The (in)efficiency of within-language variation in online communities

Jai Aggarwal **Julia Watson** **Prabuddha Senapati** **Suzanne Stevenson**

Department of Computer Science
University of Toronto
{jai, jwatson, prabuddha, suzanne}@cs.toronto.edu

Abstract

We conduct a large-scale study of online community variation in language. We show that factors of efficient communication, which have been shown to drive crosslinguistic variation in lexical semantic systems, also play a role in *within-language variation* across 1926 English-language Reddit communities. We study variation in stancetaking behaviour, a domain where efficient communication may be influenced by social motivations for language use. We find that communities indeed have efficient stancetaking systems, particularly with respect to their own communicative needs. However, contrasting with crosslinguistic work, we find that communities are often *not optimized* for their needs. Moreover, we find that community-level social factors correlate with how optimized they are. These results highlight the importance of accounting for social pressures for language use when studying how efficient communication drives variation.

Keywords: efficient communication; communicative need; within-language community variation; stance

Introduction

Variation in lexical semantic systems – how languages carve up semantic space into words – has been argued to reflect a drive for efficient communication. Evidence from a range of domains has shown that languages vary in how they optimize a trade-off between two communicative pressures: informativity – how precise or fine-grained words are in their meanings; and simplicity – such as keeping the number of lexical terms for the domain low (Kemp & Regier, 2012; Zaslavsky et al., 2018, 2021). Moreover, the general drive for efficiency interacts with language-specific communicative need – the extent to which a language needs to refer to a given meaning or domain (Gibson et al., 2017; Bradford et al., 2022; Gao & Regier, 2022; Anand & Regier, 2023). For example, languages spoken in colder climates (where people likely talk more about cold weather) more often use separate words to label snow and ice (Regier et al., 2016), yielding a more informative – but less simple – system.

Less work has considered how pressures for efficient communication shape linguistic variation across communities *within* a language (although see Sun & Xu, 2022; Watson et al., 2023). Sociolinguists study such variation across **communities of practice** – social groups defined by their shared norms, values, and interests (Cheshire, 1982; Eckert, 2000). These groups often develop variations in language that help to both signal and reinforce community membership (Eckert & McConnell-Ginet, 1992; Holmes & Meyerhoff, 1999).

One way communities linguistically distinguish themselves is through different approaches to **stancetaking** (Bucholtz & Hall, 2005) – dialogic behaviors that speakers use to position themselves, such as affectively or evaluatively, toward a topic or other speakers. For example, while someone may talk about an *amazingly good book* in everyday conversation, they know not to use *amazingly good* to describe a cited work when writing a scientific paper. A key component of this knowledge is a lexical semantic system of stance – henceforward, **stance system** – in which **stance markers** (words that express a stance, like *amazingly*) are mapped to **stance meanings** (sets of stance-related properties, such as formality and valence, that capture the speaker's position).

Communities vary greatly in their stance systems. To illustrate this, consider the variation in word choices for expressing stance in similar, low formality situations, on Reddit communities r/AskHistorians (a community for discussing history) and r/aww (a community for sharing cute animal pictures), shown in Figure 1a, (i) and (iii). Here, use of *particularly* reflects the academic culture of r/AskHistorians, while *especially* reflects the more casual, personal sense of community on r/aww. Indeed, Figure 1a shows that r/AskHistorians uses the relatively formal word *particularly* very broadly, in both informal and formal settings (i and ii), while r/aww uses it very precisely in high formality settings (iv).

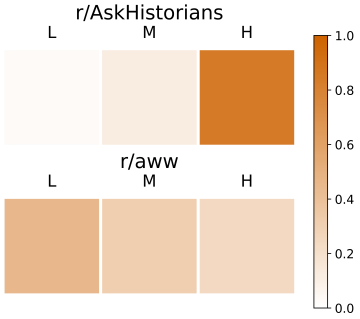
We hypothesize that, as in crosslinguistic analyses, communicative need shapes how communities trade off precision vs. simplicity in stance systems. For example, Figure 1b shows that r/AskHistorians has greater need to talk about high formality situations, compared to r/aww. In line with this, we observe that r/AskHistorians uses a substantially larger number of different stance markers to express more formal stances, including *adamantly*, *egregiously*, and *veritably*, which are never used in such situations on r/aww.

