all comments in r/Feminism written within 24 h of one another, and then (2) calculating the pairwise convergence measurement for all comments in r/MensLib and all comments in r/MensRights written within 24 h of one another. We then use a t test of independence procedure to test whether entropy is lower and statistically significant in condition (1) when compared to condition (2).

We then tested whether there is greater convergence between r/Feminism and r/MensLib via the same procedure: (1) calculating the pairwise convergence measurement for all comments in r/Feminism and all comments in r/MensLib written within 24 h of one another, and then (2) calculating the pairwise convergence measurement for all comments in r/Feminism and all comments in r/MensRights written within 24 h of one another. If r/MensLib is more closely aligned with other Feminist leaning groups, then both of these conditions must be true: r/MensLib must have lower entropy with r/Feminism, and reciprocally r/Feminism should have lower entropy with r/MensLib. Note that both of these conditions can be independently true or false.

Results

r/MensLib

Results from our LME model are reported in Table 5. A category plot showing the average entropy for comments *y* posted within a window of 10-comments from the comment *x* is shown in Fig. 3a.

We found that the effect of comments being written to the same post to be a significant predictor of lower entropy. This can be observed in Fig. 3a, where entropy is higher in condition where comments are from different posts than they are in the "same-post" condition. This observed difference is corroborated in our LME results, shown in Table 5. We interpret this as indicating that individuals tend to stay on topic in group discussions.

We found that the effect of how many up-votes or "comment-ups" that the comment *x* received was significant and associated with higher entropy. This was surprising, as it flies in the face of our initial hypothesis. There was also a complex, significant effect of "comment-ups" for the comment *x* and being in the same post, indicating that this effect was larger for comments written to the same post than between comments written to different posts. Both of these results are shown in Table 5.

We found that the effect of "comment-ups" for the comment *y* was a significant predictor of lower entropy. This result is also shown in our analyses in Table 5. This effect was stronger between comments made to different posts.

We found that the difference in time between comments was not a significant predictor of entropy on its own. However, as shown in Table 5, several complex interactions with time differences were significantly predictors of higher entropy. Those include an interaction of difference in time within the same post (increases entropy); difference in time and the number of "comment-ups" received by the comment *x* (increases entropy); the interaction of comment-ups for the comment *x* being in the same post, and difference in time (decreases entropy); the interaction of difference in time and the number of "comment-ups" for the comment *y* (decreases entropy); and the interaction of difference in time, the number of "comment-ups" for the comment *y* and being in the same post (increases entropy).

As shown in Table 5, The author of the comment *x* was not significantly predictive of differences in entropy, but the author of the comment *y* was significant and predictive of higher entropy.

Comments that converged in r/MensLib were characterized as discussing the following five terms – "movies" (one of the posts was specifically about representations of gender norms in cinema), comments discussing the concept of what is considered stereotypical "feminine", comments discussing the concept of what is stereotypically "masculine", comments that invoke numerical quantities of "10" (though why this is the case is unclear), and comments that discuss the opinions of commenters qualified with the word "like".

r/MensRights

Results from our LME model are reported below in Table 6. A category plot showing the average entropy for comments *y* posted within a window of 16-comments from the comment *x* is shown in Fig. 3b.

We found that the effect of comments being written to the same post to be a significant predictor of lower entropy. This effect is observable in Fig. 3b, and is found to be predictive of lower entropy per our LME analysis (see Table 6). It is worth noting that in both our LME results and as seen in Fig. 3b, this effect is weaker, though it is present and significant. This indicates that individuals tend to stay on topic in group discussions.

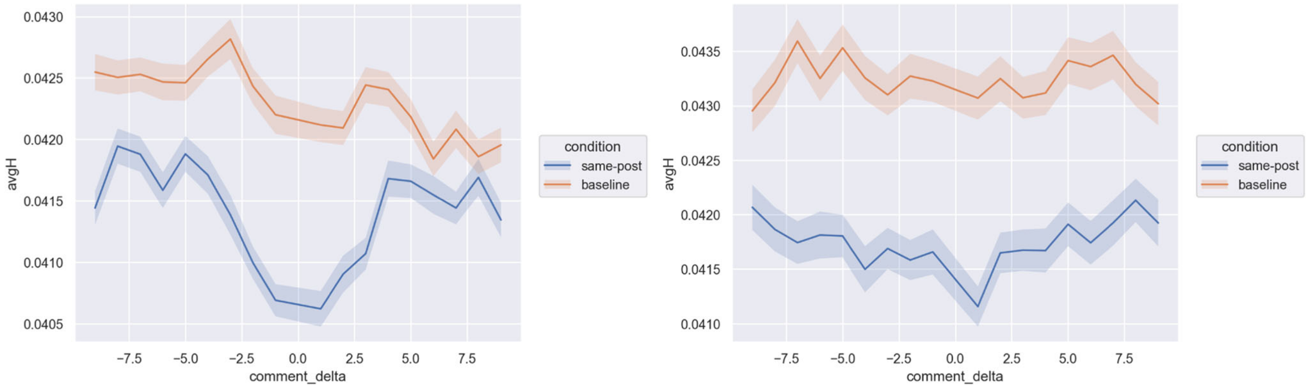
We found that the effect of how many up-votes or "comment-ups" that the comment *x* received was not significantly predictive of difference in entropy, nor were any interactions with it of note. These results are shown in Table 6.

