



UNIVERSITY OF HELSINKI



<https://helda.helsinki.fi>

Helda

---

## Digital debating cultures : Communicative practices on Reddit

Messerli, Thomas C.

Oxford University Press

2025-04

---

Messerli, T C, Dayter, D, Leuckert, S, Liimatta, A, Mahler, H, Bohmann, A, Kozma, G & Tosin, R 2025, 'Digital debating cultures : Communicative practices on Reddit', Digital Scholarship in the Humanities , vol. 40, no. 1, fqaf005, pp. 227-240. <https://doi.org/10.1093/lhc/fqaf005>

---

<http://hdl.handle.net/10138/593289>

10.1093/lhc/fqaf005

---

cc\_by\_nc

publishedVersion

---



*Downloaded from Helda, University of Helsinki institutional repository.*

*This is an electronic reprint of the original article.*

*This reprint may differ from the original in pagination and typographic detail.*

*Please cite the original version.*

# Digital debating cultures: communicative practices on Reddit

Thomas C. Messerli <sup>1</sup>, Daria Dayter<sup>2</sup>, Sven Leuckert <sup>3</sup>, Aatu Liimatta<sup>4</sup>,  
Hanna Mahler<sup>5</sup>, Axel Bohmann<sup>6</sup>, Gustavo Kozma<sup>7</sup>, Rafaela Tosin<sup>8</sup>

<sup>1</sup>English Seminar, Department of Languages and Literatures, University of Basel, Nadelberg 6, Basel 4051, Switzerland

<sup>2</sup>Languages Unit, Tampere University, Kalevantie 4, Tampere 33100, Finland

<sup>3</sup>Institute of English and American Studies, Technische Universität Dresden, Dresden 01062, Germany

<sup>4</sup>Faculty of Arts, University of Helsinki, Helsinki 00014, Finland

<sup>5</sup>English Department, University of Freiburg, Freiburg im Breisgau 79085, Germany

<sup>6</sup>English Department I, University of Cologne, Cologne 50923, Germany

<sup>7</sup>English Department, University of Freiburg, Freiburg im Breisgau 79085, Germany

<sup>8</sup>English Department, University of Freiburg, Freiburg im Breisgau 79085, Germany

\*Corresponding author. English Seminar, Department of Languages and Literatures, University of Basel, Nadelberg 6, Basel 4051, Switzerland. E-mail: thomas.messerli@unibas.ch

## Abstract

Reddit as a platform for online forum interaction hosts various communities (subreddits), which give rise to specific digital communicative practices. Communication on Reddit is on the one hand unified by a set of common affordances, and on the other hand it is characterized by a great deal of variation in terms of topics, interactional patterns, and linguistic features. In this article, we aim to map the communicative space(s) of Reddit by means of vignettes of studies from four research groups, each of them describing a different approach to one or several subreddits. Study 1 examines concessive constructs in *r/changemyview* posts, Study 2 focuses on expert-lay communication on *r/mountaineering*, Study 3 addresses register variation within a large corpus of Reddit comments from many different communities, and Study 4 looks across different subreddits to explore the function of text-length. Drawing on these four vignettes, we show the range of discourses that Reddit has hosted in recent years, while also exploring the methodological tools as well as the insights that qualitative, text-based digital humanities, critical discourse analysis and computational text analysis can provide. The map to the Reddit language research space we create here can help digital humanities researchers to make sense of the variety of different theoretical vantage points and methodologies that are currently being employed to Reddit, and it can serve as a guideline when it comes to choosing research agendas and relevant literature on current forum interaction.

**Keywords:** Reddit; computer-mediated communication; online communities; debating practices; register analysis; text-length; expert-lay communication; persuasion; digital discourse analysis; linguistic pragmatics.

## 1. Introduction

The social media platform Reddit understands itself as a ‘home to thousands of communities’ (<https://www.redditinc.com>). As researchers in the humanities, we concur that the submissions and comments posted to Reddit’s subreddits comprise authentic digital human interaction by groups of people. These groups are in some cases prototypical communities and in other cases the result of chance encounters of users who find themselves oriented towards and engaging with the same virtual space. The collective communicative acts

of Reddit users can be positioned in the tradition of computer-mediated communication (CMC)—as one key site of digitalized communication, shaped partially by the affordances provided by the platform. As data, communication on Reddit is of interest to researchers not just in terms of its linguistic content, but also its multimodal context and discursive structure.<sup>1</sup>

While different subreddits share generic structures and affordances of Reddit and may even in some cases codify communicative norms in similar regulations, it is important to note that Reddit (sub-)communities are not subject to identical communicative patterns—within

each community, user types, and individual users communicate following their own norms or even idiosyncratically. Similar to a federalist state, Reddit as a digital space is a pluricentric entity whose norms are governed macro-, meso-, and microscopically, that is, on the level of platform, subreddits, and individual posts—codified in rules and regulations for platform and subreddits and policed by moderators. Together, the sum of all subreddits, but also the sample we focus on in this study, comprises an ideal sandbox to observe the range and variation of multimodal yet text-focused digital communication as it functions on the post-third-wave or post-Phase III (Herring 2019) web.