However, this difference in communicative need between the two communities does not explain why r/AskHistorians uses *particularly* so broadly. To understand this, we need to consider speakers' social motivations: They may use a more formal word like *particularly* even in informal settings in order to fit in with the academic tenor of the community. Here, we argue that, in addition to pressures related to community-specific communicative needs, stance systems may also be shaped by such social goals related to signaling and reinforcing community membership.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

(i)	**r/AskHistorians** Low formality	Well... for much of Europe, **particularly** in Western Europe, they did!
(ii)	**r/AskHistorians** High formality	The latter definition is a **particularly** important one here, as I'll get into below.
(iii)	**r/aww** Low formality	Source: had a cat like that, **especially** in the mornings.
(iv)	**r/aww** High formality	I'm **particularly** interested in exactly how dangerous our own bacteria is to birds.

(a) Examples from subreddits r/AskHistorians and r/aww, in Low and High Formality situations.



(b) Communicative need (relative frequency) by level of formality (Low, Medium, High)

Figure 1: Online communities (subreddits) vary in (a) use of stance markers and (b) communicative needs (data from Reddit).

We study such issues at scale, using real-world data from 1926 English-language online communities on Reddit. Reddit communities are ideal for our research focus, as they have been shown to vary both in their communicative needs (Watson et al., 2023), and in their language use (Del Tredici & Fernández, 2017; Lucy & Bamman, 2021; Aggarwal et al., 2023). We address three research questions in this setting:

RQ1: Do communities efficiently trade off between simplicity and informativity in the domain of stancetaking?

RQ2: Are stance systems shaped by community-specific communicative needs?

RQ3: How do community-specific social factors influence the efficiency of communities' stance systems?

In RQ1 we find that, when assessed using Reddit-general communicative needs, communities indeed optimize this trade-off efficiently, showing how general language pressures shape their stance systems. RQ2 builds on this to show that communities are more efficient relative to their own needs than those of Reddit in general. Surprisingly, however, in contrast to crosslinguistic work, we find that they are not as efficient as they could be, suggesting that other community-related social factors may be at play. This motivates RQ3, where we find correlations of efficiency with various social structural properties of communities, such as size and community interaction. Drawing on research on social network structure and communication (Trudgill, 1997; Raviv et al., 2020), we argue that communities' lexical semantic systems are shaped by social goals, which up to now have not been integrated into frameworks of efficient communication.

Communicating Stance

In this section, we first describe the lexical semantic system of stance that we study. We then explain how we model efficient communication given such a system, framing the trade-off between informativity and simplicity as an equivalent trade-off between their inverses: communicative cost and complexity.

Stancetaking as a Lexical Semantic System

Stancetaking is a broad communicative behavior that engages linguistic devices from the phonological to the pragmatic. Here we study the lexical semantic system mapping stance markers (words) to the properties of stance being communicated (meanings). We focus on the set of English **intensifiers**, such as *particularly* and *especially*, because they play a key role in emphasizing and adding nuance to a speaker's stance. Moreover, they do so in a variety of linguistic and social contexts (Bolinger, 1972; Ito & Tagliamonte, 2003; Tagliamonte & Roberts, 2005), thereby expressing a range of stance properties, with potential for community variation; cf. Figure 1.

Defining such stance systems depends on having an appropriate semantic space for representing the meanings of intensifiers. We adopt the context-based approach of Aggarwal et al. (2023), which views the meaning of each intensifier usage as a vector of semantic features gleaned from the usage sentence. These semantic vectors capture a key set of properties relevant to the expression of stance, including affect (broken down into valence, arousal, and dominance), politeness, and formality (Jaffe, 2009; Pavalanathan et al., 2017; Kiesling et al., 2018). To apply existing information theoretic methods that require meanings to be discrete (e.g., Regier et al., 2015; Y. Xu et al., 2016), we group these continuous vectors into discrete meanings that represent regions of the stance semantic space (explained in Data and Methods below). Each intensifier can then be associated with a set of such meanings (which correspond to its usage contexts), and meanings can be (and typically are) associated with more than one intensifier, yielding a lexical semantic system of stance.

Assessing Efficiency of Stancetaking

The next step is to determine how efficiently a community deploys its lexical semantic system of stance according to its communicative needs. As in much previous work (e.g., Regier et al., 2015; Y. Xu et al., 2016), we model the communication process as speaker S in community c using word w to convey intended meaning m to listener L, who must recover the intended meaning; see Figure 2.