We found that the effect of "comment-ups" for the comment *y* was a significant predictor of lower entropy. This effect was stronger between comments made to the same post (i.e., "same-post * comment-ups"). These results are shown in Table 6.

The results of our LME predictive model as demonstrated in Table 6 shows several significant interactions of the absolute difference in time ("t-delta-abs") with other variables in the data. We found that the difference in time between comments was a significant predictor of lower entropy. Several complex interactions with time differences were also significant predictors of entropy. Those include an interaction of difference in time within the same post (increases entropy); difference in time and the number of "comment-ups" received by the comment *y* (decreases entropy); the interaction of comment-ups for the comment *y* being in the same post, and difference in time (decreases entropy). Of note: this behavior was almost the opposite of what we found for r/MensLib, with the exception of the interaction of difference in time with being in the same post.

The author of the comment *x* was significantly predictive of higher entropy, and the author of the comment *y* was significant and predictive of lower entropy. These results are reported on in Table 6.

Comments that converged in r/MensRights were characterized as discussing what group members perceived to be "good", comments discussing what is required ("require") of commenters, comments discussing the groups' normative concept of "people" (typically outgroup people), comments discussing what it means to be "single" from an MRA perspective, and comments discussing setting "boundaries"



(a) r/MensLib (b) r/MensRights

Fig. 3 Category plots of average entropy for "same-post" and "baseline" conditions for x compared to comments in range $k \pm 10$ for both r/MensLib and r/MensRights

Table 6 LME coefficients, t values, and p values for r/MensRights

Var	Coefs.	Stat	p
Intercept	.04269	8.521e+01	$< 10^{-9}$
same-post[T.True]	-1.170e-03	-2.387e+01	$< 10^{-9}$
x-comment-ups	6.199e-06	.7301	.4653
x-comment-ups:same-post[T.True]	1.935e-06	1.346	.1785
y-comment-ups	-5.812e-06	-4.834	$< 10^{-5}$
y-comment-ups:same-post[T.True]	-9.21e-06	-6.656	$< 10^{-9}$
x-comment-ups:y-comment-ups	6.879e-08	1.864	.06226
x-comment-ups:y-comment-ups:same-post[T.True]	-1.682e-07	-3.706	$< 10^{-3}$
t-delta-abs	-1.096e-10	-2.812	4.918e-03
same-post[T.True]:t-delta-abs	8.458e-10	3.458	$< 10^{-3}$
x-comment-ups:t-delta-abs	-4.334e-13	-.3840	.701
x-comment-ups:same-post[T.True]:t-delta-abs	-9.506e-12	-1.105	.2690
y-comment-ups:t-delta-abs	-3.825e-12	-3.514	$< 10^{-3}$
y-comment-ups:same-post[T.True]:t-delta-abs	5.679e-12	.66	.5093
x-comment-ups:y-comment-ups:t-delta-abs	-6.755e-14	-2.004	.04507
x-comment-ups:y-comment-ups:same-post[T.True]:t-delta-abs	1.682e-11	5.672	$< 10^{-9}$
1 | x-user	1.802e-06	1.971	.04867
1 | y-user	-1.237e-06	-3.377e+01	$< 10^{-9}$
Group Var	.5579	2.29e+01	$< 10^{-9}$

(specifically as a way of establishing constraints on the behavior of sexual partners).

r/MensLib's similarity to other feminist communities

We find that r/MensLib has lower entropy (higher convergence) with comments posted to r/Feminism than it does with comments posted to r/MensRights ($t(20, 172) = -10.64$, $p < 1e^{-9}$).

We also find that r/Feminism has lower entropy (higher convergence) with comments posted to r/MensLib than with comments posted to r/MensRights ($t(10, 663) = -20.39$, $p < 1e^{-9}$).

Our results indicate that there is sufficient evidence to reject the hypothesis that the similarity between r/MensLib and r/Feminism over other comparisons is due to random chance. Furthermore, the characteristics of our results show that there is some evidence that r/MensLib shares more information in common with r/Feminism (and vice versa) than either do with r/MensRights. This provides good evidence that content from r/MensRights is conceptually more similar, and thus more closely converges with, rhetoric espoused by other feminist groups than rhetoric espoused by other groups within the manosphere.