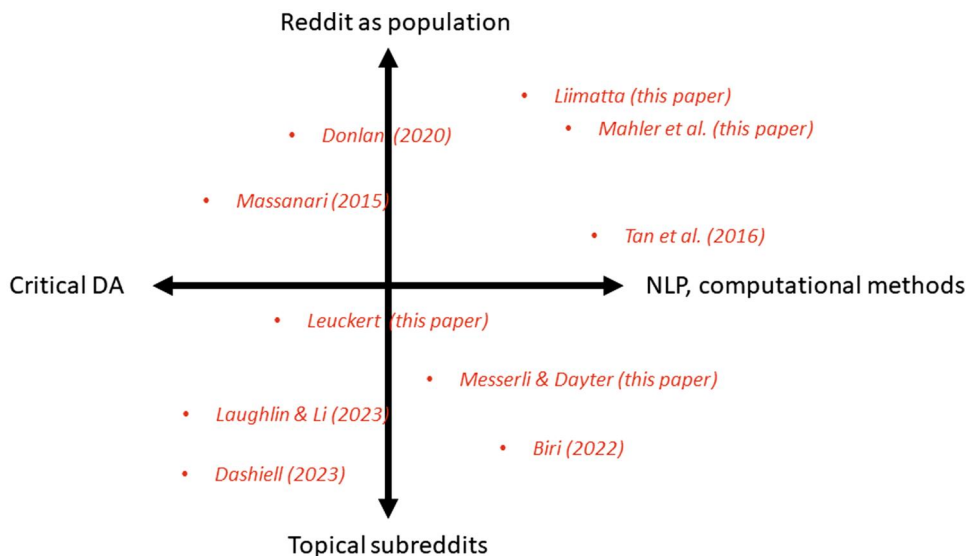
To examine the cosmos (micro- to macro-) of Reddit, digital humanities offer a wide variety of methodological approaches to study who communicates on Reddit, how, and about what. The case studies in this article approach Reddit communication from various angles that have been identified and/or fruitfully realized in Reddit research. We visualize these possible approaches as the *Reddit language research space* in Fig. 1, with a small selection of studies—including those we include in this article—that by no means cover all of the communication-related research on Reddit, but serve as examples of how different methods and interests may position research in this space.

Flattening the complexity of the research space to two dimensions, we broadly identify two axes of scholarly interest in Reddit data: methodological and content related. On the methodological axis, the studies interested in Redditesque communication range from large-scale, Big Data work using predominantly statistical methods or machine learning, to small-scale

qualitative analysis in the tradition of Critical Discourse Analysis, involving manual annotation or close reading of samples that are sometimes as small as a single post. On the second axis, research interest ranges from looking at Reddit as one population (similar to the studies of, e.g., ‘English language’ or even ‘language’ as a whole, with Reddit merely as a sample of such language); to choosing single Reddit communities or thematically-related subreddit bundles with the aim of studying how, for example, communication about politics functions.

For example, the monograph located in the most ‘qualitative’ position in the upper left quadrant, Massanari (2015), takes an ethnographic approach to study how redditors communicate through play and humour, how they use communication to form a community, and what they say about gender and politics. Although Massanari necessarily uses particular communities as examples, her work aims to make statements about Reddit as a whole and even about geek culture itself. This interest distinguishes her from the qualitative work by Dashiell (2023), who uses critical discourse analysis to study the topic of military sexual trauma (MST) and draws specifically on subreddits that discuss men as victims of MST since this is one of the very few sources of such data not affected by observer’s paradox. Similarly, Laughlin and Li’s (2023) interest is not in Reddit as a platform and its inner workings, but specifically in ‘mommy blogging’, which is instantiated on certain thematic subreddits.

In the upper right quadrant, we find methodologically motivated studies such as Liimatta (Section 5) that treat Reddit as a convenient data source to obtain social media



**Figure 1.** The Reddit language research space

texts of different lengths, or Mahler *et al.* (Section 4) that ask conceptual questions about Reddit as a whole. This conceptual level of interest is similar to that of Donland (2020), who studies lexical diffusion and community formation with Reddit as a convenient data source, but additionally resorts to manual coding along with corpus-driven methods.

Finally, the remaining two studies in this article, Dayter and Messerli (Section 2) and Leuckert (Section 3) engage closely with the specific subreddits they use. Dayter and Messerli, despite their broad interest in how persuasive discourse functions, work with a specific communicative environment of the *r/Changemyview* subreddit that operates under strict interactional rules, making it different from other linguistic environments. Leuckert is interested in expert-lay communication more generally, but his study is closely tied to the ways Reddit users on a mountaineering subreddit communicate, without attempts to extrapolate the findings from this very specific community.

The four studies presented below explore discourse(s) from different theoretical perspectives, but all with the objective of contributing to the understanding of Reddit's own communicative culture as well as the exploration of digital practices more generally. We investigate aspects of digital culture and participatory culture, manifested in the communicative acts of different (sub-)communities on the online social platform Reddit. These subreddit-communities and the digital genres they give rise to are sites of linguistic innovation as well as of new debating practices—from navigating knowledge asymmetries on *r/Mountaineering* to the vying for delta awards on *r/changemyview*. They raise questions of methodology, such as the understanding of text length as both a challenge for research and a motivated choice of text authors and they let us gain insights into the relationship between register and online genre. Finally, the employment of a mixed-methods design helps obtain results that are both driven by big data as well as by in-depth understanding of individual and collective communicative acts.