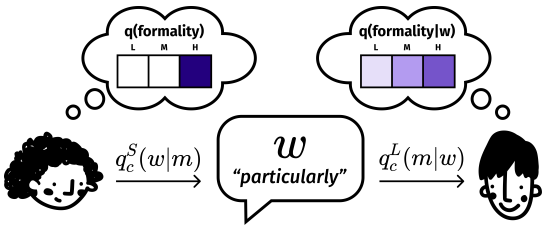


Figure 2: Overview of communication model. (Note formality is only one dimension of our meaning space.)

The speaker's choice to convey meaning m with word w is modeled probabilistically as $q_c^S(w|m)$, which we derive using population-level data from each community c. In this way, $q_c^S(w|m)$ across all stance words and meanings represents one community's lexical semantic system for stance. The listener must mentally reconstruct the speaker's intended meaning from word w. Since w may refer to multiple meanings, the listener considers a distribution $q_c^L(m|w)$ over all meanings that w could refer to; this is defined formally below.

Communicative Cost The communicative cost of a word relates to how accurately it allows the listener to reconstruct the speaker's intended meaning. Precise words (that convey similar meanings, e.g., only high formality stances) allow for more accurate reconstruction, and incur a lower communicative cost than broader words (that convey dissimilar meanings, e.g., both high and low formality stances). Thus, assuming that w labels a set of meanings C_w, the accuracy of the listener's distribution $q_c^L(m|w)$ is shaped by how dissimilar the meanings in C_w are (e.g., as in Kemp & Regier, 2012; Y. Xu et al., 2020).

Stance markers are often semantically broad, and vary in the probability that they convey any one meaning. We thus weight the similarity function used in previous work by $p_c^L(m_i|w)$, the probability that w labels meaning m_i. We assume a Bayesian listener that has access to the speaker distribution $q_c^S(w|m)$, as both are members of the same community:

$$q_c^L(m|w) \propto \sum_{m_i \in C_w} sim(m,m_i) p_c^L(m_i|w) \quad (1)$$

$$p_c^L(m_i|w) \propto q_c^S(w|m_i) p(m_i) \quad (2)$$

Here $p(m_i)$ is the communicative need for m_i and *sim* is the inverse squared Euclidean distance between the representations of two meanings.[1]

If a listener's reconstruction is less accurate, this leads to information loss in communication. A system's communicative cost is the expected amount of information loss across all words and meanings, typically modeled using KL Divergence (e.g., Y. Xu et al., 2016; Zaslavsky et al., 2018). Because we assume the speaker is certain of the meaning they intend to convey (i.e., $q_c^S(m) = 1$), we have:

$$\text{Cost} = E_q[KL[q_c^S(m)||q_c^L(m|w)]] \quad (3)$$

$$= \sum_{m,w} p(m) q_c^S(w|m) \log \frac{1}{q_c^L(m|w)} \quad (4)$$

Complexity Crosslinguistic work often formalizes complexity of a lexical semantic system as the cognitive effort in representing a semantic domain with some number of terms or rules (e.g., Kemp & Regier, 2012; Y. Xu et al., 2020). However, communities *within* a language have access to the same set of words; in this setting, we propose that cognitive effort arises in representing *multiple* lexical systems: both one's general knowledge of stance marker usage, and how they are used within a community. For example, the use of *amazingly good* would be inappropriate in a scientific context, as mentioned earlier. We propose then that complexity increases with distinctiveness: the more distinctive the uses of stance markers in a community (compared to general usage), the higher the complexity of its lexical system.

We capture word usage specific to community c using the joint distribution $q_c(w,m)$. This captures a speaker's knowledge of words in the community, $q_c^S(w|m)$, and accounts for the communicative need for meanings $p(m)$ (though this may be the community's own needs $q_c(m)$, we vary this across research questions). We estimate a speaker's general understanding of how words are used, $q^*(w,m)$, with general language data from Reddit (described in Data and Methods). We use KL Divergence between the community distribution and the general distribution to capture complexity as the additional information (over general usage) a speaker must learn in order to express stance in a particular community:

$$\text{Complexity} = KL[q_c(w,m)||q^*(w,m)] \quad (5)$$

To summarize, a community stance system is more complex to the extent that it differs from more general usage on Reddit.