Discussion

It was not surprising that in both groups the difference in time between comments was a significant predictor of rising entropy, specifically within the same post. As people comment to posts and the conversation evolves, we would expect that comments written in closer proximity to one another should contain more overlapping conceptual content. This fits our findings in the dyadic case study as well. We were also not surprised that the number of y-comment-ups was associated with lower entropy in both groups. Based on prior work in CAT it makes sense that people would converge towards comments that are saliently rewarded by other group members. Finding that these trends exist in both groups makes us think that these are general principles underlying the dynamics of convergence, at least within the current case study. Future work can leverage the same methodology to explore just how generalizable these observations are to communication practices in other groups and contexts. We welcome such extensions (and testing) of our work.

The differences in which social factors are predictive of differences in convergence between r/MensLib and r/MensRights are particularly intriguing. First, that the number of comment-ups a comment x receives is associated with *greater* entropy in one group (r/MensLib) and has no effect in the other (r/MensRights) may indicate that the incentive to write novel content differs between groups. r/MensLib appears to reward less predictable content (content that is divergent from group norms) with more salient rewards than r/MensRights does. That may indicate that r/MensLib is more welcoming of a wider range of views. Other notable differences in interactions include the difference in the way "y- and x-comment-ups" interact with various variables. Again, x-related factors are more prevalent in r/MensLib, while y-related factors are more prevalent in r/MensRights. This could mean that members of r/MensRights are paying closer attention to what others are saying within their ranks. These observations have not been reported in prior literature to our knowledge, and perhaps even more intriguingly we uncovered them via an easily scaled, quantitative approach.

In many ways, the results for what factors influence convergence behavior in r/MensRights are interesting and important on their own. Our findings can be partially explained by previous observations on the communicative norms in many far-right extremist organizations. In particular, the fact that lower entropy in r/MensRights is predicted by the author of the comment y mirrors observations of the fractious nature of far-right extremist organizations. In such groups members tend to split themselves into sub-factions based on perceived similarities and preferences for the rhetoric of specific, often opposing, leaders within the group (Velásquez et al., 2021; LaFree et al., 2016). In particular, LaFree et al. (2016) point out that "Of valid cases, by a large margin, far right extremist groups exhibited the highest levels of inter- and intra-group competition, at a little over 50%". The observation that entropy is mediated by the author of the comment y matches what one would expect when such fractiousness is present within a group: *who* writes the message you converge to becomes an important social indicator of whose ideas you support, and, potentially, of who you are willing to be led by.

Our qualitative description of what ideas are associated with convergence (our TF-IDF based analysis) in the two groups may appear to replicate findings in prior work, but the way we come to our findings is important. Prior work simply looked at r/MensRights and r/MensLib as separate entities and went no further than to ask what the differences between popular comments/posts from the two groups are. LaViolette and Hogan (2019), for example, looked only to the most popular content posted to both groups and used topic modeling to coarsely characterize group rhetoric. In the words of the authors, their method for comparison was to "examine which words are distinct between each group" (LaViolette & Hogan, 2019) via direct comparison of collocational frequencies. By necessity, their findings thus rely on contrasting the two groups against one another. Our findings show that *the ideas that get perpetuated by members from within these groups*, independent of any external information, are different. Whereas the methodology used by LaViolette and Hogan (2019) and others requires a comparison between groups, we naturally show that what gets passed along within these groups is different from what is glossed over without the need for external comparison. Ergo, our description of what gets converged to may be a better indicator of what characterizes the communicative norms in r/MensLib and r/MensRights.

Conclusions

Throughout the course of this study, we have provided a quantitative description of social identity signaling grounded in information-theoretic principles. We then provided a scalable computational model to test hypotheses based on this description of social identity signaling. We deployed this model in two case studies, validating the computational model, and later uncovering insights into both interpersonal and intergroup messaging dynamics.

There are certainly limitations to the current description and subsequent model. First, while the model can quantitatively describe the degree of convergence and divergence phenomena, it cannot directly specify motivations for them. Hypotheses in both case studies were based on previous observations surrounding the rhetoric of the groups studied.

Still, model outputs are useful in validating observations in a data-driven way and they add additional support to prior claims about interpersonal and intragroup communication.

Additionally, the model is not equipped to tell researchers the precise lexical differences that drive convergence and divergence. If a researcher is interested in pointing to the precise lexical patterns powering convergence behavior they will either need to retool the model to output the specific tokens that are similar across utterances, or else consult the data qualitatively (cf. our use of TF-IDF in "Case study 2: Convergence in internet communities"). However, it is possible to integrate our strategy here with more focused content analysis. The methods described here could be used as a data-mining tool for these lexical specifics. By identifying extreme differences in entropy – by finding the tails of this distribution – researchers could then target the source samples producing them as representative of the more extreme differences across groups. We leave this for future researchers to try.

This raises another similarity to recurrence analysis, which we noted earlier. In particular, conceptual recurrence plots (Angus et al., 2012) are semantic modeling done over time in interactions, and it is possible to find high regions of semantic similarity. This would indicate that conversation partners are using similar words (or meanings) at particular points in time. This can help to identify the particular word forms that are generating semantic similarity, and thus reveal the meanings used by parties in an interaction. Our method could be adapted for a similar purpose. Very high (or, conversely, very low) entropy may signal a particular form x that deviates (or converges) sharply with y. We could scan such extreme instances to mine for the underlying words (and their corresponding conceptual basis).