## 2. Study 1: Share my view: Harmonious debating culture on *r/changemyview* (Daria Dayter and Thomas C. Messerli)

Within social media more generally and Reddit more specifically, the subreddit *r/Changemyview* (CMV) holds a special place: Rather than following the trends of outrage culture and polarization, *r/Changemyview*'s original posters invite commenters to change their view through respectful debate. The interaction norms are codified in extensive subreddit rules (tokens  $N = 7338$ ), which specify that a view needs to be

backed with argumentation and the poster genuinely open to changing their view. As an incentive to comply with these norms, social capital in the community is awarded through the 'delta' validation system. A delta marker is obtained if the commenter successfully achieves a 'change of view' of a thread participant following the community rules. Due to the delta system, CMV makes for a unique data source where argumentative threads have been pre-annotated by the participants as successful or unsuccessful.

Existing research on persuasive language in CMV comes predominantly from the fields of computational linguistics and NLP. Such studies identify three promising directions of research: microlinguistic features; argumentative structure; and malleability of opinion. The results are somewhat contradictory when it comes to linguistic features that had been identified as predictors of persuasiveness. Tan *et al.* (2016), for example, in a study that used 2013–2015 CMV data to train logistic regression models, established that successful arguments are less similar to the original post in content words, but more similar in 'stopwords' (such as *the* or *of*). Longer replies are strongly correlated with success. In contrast, Wei, Liu, and Li (2016), training a Machine Learning-based classifier, found that surface text features (number of words, Part-Of-Speech tags, URLs, punctuation) are unimportant, whereas argumentation-related features result in the best performance of their algorithm.

Two further Natural Language Processing studies applied aspects of argumentation theory to detect differences between delta and non-delta posts. Hidey *et al.* (2017) (who reused Tan *et al.*'s 2016 dataset) found that delta comments rely less on rational evaluations and more often combine emotional and logical arguments. Priniski and Horne (2018) examined the role of citing evidence on CMV (references to URLs or use of statistical language) and found that more evidence is cited in sociomoral contexts, that the likelihood of changing views is roughly the same in sociomoral and non-sociomoral contexts, but that a delta was more likely awarded when evidence was provided.

Tan *et al.* (2016) tie the frequency of the first person singular pronouns to the 'malleability of opinion'—the readiness of a person to change their point of view. Similar findings were reported by Jhaver, Vora, and Bruckman (2017) on the basis of a different method—participant observation and structured interviews—to find the main motivations for contributing to CMV. The study raised the question whether CMV would inevitably only cater to those who are already open minded, whereas those who are set in their views would simply not post on this subreddit.

Although these reports are undoubtedly relevant when trying to approach the role of language factors in persuasion, they leave ample space for more in-depth study in

the linguistically informed strand of digital humanities research. One such study has been carried out as a part of the project described in this article (Dayter and Messerli 2022). It involved the creation of a large 6-million-word searchable corpus of CMV posts over a 5-year period. The design of the study was contrastive, looking to explore any potential register differences (with a specific interest in formality/informality dimension) between the posts that had been awarded a delta, and non-delta posts. While no linguistic differences between these two groups were identified, the keywords suggested that there might be a topic-related explanation for malleability of opinion—an avenue of research to be explored in a more detailed study.

To sum up, while some work has been done on the specific strategies that commenters use to persuade, little attention has been paid to the question how prepared original posters actually are to change their mind and how this ‘malleability of opinion’ is discursively constructed. This is the interest Dayter and Messerli pursue in the present study. They draw on a subsample of posts from their large CMV corpus (Dayter and Messerli 2022) to explore how authors construct persuadability. Dayter and Messerli posit that the authors of submissions on r/Changemyview pursue two main aims: (1) constructing malleability of opinion; and (2) providing a convincing argument to demonstrate expertise and authenticity. The analysis proceeds in two steps: first, a taxonomy of strategies is developed on the basis of existing pragmatic literature that deals with advice, a speech act that shares the second aim listed above (Miller and Gergen 1998, Locher 2006, Rudolf Von Rohr 2018). This taxonomy is operationalized with discursive moves as coding units and applied exploratively to CMV submissions. Using an iterative manual annotation process, a small sample of 18 CMV posts proved sufficient to establish a clear rhetorical structure and to identify a robust inventory of typical discursive moves.

In the second analytical step, Dayter and Messerli zoom in on one of the strategies, concession, that can be identified as particularly interesting for CMV and is frequent in CMV original posts across topics.

**Table 1.** Incidence of discursive moves in the CMV subsample.

Move	Dispersion (N = 18)
CMV directive	100%
CMV statement	100%
Pro context	78%
Contra negative evaluation	67%
Pro positive evaluation	39%
Elaboration pro context	83%
Concession	61%
CMV invitation	22%

Table 1 presents the main discursive moves and their incidence in the posts in the small manually annotated sample (e.g., concessions were found in 61% of all original submissions in the sample).<sup>2</sup>

The moves ‘CMV directive’ and ‘CMV statement’ are unique to the subreddit and compulsory for every post. The first one is purely formal, while the second one fulfils dual format- and content-oriented functions by marking the beginning of an original submission and succinctly summarizing the view to be changed. In ‘pro context’, the authors provide background information on the stated view (corresponding loosely to ‘general information’ in advice, see e.g. Locher 2006: 117). Further reasoning as to why the author holds this view and considers it the most applicable can be presented in ‘elaboration of pro context’ (corresponding to ‘explanation’ in advice, *ibid.*). Two evaluative moves, ‘pro positive evaluation’ and ‘contra negative evaluation’, explicate the author’s stance on their own (pro) view and hypothetical alternative (contra) views, corresponding to advice’s ‘assessment’. The concessive move (corresponding to ‘disclaimer’ in advice) constitutes a special kind of evaluation of the author’s view, which aims to pre-empt criticism or counterarguments in various ways: admitting the existence of alternative views or contradicting evidence, incompleteness of information, or the author’s lack of expertise. Despite the fact that concessive moves are thus a conceptually narrower category than categories such as context or evaluation, they were surprisingly frequent in submissions. Finally, the post is bracketed by the concluding move, CMV invitation (‘if anyone has good counterarguments or resources, I’m all ears’).