Efficiency Trade-off We explore whether communities trade off between the communicative cost and complexity of their stance systems: do they develop a system more distinct from the general understanding of intensifier usage if it enables community members to infer stance meanings more accurately? Moreover, the efficiency of this trade-off may be shaped by the communicative need distribution $p(m)$; in our research questions, we consider needs both from general language and from the specific community. As in previous work (e.g., Kemp & Regier, 2012; Regier et al., 2015), we assess the efficiency of attested lexical semantic systems by comparing them to hypothetical systems – a large set of realistic, logically possible partitions of the stance meaning space (described in Data and Methods).

Data and Methods

Representing Stance Our stance markers are the set of 252 single-word English-language intensifiers released by Luo et al. (2019). As noted, we use context-based semantic representations of intensifier usages (Aggarwal et al., 2023), which

[1]To have a continuous similarity function, each discrete stance meaning is represented here as the average of all the continuous semantic vector representations mapped to it.

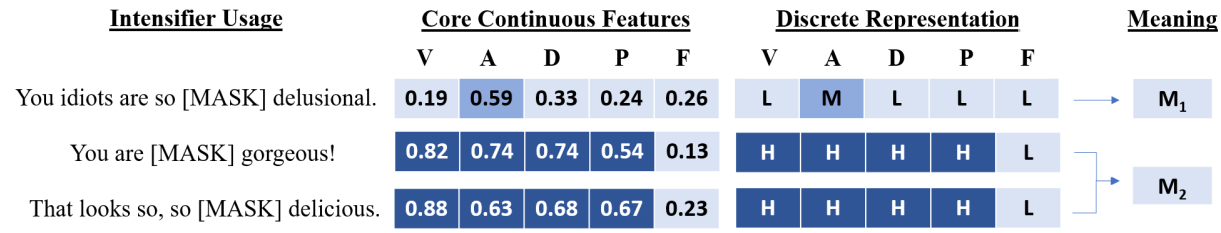


Figure 3: Examples of stance meanings assigned to intensifier usages (the intensifier itself is masked out). Continuous representations show raw values for the 5 core features: valence (V), arousal (A), dominance (D), politeness (P), and formality (F) (extremeness features not shown). Discrete representations assign each score to the low (L), medium (M), or high (H) tercile. Matching discrete representations are assigned to the same meaning in the final columns, as is the case for the two bottom rows.

are 10-dimensional vectors. Five core features encode the valence, arousal, dominance, politeness, and formality of the sentential context of use of intensifiers. Another 5 features indicate the level of extremeness (distance from the mean) of each of these features, because extremeness is relevant for intensifiers given their role in expressing emphasis. The 5 core features are calculated by applying regression models to infer values for the SBERT-based embeddings (Reimers & Gurevych, 2019) of the sentence contexts (with the intensifier masked), extrapolating from ground-truth data resources for these properties; see Aggarwal et al. (2023) for details.

We create discrete meanings by splitting each of the 5 core features into terciles of low, medium, and high values across the dataset, resulting in $3^5 = 243$ distinct stance meanings that capture combinations of these features. Each sentence thus has a 10D vector representing the continuous meaning of an intensifier's usage in that sentence, which is binned to one of the 243 stance meanings; see Figure 3. We study variation in the distribution of each intensifier's usages across these 243 meanings. We also use the distribution of the meanings themselves, within each community, to estimate communicative need (as described in the next section).

Extracting Reddit Data Reddit is an ideal source of large-scale data for studying community variation, as it is divided into communities of practice called *subreddits* (providing community-specific data), whose posts are unconstrained in length (thus fairly representative of how people naturally speak). We extract data from 2019 using the Pushshift Data Dumps (Baumgartner et al., 2020), retaining sentences that have exactly one intensifier and are at least 6 tokens long (for reliable assessment of the semantic features). To ensure plentiful data for each subreddit, we focus on those with at least 10K sentences. For each such subreddit (except r/AskReddit, as described next), we sample exactly 10K sentences, so that all communities have a comparable amount of data. This yields a dataset of 1926 distinct subreddits having a total of 19.26M intensifier usages. We assign a stance meaning to each intensifier usage as described above, and use this to derive a joint distribution of stance behaviour $q_c(w,m)$ for each subreddit c. We then factor $q_c(w,m)$ into a subreddit's **stance system** $q_c(w|m)$ and **need distribution** $q_c(m)$.