In spite of these drawbacks, our description and model represent two major contributions to the study of social identity signaling and language use. First, by providing a quantitative definition of convergence, we provide a strong foundation for future researchers to test their hypotheses about both of these phenomena. Whether individuals and/or groups converge or diverge over time becomes a function of testable changes in entropy. Second, deployment of our computational model can facilitate analysis of datasets that are unique to some domains (such as distinct Reddit communities) but are large enough to make it difficult to yield a bespoke semantic model for them. Our study analyzed convergence for 225 conversational transcripts (consisting of 100–200 turns per transcript) and 2178 unique Reddit comments from three different groups. We performed all of these analyses in a matter of hours as opposed to the days and weeks that would be necessary to handle this amount of data in another setting. Flexible application of a bidirectional encoder like BERT promotes vectors that adapt to such unique linguistic surroundings without having to hand-code or carry out other data cleaning or processing.

While it is not a primary focus in the case studies we have provided, there is utility in measuring convergence within groups as a window into the concept of "echo chambers" and related phenomena in web-based discourse. The methodology described in this paper can come to bear in such work in two distinct ways. First, as shown in the Reddit case study, (1) it is possible to study the social factors that lead to consolidation of individuals' views within larger group rhetorical norms. By understanding the social influences that yield greater convergence one can better understand why and how echo chambers form while focusing on the identities managed by members of a group. Additionally, (2) our entropy based measurement of convergence can also be deployed in lieu of many of the coarse grained measurements of content similarity used in studies of echo chambers and online polarization. Only a few studies interested in quantitatively measuring polarization in online communities leverage user-generated discursive data (i.e., text: Terren and Borge-Bravo, 2021). Those that do often focus on much coarser units of semantic similarity between texts generated by members of a group. Studies that leverage simple topic modeling and sentiment classification techniques as a description of textual similarity (Terren & Borge-Bravo, 2021; Villa, Pasi & Viviani, 2021), analysis of Twitter hashtags (often manually: Cota, Ferreira, Pastor-Satorras and Starnini, 2019; Garimella, Morales, Gionis and Mathioudakis,2017) or may in some more contemporary cases use sentence-level embeddings (Alatawi, Sheth & Liu, 2023)[7] will ultimately miss lexical units and patterns that are used to signal relevant identities of participants in web-based discussions.

Perhaps more importantly, our approach reinforces the importance of theory and avoiding strictly data-science based approaches to linguistic analysis. This has been recognized for some time, from the influential "Plato's problem" framing of Landauer and Dumais (1997) to more recent debates about semantic representation, as in Jamieson, Avery, Johns and Jones (2018). Our measurement and its conceptual motivation show that statistical modeling of this kind need not be theory agnostic. Theory-driven data science is a powerful paradigm for developing empirical methods for studying language and psychological phenomena, and it echoes a long-standing recognition that the massive scale of our envi-

[7] The use of SBERT as a means of measuring semantic similarity (see: Alatawi, Sheth and Liu, 2023) has a major limitation however when compared to the method proposed here. Namely, SBERT is designed and trained to classify sentences with global semantic similarity-sentences that have been labeled as meaning the same thing, irrespective of their constituent components-as more similar to one another (Reimers & Gurevych, 2019). However, it is easy to imagine a case wherein individuals from a group, engaging with one another in ongoing discourse, should be expected to author utterances that imply wildly different meanings while leaning into group lexico-semantic norms for the subcomponents of their utterances.

ronment and how the human mind learns it represent deep theoretical problems (Johns, Jamieson & Jones, 2023). When we step away from theory in favor of bigger models and massive data sets, we may lose more than we gain. There is already a robust trend in machine learning (ML) and related literatures to cover computation with a veil of theoretical agnosticism with respect to how models align with empirical studies. In many of these works, including in NLP, the focus may be on the benchmark performance rather than what an analysis says about language or social cognition. While this does not characterize all contemporary work in the field, "models" tend to function by using large neural networks that map inputs to some defined subset of outputs. Importantly, neural network models are a kind of general purpose approximator (Hornik, Stinchcombe & White, 1989). That is, you can use the same neural network design to classify documents as you could to identify the metaphoricity of a text. Thus, contemporary ML engineers often do not concern themselves with how the workings of their models align with theoretical and empirical assumptions about human data and the social processes that generate it. There are several consequences of this supposed agnosticism, not the least of which is that contemporary ML models require genuinely massive quantities of data to achieve good results (Villalobos et al., 2022; Kaplan et al., 2021; Hoffmann et al., 2022).