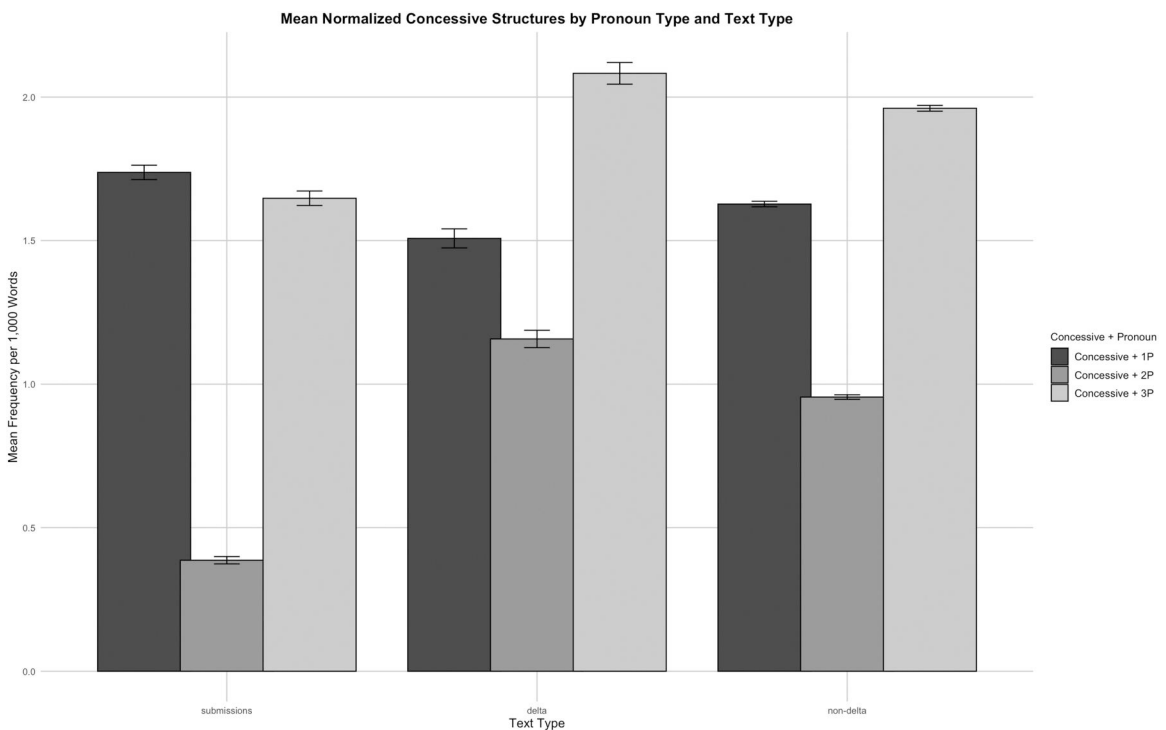
Table 2 illustrates most of these moves on the example of a submission from the CMV submission corpus (as one can see, several moves are optional and have not been used here).

As the final step of the analysis, corpus linguistic methodology was used to examine the concession move—a key strategy in constructing oneself as a reasonable and persuadable poster. A great proportion of concession moves can be automatically extracted from the corpus since they incorporate concessive connectives (*even though, though, although, even if, while*). Concessives, on the pragmatic level, mark ‘precisely those areas where speakers have something to gain or lose by their addressee’s acceptance or rejection of what they are saying’ (Hoye 2005: 1484) and are therefore of particular interest in studying the discursive means of obtaining social capital.

In order to examine the object of concessions in CMV postings, the extracted concessive moves were split into three categories depending on whether the connective is followed by a first-, second-, or third-person pronoun (e.g. *although I; although you;*

**Table 2.** An excerpt from an original post in r/changemyview annotated for the main discursive moves.

1	CMV:	CMV directive
2	<b>I don't believe suffering the death of a spouse is inherently more painful than being left by your spouse.</b>	Title = CMV statement
3	There seems to be a much stronger sympathy in general for widows/widowers than for people who survived an unwanted divorce,	Pro context
4	while I think each situation is different	Concession
5	I don't think one is necessarily worse than the other.	Pro context
6	A divorce doesn't necessarily mean the relationship was not good (especially for the party who wants to stay in the marriage).	Contra negative evaluation
7	The aftermath for those with children can farther substantiate my argument as the person now has split time with their children	Pro positive evaluation
8	which can be very painful.	Pro positive evaluation



**Figure 2.** Concessives followed by personal pronouns in the CMV corpus, frequency per 1,000 words

*although they*). Such clustering served as a proxy for the concessive moves modifying the author's own view; the imagined audience's or addressee's view; or the views of an imaginary third party or opponents' group. To complement the findings on concessions in original submissions in the corpus, Dayter and Messerli obtained this data for delta and non-delta comments as well (see Fig. 2). Not making assumptions about normal distribution of the data, the authors conducted non-parametric statistical tests to establish the significance of differences between the three types of concessive + pronoun structures in each

type of texts. Kruskal–Wallis tests provided evidence of a significant difference in all three test types.<sup>3</sup> Pairwise Wilcoxon rank sum tests confirmed that all differences between pairs of structure (e.g. concessive + first person vs. concessive + second person) in each text type where significant.