Recall that our notion of complexity is how distinct a community's stance system is from the general usage of the lexical system of intensifiers. We assume that data from r/AskReddit is apt for capturing knowledge of the latter, given the massive size, diverse user base, and broad topical scope of this subreddit.[2] We use the 6.8M sentences from r/AskReddit that satisfy the criteria above to compute the general distribution, $q^*(w,m)$, used in Equation (5).

Generating Hypotheticals To show that attested systems in our communities are efficient, we require a large set of hypothetical systems that represent diverse solutions to the cost–complexity tradeoff. Moreover, these solutions should be realistic, such that words within a system vary in their frequency of usage and their semantic breadth, as occurs in our attested systems. To meet these criteria, we generate hypotheticals using a Gaussian mixture model.

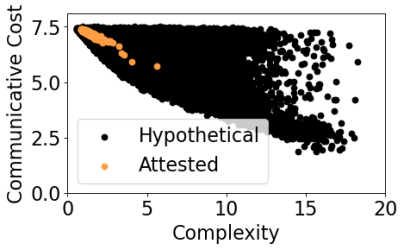
For each hypothetical stance system, we model each intensifier as a Gaussian distribution over meanings. The mean of each Gaussian is the continuous representation x_w of some stance meaning m_w (sampled randomly with replacement from our set of stance meanings). To model variation in semantic breadth, we draw each Gaussian's variance from $N(0,v)$ (where v is a system-level parameter that we vary). To model variation in word frequency, we draw mixture weights $q(w)$ from a Zipfian distribution over the range $[1,1000]$. After normalizing these weights, we use them to compute the likelihood of a speaker producing each word for each meaning, drawing on the approach of Carlsson et al. (2023):

$$q(w|m) \propto sim(m,m_w)q(w) \tag{6}$$

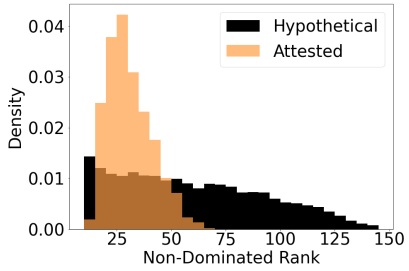
where *sim* is defined as it was for Equation (1). We generate 10K hypothetical systems using this approach.

Measuring Efficiency Following A. Xu et al. (2022), we assess efficiency – how well a system trades off complexity and cost – using non-dominated (ND) rank (Jensen, 2003). Intuitively, a system has lower (better) rank if it is less costly than other systems and no more complex than them (or vice

[2] An alternative approach of averaging usage patterns across all subreddits is less appropriate because general language use is not simply an aggregate of community-specific patterns.

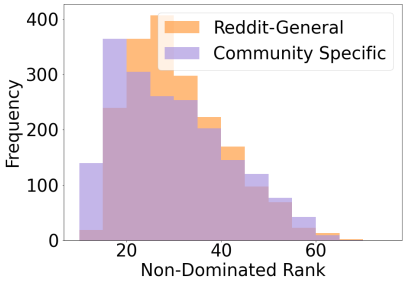


(a) Complexity vs. communicative cost

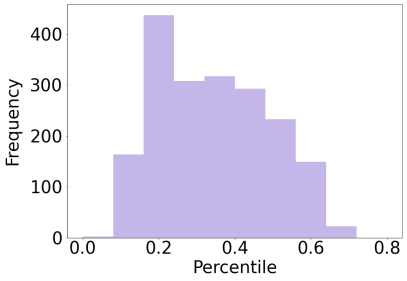


(b) Comparison of ND Ranks

Figure 4: (a) Efficiency of stancetaking systems according to Reddit-general need probabilities and (b) a comparison of ND ranks for attested and hypothetical systems.



(a) Comparison of ND Ranks



(b) Distribution of Efficiency Percentiles

Figure 5: (a) Efficiency of attested systems using community-specific need probabilities compared to Reddit-general needs. (b) The percentile score of each attested community relative to its own needs compared to all other systems.

versa); that is, the lower the ND rank of a system, the fewer systems there are that are more efficient than it. We make all code and data available at `https://github.com/jaikaggarwal/cogsci_2024_efficiency_online`.