The measurement proposed in this paper is not simply an ML classifier. This is, of course, by design. The measurement of convergence we describe is significantly more granular than other approaches, and can be used to assess the relative influence of conversational and social dynamics on human communicative behavior (at least within the domain of linguistic convergence). Its measurements are tailored towards this kind of analysis. By focusing on quantitatively describing convergence, we also circumvent the need for immense quantities of training data to "fit a model" with. We demonstrate this point throughout this paper. In layperson's terms: We do not engineer a general purpose approximator. We propose a theory specific measurement that researchers can use at their discretion. Thus, our objective is different from most work in contemporary ML, but is strongly rooted in data-driven approaches to theoretical cognitive science (Johns, Jamieson & Jones, 2023; Jones, 2016).

Linguistic convergence is present in nearly all examples of human communication. It may serve generally as an organizing principle for how human beings use language, and perhaps to signal many important aspects of our social world (Pickering & Garrod, 2004; Branigan et al., 2000; Reitter & Moore, 2014; Giles et al., 2007; Angus et al., 2012; Brennan & Clark, 1996). Even if interlocutors can't use "vogueing" in a grammatically appropriate way themselves, they can easily recognize at least one of the myriad social identities held by speakers who can. But what is perhaps more fascinating: an interlocutor can *learn* how to use "vogueing" correctly in a waltz of ever decreasing entropy – adapting the way they speak as they themselves become more and more integrated within a community that they grow into.

Open Practices Statement

The code and data required to reproduce the analyses in this paper can be found on the Open Science Foundation at the following repository:

https://osf.io/at85c/?view_only=e3879dbf4119465085a8f8b2e7665a25

These analyses were not preregistered.

References

Adams, A., Miles, J., Dunbar, N. E., & Giles, H. (2018). Communication accommodation in text messages: Exploring liking, power, and sex as predictors of textisms. *The Journal of Social Psychology, 158*(4), 474–490. https://doi.org/10.1080/00224545.2017.1421895

Alatawi, F., Sheth, P., & Liu, H. (2023). Quantifying the echo chamber effect: An embedding distance based approach. Retrieved September 10, 2023, from arXiv:2307.04668

Albert, S., de Ruiter, L. E., & De Ruiter, J. (2015). Cabnc: The Jeffersonian transcription of the spoken British national corpus. CABNC: The Jeffersonian transcription of the Spoken British National Corpus.

Angus, D., Smith, A., & Wiles, J. (2012). Conceptual recurrence plots: revealing patterns in human discourse . *IEEE Transactions on Visualization and Computer Graphics, 18*(6), 988–997. https://doi.org/10.1109/TVCG.2011.100 [Conference Name: IEEE Transactions on Visualization and Computer Graphics]

Angus, D., Watson, B., Smith, A., Gallois, C., & Wiles, J. (2012). Visualising conversation structure across time: Insights into effective doctor-patient consultations. *PloS one, 7*(6), e38014.

Ba, L., & Zhao, W. G. W. (2021). Symbolic convergence or divergence? Making sense of (the Rhetorical) senses of a university-wide organizational change. *Frontiers in Psychology, 12*, 690757. https://doi.org/10.3389/fpsyg.2021.690757

Bäck, E. A., Bäck, H., Sendén, M. G., & Sikström, S. (2018). From I to we: Group formation and linguistic adaption in an online xenophobic forum [Number: 1]. *Journal of Social and Political Psychology, 6*(1), 76–91. https://doi.org/10.5964/jspp.v6i1.741

Bailey, B. (2000). Language and negotiation of ethnic/racial identity among Dominican Americans. *Language in Society, 29*(4), 555–582. https://doi.org/10.1017/S0047404500004036, [Publisher: Cambridge University Press]

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics, 18*(2), 135–160. https://doi.org/10.1111/josl.12080, [_eprint: https://www.onlinelibrary.wiley.com/doi/pdf/10.1111/josl.12080]

Barker, J. O., & Rohde, J. A. (2019). Topic clustering of E-Cigarette submissions among reddit communities: a network perspective. *Health Education & Behavior, 46*(2_suppl), 59S–68S. https://doi.org/10.1177/1090198119863770 [Publisher: SAGE Publications Inc]

Bormann, E. G. (1982). The symbolic convergence theory of communication: Applications and implications for teachers and consultants. *Journal of Applied Communication Research, 10*(1), 50–61. https://doi.org/10.1080/00909888209365212

Bradac, J. J., Mulac, A., & House, A. (1988). Lexical diversity and magnitude of convergent versus divergent style shifting-: Perceptual and evaluative consequences. *Language & Communication, 8*(3), 213–228. https://doi.org/10.1016/0271-5309(88)90019-5

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition, 75*(2), B13–B25. https://doi.org/10.1016/S0010-0277(99)00081-5

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 22* (6), 1482–1493. https://doi.org/10.1037//0278-7393.22.6.1482

Brennan, S., Galati, A., & Kuhlen, A. (2010). Two minds, one dialog: Coordinating speaking and understanding. *The Psychology of Learning and Motivation: Advances in Research and Theory* (pp. 301–344). Elsevier. https://doi.org/10.1016/C2009-0-62209-1