It is important to note that this method has significant limitations, as some of the forms used to extract the relevant moves may be conceivably used in functions other than concessive ones (e.g. *while*); similarly, a cluster of concessive + pronoun may have other function than indicating a modification of the pronoun

referent's belief. The study needs to be developed by adding manual annotation of a sample of automatically extracted data to determine the error rate of such assumptions.

The first study finds that commenters on CMV qualify third party beliefs more often than their own and that they use concessive structures least often with their second-person interlocutors. Original posters also use concessives in third person sentences, but tend to qualify their own beliefs even more often and only rarely contain qualifications by means of concessives in structures with second-person pronouns. Successfully persuasive delta-awarded comment (DAC) writers make concessions about interlocutors' beliefs more than non-DAC writers. However, most frequently they talk about an imaginary third party and qualify those beliefs—a discursive strategy which may be less face-threatening than addressing interlocutors with a second person pronoun, and therefore more likely to garner support. Finally, non-DAC writers qualify their own beliefs more than DAC writers, a discursive behaviour that lets us tentatively posit that a lack of certainty may make a persuader less successful.

To summarize, Study 1 established that original submissions to r/Changemyview follow a stable set of discursive moves: CMV directive and statement, pro-context, elaboration of the pro-context, positive evaluation of own view and negative evaluation of contra view, concession, and CMV invitation. While not all of these moves are compulsory, all of them besides CMV invitation and pro positive evaluation occur in the majority of posts in the sample. Zooming in specifically on the concession move, since its pragmatics function is to mark the high-stakes element of persuasive discourse, the study examines the frequencies of concessives co-occurring with personal pronouns. Preliminary results indicate that the writers of original submissions often qualify their own beliefs, thus constructing their own identity as persuadable and having malleable opinions.

### 3. Study 2: Expert-lay communication on r/Mountaineering (Sven Leuckert)

The second study derives from Leuckert's project on expert-lay communication on r/Mountaineering, a subreddit by and for (mostly) high-altitude climbers. The research questions of this project are:

- 1) Which pragmatic strategies do interactants on r/Mountaineering employ to balance knowledge asymmetries and how frequently are the different strategies employed?
- 2) Which of these strategies lead to successful communicative outcomes and which do not—and what are potential explanations for differing outcomes?

This section gives an overview of the project as well as first qualitative findings, as the quantitative analysis is still work in progress. Part of the reason why this investigation is relevant is that, despite the increasing global popularity of mountaineering, there is no official terminology or guidelines that climbers may rely on (see [Leuckert 2024](#) for a discussion). This means that significant portions of mountaineering knowledge, relating to both the field's specialized register (in the sense of [Biber and Conrad 2009: 31](#)) and various other aspects of the sport, are typically acquired individually or, for instance, by learning from local communities and guides. This also means that knowledge levels between mountaineers may vary drastically, and knowledge may be based on local guidelines as opposed to universally agreed-upon 'facts'.

While, on the one hand, a layperson's knowledge typically constrains the conceptual planning and the language planning of experts ([Hayes and Nash 1996](#)), 'asynchronous communication allows for the careful planning of a message before it is sent [...] and gives] time to reflect upon a communication partner's background knowledge and communicational needs' ([Nückles et al. 2006: 92](#)). A previous study by [Rudolf von Rohr, Thurnherr, and Locher \(2019: 233-234\)](#) has identified a range of strategies that users of a self-help forum for smokers who would like to quit make use of in order to frame themselves as experts:

- status-related (e.g. address terms; self-reference)
- information-related (e.g. referring to further help; use of jargon)
- assessment-related (e.g. criticism; personal experience)
- interpersonal (e.g. showing empathy; expressing confidence)
- interactive (e.g. requesting information from advice-seeker)

Strategies used to indicate a lower degree of expertise may range from hedges over 'oh-prefacing' ([Heritage 1998](#)) to overtly describing a lack of knowledge in a given context. While the strategies employed between experts and laypeople obviously differ, all participants in expert-lay exchanges effectively negotiate epistemic stance, that is, their attitude towards a certain object of knowledge. The role of stance in online communication, and specifically on Reddit, is highlighted by [Kiesling et al. \(2018: 710\)](#), who point out that '[s] peakers and listeners often disagree about whether a single utterance is rude, offensive, or any other interpretation'. Apart from making additional studies in this area worthwhile in general, it also adds further relevance to investigating expert-lay communication in the context of mountaineering. As mentioned above, many aspects of mountaineering do not represent

'streamlined', official knowledge, which means that they are more likely to be subject to (potentially controversial) debate (as the first example below illustrates). A successful communicative outcome would be, for instance, if a user receives an answer that pertains to the question they asked (for instance, if they asked for information on climbing gear), or if their input to a thread is seen as trustworthy and valuable.

The subreddit used for this study is *r/Mountaineering*, which had 178,067 active users as of 25 May 2023; the subreddit was crawled using Python and spans all contributions posted between 2012 and 2021 amounting to a word count of ca. 4.9 million. Relevant pragmatic strategies in the corpus will be identified by close reading of ten randomly selected threads per year in order to establish a set of (ideally recurring) linguistic expressions that can then be searched for using AntConc. After identifying exchanges in which knowledge asymmetries are balanced by interactants, the employed strategies will be identified and classified in a similar fashion to the framework developed by Rudolf von Rohr, Thurnherr, and Locher (2019). Quantitative analyses will mostly focus on descriptive statistics in order to provide an overview of which strategies are popular among the *r/Mountaineering* community; the focus, however, will be on discourse-analytic descriptions of the exchanges.