RQ1: Assessing Efficiency

Analyses Our first research question investigates whether Reddit communities efficiently trade off between complexity and cost in the domain of stancetaking; that is, do communities develop more complex (distinctive) systems if it allows them to communicate stance more precisely? To answer this, we first consider whether stance systems show a complexity–cost trade-off (using a Spearman correlation), assuming a communicative need distribution that represents general stancetaking needs on Reddit (specifically $q^*(m)$, derived from r/AskReddit). We then test whether attested systems have lower average ND ranks under $q^*(m)$ than (1) hypothetical systems under $q^*(m)$ (using a t-test), and (2) attested systems under an uninformative uniform distribution (using a paired t-test). Recall that a lower ND rank means greater efficiency (i.e., fewer systems are more efficient). We report effect sizes for each t-test using Cohen's d.

Results Figure 4a shows how the 1926 attested community systems on Reddit compare to the 10K hypothetical systems, in terms of complexity and communicative cost. It is notable that attested systems largely occupy the top-left corner of this space, reflecting that Reddit communities favour simplicity (being closer to general language) over informativity (being more precise). Yet communities in this domain do show a clear trade-off between the two communicative pressures, as shown by a strong negative correlation between complexity and cost ($r = -0.85, p < 0.001$; the correlation is the same with outliers removed).

Moreover, Figure 4b shows that they perform this trade-off efficiently, as the average ND rank of attested systems is much lower than that of hypothetical systems ($t = 25.92, p < 0.001, d = 0.81$). Finally, attested systems perform this trade-off more efficiently relative to Reddit-general needs than those of a uniform distribution, as their average ND rank is lower according to the Reddit-general need distribution ($t = 48.61, p < 0.001, d = 1.01$). This suggests that communities have developed stancetaking behaviour sensitive to the general communicative needs of this domain.

RQ2: Community-Specific Needs

Analyses Here we further assess the role of communicative need in shaping the efficiency trade-off, focusing on community-specific needs. First, we test whether attested communities attain lower ND ranks (i.e., are more efficient) under their own need distribution than the Reddit-general needs (assessed using a paired t-test). Second, we consider the extent to which a community is optimized for its needs by seeing how many other attested and hypothetical systems are more efficient than it in fitting its needs. We assess this by examining each attested system's **efficiency percentile** – the proportion of the other 11925 systems that achieve a lower rank than that system, under its need distribution. Lower percentiles indicate greater efficiency relative to other systems.

	Diversity of Contact	Size	Density
Efficiency	-0.36	0.18	-0.22

Table 1: Spearman correlations between social factors and the percentile efficiency score computed in RQ2. All values are significant at $p < 0.001$.

Results As expected, communities attain lower ND ranks (i.e., are more efficient) according to their own need distribution than the Reddit-general need distribution ($t = 5.27, p < 0.001, d = 0.17$); see Figure 5a. This suggests that community stance systems reflect their own communicative needs better than those of Reddit in general, analogous to crosslinguistic work (Gao & Regier, 2022; Anand & Regier, 2023).

However, many systems are not optimized for their needs: Figure 5b shows that most attested systems are only in the top $20-60\%$ most efficient possible stance systems for their own needs. This is somewhat surprising, given findings in crosslinguistic work that languages better fit their own needs than most other possible languages (Gao & Regier, 2022). We suggest that when considering within-language community variation, social motivations may shape language use in ways not typically considered in efficiency analyses (as discussed in Kemp et al., 2018; cf. the pragmatic influences explored in Watson et al., 2023). We turn to such social influences next.

RQ3: Social Factors

Analyses Given that our communities do not optimize for stance as efficiently as possible for its own needs, here we investigate the relationship between a community's efficiency and properties of its social environment. We draw on previous work showing that such properties correlate with variation in the complexity of languages' grammatical systems (e.g., Trudgill, 1997; Raviv et al., 2019). We explore three such factors: diversity of linguistic contact, the size of a community, and the density of interactions within the community.

Diversity of linguistic contact of a community c captures how much c may be influenced by other, very distinctive communities. It is operationalized as the dissimilarity between c and all other communities c', weighted by the potential influence of each c' on c. Dissimilarity is measured using cosine distance over textual embeddings for the communities, as in Lucy & Mendelsohn (2019); the weighting is given by the number of users that participate in both c and c' (possibly 0). Size is the number of unique authors in c, and density captures the proportion of author_i–author_j interactions (in comments and replies) out of all possible interactions among authors on c (as in Lucy & Bamman, 2021).