Cota, W., Ferreira, S. C., Pastor-Satorras, R., & Starnini, M. (2019). Quantifying echo chamber effects in information spreading over political communication networks [Number: 1 Publisher: SpringerOpen]. *EPJ Data Science, 8*(1), 1–13. https://doi.org/10.1140/epjds/s13688-019-0213-9

Dale, R., Duran, N. D., & Coco, M. (2018). Dynamic natural language processing with recurrence quantification analysis. Retrieved January 2, 2022, from arXiv:1803.07136

Dale, R., & Spivey, M. J. (2005). Categorical recurrence analysis of child language. *Proceedings of the Annual Meeting of the Cognitive Science Society, 27*(27), 530–535.

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011). Mark my words!: Linguistic style accommodation in social media. *Proceedings of the 20th international conference on World wide web - WWW* '11, 745. https://doi.org/10.1145/1963405.1963509

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. Retrieved March 31, 2021, from arXiv:1810.04805

DiBranco, A. (2020). The Men' rights movement and violence - institute for research on male supremacism. Retrieved June 15, 2022, from https://www.malesupremacism.org/2020/07/21/the-mens-rightsmovement-and-violence/

Dougherty, D. S., Kramer, M. W., Klatzke, S. R., & Rogers, T. K. K. (2009). Language convergence and meaning divergence: A meaning centered communication theory. *Communication Monographs, 76*(1), 20–46. https://doi.org/10.1080/03637750802378799

Dougherty, D. S., Mobley, S. K., & Smith, S. E. (2010). Language convergence and meaning divergence: A theory of intercultural communication. *Journal of International and Intercultural Communication, 3*(2), 164–186. https://doi.org/10.1080/17513051003611628

Doyle, G., Goldberg, A., Srivastava, S., & Frank, M. (2017). Alignment at work: using language to distinguish the internalization and self-regulation components of cultural fit in organizations. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Vol. 1:Long Papers), 603–612. https://doi.org/10.18653/v1/P17-1056

Dragojevic, M., & Giles, H. (2014). Language and interpersonal communication: Their intergroup dynamics. In C. R. Berger (Ed.), *Interpersonal Communication* (pp. 29-51). DE GRUYTER. https://doi.org/10.1515/9783110276794.29

Dumais, S., Furnas, G., Landauer, T., Deerwester, S., & Harshman, R. (1996). Using latent semantic analysis to improve access to textual information. *Proceedings, CHI, 88*. https://doi.org/10.1145/57167.57214

Duran, N. D., Paxton, A., & Fusaroli, R. (2019). Align: Analyzing linguistic interactions with generalizable techniques-a python library. *Psychological Methods, 24*(4), 419.

Ethayarajh, K. (2019). How contextual are contextualizedword representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. Retrieved March 19, 2023, from arXiv:1909.00512

Gallois, C., Ogay, T., & Giles, H. (2005). Communication accommodation theory: A look back and a look ahead. *Theorizing About Intercultural Communication*.

Gallois, C., Gasiorek, J., Giles, H., & Soliz, J. (2016). Communication accommodation theory: integrations and new framework developments. In H. Giles (Ed.), *Communication Accommodation Theory* (1st ed., pp. 192-210). Cambridge University Press. https://doi.org/10.1017/CBO9781316226537.010

Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2017). Quantifying controversy in social media. https://doi.org/10.48550/arXiv.1507.05224, arXiv:1507.05224

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination*. *Cognition, 27*,. https://doi.org/10.1016/0010-0277(87)90018-7

Giles, H., Willemyns, M., Gallois, C., & Anderson, M. C. (2007). Accommodating a new frontier: The context of law enforcement. *Social Communication* (p. 35). Psychology.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ...& Nastase, S. A. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience, 25*(3), 369–380. https://doi.org/10.1038/s41593-022-01026-4

Hamilton, W. L., Clark, K., Leskovec, J., & Jurafsky, D. (2016). Inducing domain-specific sentiment lexicons from unlabeled corpora. *Conference on Empirical Methods in Natural Language Processing, 2016*, 595–605. https://doi.org/10.18653/v1/D16-1057

Hassan, S. A., & Shah, M. J. (2019). The anatomy of undue influence used by terrorist cults and traffickers to induce helplessness and trauma, so creating false identities. *Ethics, Medicine and Public Health, 8*, 97–107. https://doi.org/10.1016/j.jemep.2019.03.002

Hassan, S. (2017). Brainwashing young people into violent extremist cults. Freedom from Fear, 2017 (13), 18–22. https://doi.org/10.18356/d37f4d01-en [Publisher: United Nations]

Hilte, L. (2023). How is linguistic accommodation perceived in instant messaging? A survey on teenagers' evaluations and perceptions. *Journal of Language and Social Psychology, 0261927X2311671*. https://doi.org/10.1177/0261927X231167108

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Casas, D.d.L. (2022). Training Compute-Optimal Large Language Models. https://doi.org/10.48550/arXiv.2203.15556, arXiv:2203.15556

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior, 1*, 119–136.