While there is an abundance of affordances provided by CMC, there are also challenges for both communication between users and researchers analysing this communication. First, non-verbal and paraverbal cues are absent from Reddit. Emoji, images, and other communicative tools may be used to fill in for prosodic tools and gestures, but they are hardly able to fulfil the same role. This means that users are required to plan their contributions rather carefully in order to avoid disrupting an exchange. Second, 'expert' and 'layperson' exist on a continuum; they are not strictly binary terms. A user coming to *r/Mountaineering* may have reached a high level of expertise in one specific area related to climbing but lack expertise in another, or they may simply be interested in exchanging opinions with others at a similar level of expertise.

To illustrate the kind of discussion in focus in this project, consider the following exchange about the meanings of the terms 'alpinism' and 'mountaineering' (usernames have been removed):

(3.1) Eh that's alpinism (hence the name alpinist), mountaineering is more glaciated walk-ups with the occasional step of easy technical climbing. They are closely related but when you start climbing multiple continuous pitches of hard rock/ice/mixed pitches it becomes something different.

(3.2) Huh? That is NOT the definition of 'mountaineering' in any dictionary but nice try trying to create your own glossary, I guess.

The first user begins their contribution with a discourse marker that here likely functions as a turn-taking device that starts the supposed correction of a previous comment. Noticeably, all copulas in the post are forms of 'to be', indicating that the information presented is factual and not opinion-based; no hedging is used at any point. The response to this post uses multiple strategies to convey their disapproval, starting with the interrogative 'huh?' that may be used, for instance, when 'one has not clearly heard what someone just said' (Dingemans, Torreira, and Enfield 2013). However, while 'huh?' may potentially invite further clarification, in this case, the user goes on to express that, again in supposedly factual terms, the provided definition is not part of any dictionary, suggesting that 'huh?' expresses disbelief and surprise in this instance. The use of sarcasm by the second user ('nice try') may be taken as an offence and effectively shut down the conversation. Both users suggest that they have a high degree of expertise regarding how core terms such as 'alpinism' and 'mountaineering' need to be defined. Interestingly, defining these terms seems to be an aspect that, arguably, would be expected to be part of the common ground on the subreddit. While both users employ what could be classified as 'information-related' strategies in Rudolf von Rohr, Thurnherr, and Locher's (2019) taxonomy, the first user does not relate to any official information, while the second user does—albeit only vaguely by referring to 'any dictionary'.

A second example, this time of a person framing themselves as lacking knowledge on a certain topic, is provided below (the example has been shortened for reasons of space):

### (3.3) Double boots—am I making the wrong choice?

TLDR: Boreal G1 Lite—a solid double boot?

Hello everyone! First of all I just want to thank everyone in this group for the amazing resources and insights that you all share, it's the best place in the internet for everything mountaineering.

[...]

I would love if anyone could give any pointers or suggestions (unfortunately, here in London UK, there are few places with double boots for sale that I can try in store, which has been very limiting for me).

Thanks!

In their request for feedback on suitable boots, the user in this example uses a range of strategies to appeal

to the other users and increase the likelihood of a successful communicative outcome. The title of the post is phrased as a question and indicates that the user is, in fact, unsure of their decision. Thus, both responses—‘yes’ and ‘no’—would be acceptable, and having experts step in is explicitly invited. This is followed by an introduction, including a greeting, which has varying degrees of propositional content across linguacultural contexts (cf House *et al.* 2022) but is, at the very least, indicative of courtesy (Searle and Vanderveken 1985: 215–216). The post continues with a positive evaluation of the subreddit’s contributions, suggesting that the subreddit is indeed a community with shared goals (see Leuckert and Leuckert 2020). In the second part of the excerpt in (3.3), the user again invites feedback, explains their situation, and shows politeness by thanking others in advance for their contributions. Overall, this contribution is marked by the user indicating an absence of knowledge in a specific area; they are seeking advice by combining multiple politeness strategies and carefully phrased speech acts.

The contrast between the exchange in (3.1) and (3.2) and the construction of the post in (3.3) illustrates that investigating the strategies employed in expert-lay communication on Reddit has great potential. Developing a better understanding of these strategies and which outcomes they lead to will provide an important piece to the puzzle of how Reddit works and how knowledge asymmetries are balanced in the absence of non- and paraverbal signals. R/ Mountaineering represents an ideal example in this context, since members of the mountaineering community may range anywhere from absolute beginners to seasoned experts in the field. In addition, r/ Mountaineering (and similar climbing-related subreddits) may represent the only ‘truly’ international forum on the topic, which means that people at different levels of expertise, but also from different countries and potentially vastly different climbing contexts, converge on the subreddit.

#### **4. Study 3: Register variation in Reddit comments—A multidimensional analysis (Hanna Mahler, Axel Bohmann, Gustavo Kozma, Rafaela Tosin)**

The register status of Reddit discourse—its differentiation from other registers as well as its internal variability or homogeneity—is a question of current CMD (computer-mediated discourse) research. In the present contribution, Mahler, Bohmann, Kozma, and Tosin describe their ongoing work in this context, which aims to answer the following research questions:

1. Which dimensions describe the linguistic variation found on the platform Reddit?

2. Can the topical sub-fora (‘subreddits’) reasonably be described as different registers according to the result of a multi-dimensional analysis?