We compute the Spearman correlation of each of these three properties with the negative of the efficiency percentile scores from RQ2. Because lower percentiles mean higher efficiency, we negate the percentile scores here, such that positive correlations imply a positive relationship between the property and efficiency.

Results Table 1 shows that diversity and density are negatively correlated with efficiency, while size is positively correlated. (Note that size and density are themselves inversely correlated, $r = -0.88$, $p < 0.001$.)

Communities with more diverse linguistic contact may be less efficient because the diverse contact brings in a mixture of stance usages from dissimilar communities, resulting in a stance system that is less tailored to its communicative needs. Community contact is particularly important for understanding within-language variation, as speakers regularly participate in multiple communities (Wenger, 1999).

The correlations of size and density with efficiency align with work exploring how social factors influence linguistic structure (Trudgill, 1997; Lupyan & Dale, 2010; Raviv et al., 2019). Such work argues that smaller, tight-knit social groups have the shared context to support more complex and less transparent grammatical systems, while larger communities with fewer social ties require simpler, more informative systems to communicate clearly. We argue that similar pressures may drive larger, less dense communities on Reddit to have simpler and more informative (i.e., more efficient) stancetaking systems, extending these insights to online communities. Further work is required to tease apart the effects of size and density individually (as argued by Raviv et al. 2019, 2020).

Discussion

In a large-scale study of almost 2000 online communities, we show that similar pressures of communicative efficiency are at play in within-language variation as across languages. However, some of our results contrast with crosslinguistic findings (Gao & Regier, 2022; Anand & Regier, 2023): Although communities' lexical systems are tailored to meet their own communicative needs, they do not often converge on the most efficient way to do so. We find that different patterns of social interaction among users, both within and between communities, help to make sense of these results, highlighting the need to consider broader social factors in assessing communicative efficiency.

These novel insights are made possible by studying stance systems in online communities of practice. First, because stancetaking expresses social positioning and values (e.g., Du Bois, 2007), stance systems vary greatly across communities, making this an ideal domain for studying how social goals shape efficiency. Second, online social media platforms provide very large-scale data across structured communities with substantial user overlap. This enables richer studies of how social network structure shapes efficiency, complementing previous work relating efficiency to environmental factors (Regier et al., 2016).

New opportunities also arise for diachronic research, which is limited for studies of crosslinguistic efficiency by a lack of historical data (cf. Zaslavsky et al., 2022). Using Reddit data can enable investigation of efficiency in the presence of dynamic social structures and shifting communicative needs, in a rapidly changing communicative environment.

Acknowledgments

We acknowledge the support of NSERC of Canada (through grant RGPIN-2017-06506 to SS), as well as the support of the Data Sciences Institute, University of Toronto (through a Catalyst Grant to SS and JW).

References

Aggarwal, J., Diep, B., Watson, J., & Stevenson, S. (2023). Investigating online community engagement through stancetaking. In *Findings of the association for computational linguistics: Emnlp 2023* (pp. 5814–5830).

Anand, G., & Regier, T. (2023). Kinship terminologies reflect culture-specific communicative need: Evidence from hindi and english. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit Dataset. In *Proceedings of the international aaai conference on web and social media* (Vol. 14, pp. 830–839).

Bolinger, D. (1972). Degree words. In *Degree words.* De Gruyter Mouton.

Bradford, L., Thomas, G., & Xu, Y. (2022). Communicative need modulates lexical precision across semantic domains: A domain-level account of efficient communication. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, *7*(4-5), 585–614.

Carlsson, E., Dubhashi, D., & Regier, T. (2023). Iterated learning and communication jointly explain efficient color naming systems. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Cheshire, J. (1982). Variation in an English dialect: A sociolinguistic study. *Cambridge Studies in Linguistics London*, *37*.

Del Tredici, M., & Fernández, R. (2017). Semantic variation in online communities of practice. In *Proceedings of the 12th international conference on computational semantics.*

Du Bois, J. W. (2007). The stance triangle. *Stancetaking in discourse: Subjectivity, evaluation, interaction*, *164*(3), 139–182.