Johns, B. T. (2021). Distributional social semantics: Inferring word meanings from communication patterns. *Cognitive Psychology, 131*, 101441.

Johns, B. T., Jamieson, R. K., & Jones, M. N. (2023). Scalable cognitive modelling: Putting Simon's (1969) ant back on the beach. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*.

Jones, M. N. (2016). *Big data in cognitive science*. Psychology Press.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*(1), 1.

Jones, S., Cotterill, R., Dewdney, N., Muir, K., & Joinson, A. (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* (pp. 455–465). https://aclanthology.org/C14-1044

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Gray, S. (2020). Scaling laws for neural language models. arXiv:2001.08361

Keblusek, L., Giles, H., & Maass, A. (2017). Communication and group life: How language and symbols shape intergroup relations. *Group Processes & Intergroup Relations, 20*(5), 632–643. https://doi.org/10.1177/1368430217708864 [Publisher: SAGE Publications Ltd]

Khan, A. (2019). Text mining to understand gender issues: Stories from the red pill, men's rights, and feminism movements [Accepted: 2019-08-28T15:36:59Z Publisher: University of Waterloo]. Retrieved May 29, 2022, from https://uwspace.uwaterloo.ca/handle/10012/14973

Krendel, A., McGlashan, M., & Koller, V. (2021). The representation of gendered social actors across five manosphere communities on Reddit [Number: 2]. Corpora, 17 (2). Retrieved May 25, 2022, from https://eprints.lancs.ac.uk/id/eprint/155332/

LaFree, G., Atwell-Seate, A., Pisoiu, D., Stevenson, J., Tinsley, H., Manager, G., & Picarelli, J. (2016). Final report: Empirical assessment of domestic radicalization (EADR) (tech. rep. No. 250481). National Institute of Justice, Office of Justice Programs, U.S. Department of Justice.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2013). *Handbook of latent semantic analysis*. Psychology Press.

LaViolette, J., & Hogan, B. (2019). Using platform signals for distinguishing discourses: The case of Men's rights and Men's liberation on reddit. *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media* (ICWSM), 323–334.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*(2), 203–208.

MacIntyre, P. D. (2019). Anxiety/uncertainty management and communication accommodation in women's brief dyadic conversations with a stranger: an idiodynamic approach. *SAGE Open, 9*(3), 215824401986148. https://doi.org/10.1177/2158244019861482

Male Supremacy. (2021). Retrieved July 30, 2022, from https://www.splcenter.org/fighting-hate/extremistfiles/ideology/male-supremacy

Mange, J., Lepastourel, N., & Georget, P. (2009). Is your language a social clue? Lexical markers and social identity. *Journal of Language and Social Psychology, 28*(4), 364–380. https://doi.org/10.1177/0261927X09341956 [Publisher: SAGE Publications Inc]

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space.

Nganga, S. W. (2020). Winning through (Dis)Alignment: The language of the Kenyan female politicians. In C. Kioko, R. Kagumire, & M. Matandela (Eds.), *Challenging patriarchy: The role of patriarchy in the roll-back of democracy* (pp. 150–162). Heinrich-Böll-Stiftung.

Nishida, S., Blanc, A., Maeda, N., Kado, M., & Nishimoto, S. (2021). Behavioral correlates of cortical semantic representations modeled by word vectors. *PLOS Computational Biology, 17*(6), e1009138. https://doi.org/10.1371/journal.pcbi.1009138 [Publisher: Public Library of Science]

Park, A., & Conway, M. (2018). Harnessing reddit to understand the written-communication challenges experienced by individuals with mental health disorders: Analysis of texts from mental health communities. *Journal of Medical Internet Research, 20*(4), e8219 . https://doi.org/10.2196/jmir.8219 [Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada]

Paxton, A., Dale, R., & Richardson, D. C. (2016). Social coordination of verbal and nonverbal behaviours. In P. Passos, K. Davids, & J.Y. Chow (Eds.), *Interpersonal coordination and performance in social systems* (p. 259). Routledge.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Pérez-Sabater, C., & Maguelouk, M. G. (2019). Managing identity in football communities on Facebook: Language preference and language mixing strategies. *Lingua, 225*, 32–49. https://doi.org/10.1016/j.lingua.2019.04.003

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(2), 169–190. https://doi.org/10.1017/S0140525X04000056 [Publisher: Cambridge University Press]

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. Retrieved May 1, 2020, from arXiv:1908.10084

Reitter, D., & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language, 76*, 29–46.