3. Which other factors help to explain register differences between individual comments?

The project adds to the current state of knowledge by providing an extension of Biber’s (1988) multi-dimensional analysis (MDA) in Python, implementing an innovative approach to bot-removal, and considering register variation at the fine granularity of individual comments (as one of the first studies to do so).

The study is concerned with the variation in register (defined as ‘a variety associated with a particular situation of use’, Biber and Conrad 2009: 6) on Reddit. One of the central approaches to describe register variation is the MDA framework (Biber 1988), which uses factor analysis to identify frequently co-occurring groups of linguistic features within a corpus of texts. These groups of features are interpreted in functional terms. Originally designed for the description of spoken and written language, the approach has since been applied to different varieties of English (Bohmann 2019), languages other than English (overview in Biber 2014), and to language change (e.g. Biber and Finegan 1989). The method has also been utilized in computer-mediated contexts (e.g. Berber Sardinha and Pinto 2014; Biber and Conrad 2009; Biber and Kurjian 2007; Biber and Egbert 2016; Clarke and Grieve 2017; Bohmann 2020; Liimatta 2016; Grieve *et al.* 2010). Especially relevant for the study at hand is Liimatta (2016), who performs MDA on a small number of selected subreddits. He finds that ‘there are register differences between subreddits’ (2016: 67) along the dimensions ‘Present-Abstract vs. Past-Narrative Focus’, ‘Personal vs. Factual Focus’, and ‘Informational vs. Involved Style’ (2016: 68). His level of analysis is individual subreddits in their entirety, and he raises the important question whether more meaningful variation might be found on the comment level (2016: 68).

Previous studies show how the development of CMC formats gives rise to new types of registers (Bohmann 2020, Biber and Conrad 2009, Biber and Egbert 2016, 2018), whose linguistic properties are gradually, rather than radically, distinct from offline registers (Berber Sardinha and Pinto 2014). Register studies of online discourse also provide support for the alleged universality of some of Biber’s original dimensions (Biber and Egbert 2016: 130, Biber and Egbert 2018: 70, Berber Sardinha 2018). They furthermore emphasize that the internet contains a variety of different registers and cannot be regarded as homogeneous (Biber and Conrad 2009: 208, Biber and Kurjian 2007). The same is true for discourse on individual

platforms (Bohmann 2020, Clarke and Grieve 2017, Grieve *et al.* 2010, Liimatta 2016).

Existing research on Reddit already indicates considerable linguistic variability. While Liimatta (2016), Mahler (2020), and Cole, Ghafurian, and Reitter (2017) emphasize linguistic differences between subreddits as distinct communities, Flesch (2018) illustrates variation by socio-demographic characteristics, especially ethnicity and age. The question of how variation between subreddits intersects with user-specific (e.g. socio-demographic) variation remains largely unaddressed. In order to investigate this aspect of register on Reddit, an analysis at a finer level of granularity than whole subreddits is required. This is the perspective Mahler *et al.* aim to develop in this project.

The data used in this study was taken from the 'Pushshift Reddit Dataset' (Baumgartner *et al.* 2020). The dataset is divided into two components: submissions (known as 'posts') and comments. For the study at hand, only the comments were used. The data is fully open access, and the authors see no ethical problems as they are not focusing on the linguistic profile of individual users, but are interested in stylistic differences on the platform as a whole. Such large-scale approaches are in line with internet user's preferences (as identified by Fiesler and Proferes 2018: 6).

The analysis by Mahler *et al.* focuses on variation among comments. However, there is considerable variation in comment length, and this factor has been shown to influence results of an MDA (Liimatta 2016; see also Section 5). This issue is addressed by adopting a modified version of short-text MDA (Clarke and Grieve 2017), in which every sentence is treated as a text. Details about the specific implementation follow below.

As a first step, a script cleans the data. This includes removing messages that were sent by bots (by checking each comment against a list of usernames which have been suspended by Reddit due to violations of the platform's content policy), deletion of messages where the body text is only "deleted", and removing non-English comments (by checking the number of words that appear in a standard dictionary of English). Apart from this, no further standardization is implemented at this point, since the tagger is expected to handle common spelling mistakes, etc.

As a next step, every sentence is part-of-speech (POS) tagged. Accurately tagging internet data poses considerable problems for most common POS taggers (Eisenstein 2013). Since precise tagging is a crucial requirement for the accurate identification of linguistic features, a FLAIR POS tagger (Akbik, Blythe, and Vollgraf 2018) was trained using a manually tagged dataset of Reddit data from Behzad and Zeldes (2020). The FLAIR tagger was trained over 150 iterations of the dataset, which consisted of a total of 11,182 tokens. The performance of

the FLAIR tagger and of the commonly used Stanford tagger was compared against a set of 400 sentences (randomly sampled from Reddit) manually tagged by members of the project team. Since the two taggers employ different tokenization strategies, it was not possible to calculate simple accuracy measures. Instead, the output was compared by means of Levenshtein distance, that is, the number of (delete, insert, replace) operations minimally required to bring an automatically tagged sentence in line with its human-tagged gold-standard version. Results show that the NLTK tagger achieves an average normalized Levenshtein distance of 0.282, whereas the FLAIR tagger's mean is 0.226, where lower values indicate better performance. The difference in means is statistically highly significant, according to a paired t-test ( $t = 7.144$ ,  $df = 394$ ,  $p\text{-value} = 4.415e-12$ ). These results confirm that training on in-domain data is indeed essential for social media data.