Eckert, P. (2000). *Language variation as social practice: The linguistic construction of identity in belten high*. Wiley-Blackwell.

Eckert, P., & McConnell-Ginet, S. (1992). Think practically and look locally: Language and gender as community-based practice. *Annual review of anthropology*, *21*(1), 461–488.

Gao, S., & Regier, T. (2022). Culture, communicative need, and the efficiency of semantic categories. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Gibson, E., Futrell, R., Jara-Ettinger, J., Mahowald, K., Bergen, L., Ratnasingam, S., ... Conway, B. R. (2017). Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, *114*(40), 10785–10790.

Holmes, J., & Meyerhoff, M. (1999). The community of practice: Theories and methodologies in language and gender research. *Language in society*, *28*(2), 173–183.

Ito, R., & Tagliamonte, S. (2003). Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. *Language in society*, *32*(2), 257–279.

Jaffe, A. (2009). *Stance: Sociolinguistic perspectives*. Oxford University Press.

Jensen, M. T. (2003). Reducing the run-time complexity of multiobjective eas: The nsga-ii and other algorithms. *IEEE Transactions on Evolutionary Computation*, *7*(5), 503–515.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054.

Kemp, C., Xu, Y., & Regier, T. (2018). Semantic typology and efficient communication. *Annual Review of Linguistics*, *4*, 109–128.

Kiesling, S. F., Pavalanathan, U., Fitzpatrick, J., Han, X., & Eisenstein, J. (2018). Interactional stancetaking in online forums. *Computational Linguistics*, *44*(4), 683–718.

Lucy, L., & Bamman, D. (2021). Characterizing English variation across social media communities with BERT. *Transactions of the Association for Computational Linguistics*, *9*, 538–556.

Lucy, L., & Mendelsohn, J. (2019). Using sentiment induction to understand variation in gendered online communities. In *Proceedings of the society for computation in linguistics (scil) 2019* (pp. 156–166).

Luo, Y., Jurafsky, D., & Levin, B. (2019). From insanely jealous to insanely delicious: Computational models for the semantic bleaching of English intensifiers. In *Proceedings of the 1st international workshop on computational approaches to historical language change* (pp. 1–13).

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, *5*(1), e8559.

Pavalanathan, U., Fitzpatrick, J., Kiesling, S. F., & Eisenstein, J. (2017). A multidimensional lexicon for interpersonal stancetaking. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 884–895).

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Larger communities create more systematic languages. *Proceedings of the Royal Society B*, *286*(1907), 20191262.

Raviv, L., Meyer, A., & Lev-Ari, S. (2020). The role of social network structure in the emergence of linguistic structure. *Cognitive Science*, *44*(8), e12876.

Regier, T., Carstensen, A., & Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PloS one*, *11*(4), e0151138.

Regier, T., Kemp, C., & Kay, P. (2015). 11 word meanings across languages support efficient communication. *The handbook of language emergence*, *87*, 237.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv preprint arXiv:1908.10084*.

Sun, Z., & Xu, Y. (2022, December). Tracing semantic variation in slang. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 1299–1313).

Tagliamonte, S., & Roberts, C. (2005). So weird; so cool; so innovative: The use of intensifiers in the television series friends. *American speech*, *80*(3), 280–300.

Trudgill, P. (1997). Typology and sociolinguistics: Linguistic structure, social structure and explanatory comparative dialectology.

Watson, J., Walker, S., Stevenson, S., & Beekhuizen, B. (2023). Communicative need shapes choices to use gendered vs. gender-neutral kinship terms across online communities. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Wenger, E. (1999). *Communities of practice: Learning, meaning, and identity*. Cambridge university press.

Xu, A., Kemp, C., Frermann, L., & Xu, Y. (2022). Word formation supports efficient communication: The case of compounds. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Xu, Y., Liu, E., & Regier, T. (2020). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind*, *4*, 57–70.

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive science*, *40*(8), 2081–2094.

Zaslavsky, N., Garvin, K., Kemp, C., Tishby, N., & Regier, T. (2022). The evolution of color naming reflects pressure for efficiency: Evidence from the recent past. *Journal of Language Evolution*, *7*(2), 184–199.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942.

Zaslavsky, N., Maldonado, M., & Culbertson, J. (2021). Let's talk (efficiently) about us: Person systems achieve near-optimal compression. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).