Ribeiro, M. H., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., Long, S.,... & Greenberg, S. (2021). The evolution of the manosphere across the web (tech. rep. arXiv:2001.07600). arXiv. Retrieved June 14, 2022, from arXiv:2001.07600

Roozenbeek, J., & Salvador Palau, A. (2017). I read it on reddit: Exploring the role of online communities in the 2016 US elections news cycle. In G. L. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.), *Social Informatics* (pp. 192-220). Springer International Publishing. https://doi.org/10.1007/978-3-319-67256-4_16

Rosen, Z. P. (2022). A BERT's eye view: A big data framework for assessing language convergence and accommodation. *Journal of Language and Social Psychology*. https://doi.org/10.1177/0261927X221095865

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 55.

Shatz, I. (2017). Fast, free, and targeted: Reddit as a source for recruiting participants online. *Social Science Computer Review, 35*(4), 537–549. https://doi.org/10.1177/0894439316650163 [Publisher: SAGE Publications Inc]

Shin, H., & Doyle, G. (2018). Alignment, acceptance, and rejection of group identities in online political discourse. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop,* 1–8. https://doi.org/10.18653/v1/N18-4001

Smaldino, P. E. (2019). Social identity and cooperation in cultural evolution. *Behavioural Processes, 161*, 108–116. https://doi.org/10.1016/j.beproc.2017.11.015

Smaldino, P. E., Flamson, T. J., & McElreath, R. (2018). The evolution of covert signaling. *Scientific Reports, 8*(1), 4905. https://doi.org/10.1038/s41598-018-22926-1

Soler, A. G., & Apidianaki, M. (2021). Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. Retrieved September 13, 2022, from arXiv:2104.14694

Soliz, J., Giles, H., & Gasiorek, J. (2021). Communication accommodation theory: Converging toward an understanding of communication adaptation in interpersonal relationships. In D.O. Braithwaite, & P. Schrodt (Eds.), *Engaging Theories in Interpersonal Communication: Multiple Perspectives* (3rd ed., pp. 130-142). Routledge. https://doi.org/10.4324/9781003195511

Stine, Z. K., & Agarwal, N. (2020). Comparative discourse analysis using topic models: Contrasting perspectives on china from reddit. *International Conference on Social Media and Society, 73–84*. https://doi.org/10.1145/3400806.3400816

Tajfel, H. (1979). Individuals and groups in social psychology*. *British Journal of Social and Clinical Psychology, 18*(2), 183–190. https://doi.org/10.1111/j.2044-8260.1979.tb00324.x [_eprint: https://www.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8260.1979.tb00324.x]

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology, 1*(2), 149–178. https://doi.org/10.1002/ejsp.2420010202 [_eprint: https://www.onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2420010202]

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Terren, L., & Borge-Bravo, R. (2021). Echo chambers on social media: A systematic review of the literature. *Review of Communication Research, 9*, 99–118. Retrieved September 10, 2023, from https://rcommunicationr.org/index.php/rcr/article/view/94

Tolston, M. T., Riley, M. A., Mancuso, V., Finomore, V., & Funke, G. J. (2019). Beyond frequency counts: Novel conceptual recurrence analysis metrics to index semantic coordination in team communications. *Behavior Research Methods, 51*, 342–360.

Utsumi, A. (2020). Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science, 44*(6), e12844. https://doi.org/10.1111/cogs.12844 [_eprint: https://www.onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12844]

Velásquez, N., Manrique, P., Sear, R., Leahy, R., Restrepo, N. J., & Illari, L.,...& Lupu, Y. (2021). Hidden order across online extremist movements can be disrupted by nudging collective chemistry. *Scientific Reports, 11*(1), 9965. https://doi.org/10.1038/s41598-021-89349-3

Villa, G., Pasi, G., & Viviani, M. (2021). Echo chamber detection and analysis. *Social Network Analysis and Mining, 11*(1), 78. https://doi.org/10.1007/s13278-021-00779-3

Villalobos, P., Sevilla, J., Heim, L., Besiroglu, T., Hobbhahn, M., & Ho, A. (2022). Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning. https://doi.org/10.48550/arXiv.2211.04325, arXiv:2211.04325

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. Retrieved September 13, 2022, from arXiv:1909.10430

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A.,...& Cistac, P. (2020). HuggingFace's transformers: State-of-the-art natural language processing. Retrieved September 30, 2021, from arXiv:1910.03771

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language, 72*, 37–48. https://doi.org/10.1016/j.jml.2013.12.003

Xu, Y., & Reitter, D. (2015). An evaluation and comparison of linguistic alignment measures. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics* (pp. 58–67). https://doi.org/10.3115/v1/W15-1107

Yenicelik, D., Schmidt, F., & Kilcher, Y. (2020). How does BERT capture semantics? A closer look at polysemous words. *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP,* 156–162. https://doi.org/10.18653/v1/2020.blackboxnlp-1.15

Zhang, D., Lin, H., Liu, X., Zhang, H., & Zhang, S. (2019). Combining the attention network and semantic representation for Chinese verb metaphor identification. *IEEE Access, 7*, 137103–137110. https://doi.org/10.1109/ACCESS.2019.2932136

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.