As a next step, linguistic features were extracted. This was done based on Biber's (1988: 71–72) original set of features but including slight modifications to the precise operationalization for some of the features and addition of several new ones. Many of these additional features are taken from Clarke and Grieve (2017); this includes hashtags, URLs, capitalization, imperatives, comparatives, and superlatives.

Mahler *et al.* consider it an important step to empirically test the accuracy of the feature extraction script, something that existing MDA studies rarely do. This study is therefore one of the first MDA studies to conduct a detailed test on the accuracy of the feature extraction process. First, a sample of a total of 1,251 sentences was randomly selected from the Reddit comments corpus. For each feature, the project team manually determined whether a sentence contained the feature or not. The accuracy was then calculated in the form of an F-score, calculated from precision and recall. All features with an F-score lower than 0.6 were discarded from the analysis. At the moment this leaves sixty-six features, with several operationalizations still awaiting coding improvement.

In his original study, Biber (1988) applies exploratory factor analysis (Gorsuch 1988) to the matrix of normalized feature frequencies by texts. Since Reddit comments differ from Biber's original texts in several ways (especially in length), this is not feasible for the study at hand. Instead, short-text MDA (Clarke and Grieve 2017; Clarke 2022) was used. Rather than estimating normalized text frequencies, this method codes for the presence or absence of each feature in a sentence and employs multiple correspondence analysis (MCA) to develop dimensions of variation. When the aim is to develop an MDA for an extremely large corpus, the additional problem arises that storing information about each sentence in the data is beyond the

memory capacity that most researchers have access to. This problem was solved by only updating the Burt table, that is, the matrix of all two-way cross-tabulations between the categorical variables, from which an MCA can be constructed. The advantage of this approach is that it easily scales up to very large amounts of data. At any point in the analysis, only one sentence needs to be processed by the script, and the only permanent data structure that is required to be stored is a  $N^*2 \times N^*2$  Burt table (where  $N$  is the number of features). The disadvantage is that information about individual sentences is not retained. As such, the analysis is able to develop the general dimensions of variation on Reddit, but for the analysis of a specific set of data (comments from one subreddit or the discourse of a user across subreddits, etc), an additional, more targeted round of feature extraction is required.

An alternative method for addressing variable comment lengths is Liimatta's (2020) approach of length-wise scaling. Here, feature frequencies are not normalized by dividing number of occurrences by number of words; instead, feature counts are compared across texts of the same lengths. This is an elegant solution, but it is unable to account for any systematic differences between texts of different lengths.

The methodological steps reported above constitute ongoing work. While Mahler *et al.*'s feature extraction script is largely complete, some individual features still require optimization for precision and recall. The downstream steps—constructing a Burt matrix and implementing MCA on its basis—have been implemented, but as of yet, the authors have not processed significant amounts of Reddit posts to construct a full MDA. However, the scripts for all individual steps are publicly accessible in their current state of development (at [https://github.com/PythonGoesReddit/Reddit\\_MDA/](https://github.com/PythonGoesReddit/Reddit_MDA/)).

Once relevant dimensions for (a subset of) Reddit comments are identified, these can be compared to the existing analysis of Reddit by Liimatta (2016: 39) and other studies investigating online discourse. Mahler *et al.*'s approach to quantifying register variation in Reddit data opens up many interesting avenues for further research. Among them are the investigation of diachronic change and register diversification as the platform develops, the propagation of linguistic features over time, and the linguistic development of individual users over their life cycle on the platform.

## 5. Study 4: Functions of text length on Reddit (Aatu Liimatta)

Reddit datasets have properties which make them extremely useful for shedding light on various questions

and issues which were very difficult or even all but impossible to study before the advent of large-scale datasets of social media and other internet texts.

One of these issues is the role which text length plays in linguistic variation. In a lot of quantitative corpus-linguistic research, variation in the length of the texts in the analysed dataset has mostly been considered a confounding factor, something which can cause problems unless it is dealt with in some way. However, using the kinds of data available from Reddit can turn this significant issue into a research question.

For instance, Liimatta (2022a, 2022b) hypothesizes that the length of a text is not entirely random; instead, a text must be of a length which the writer has considered suitable for the intended function of the text. In order to study a potential connection between the length of a text and its function, Liimatta (2022a, 2022b) uses a methodology inspired by ideas from the multi-dimensional method of register analysis, originally developed by Biber (1988), and from the text-linguistic framework, the theoretical framework surrounding the multi-dimensional methodology (cf Biber, Egbert, and Keller 2020).

If we take the hypothesis by Liimatta (2022a, 2022b), that text length is connected with the function of the text, and combine it with the assumption, supported the body of research on register variation, that the frequency of functional linguistic features varies by the function of the text, then it should also be the case that the frequency of functional linguistic features varies by the length of the text. Furthermore, this means that the functional tendencies of texts of different lengths can be examined by analysing the observed variation in the frequencies of the linguistic features between texts of various lengths.

These resulting hypotheses would be very difficult to study using more traditional or typical linguistic corpora. In order to be able to analyse quantitatively the relationship between the length of the text and the functional linguistic features appearing in the text, the textual dataset needs to fulfil two main conditions. The dataset needs to cover a wide enough range of text lengths so that the differences between texts of different lengths become obvious enough in the analysis. Additionally, it is not enough to simply cover a range of text lengths. There also need to be enough texts of similar lengths so that it is possible to say something more general about the texts which are close to each other in length. However, traditional corpora and even many smaller internet datasets are relatively sparse in terms of text length coverage, that is, the range of text lengths covered by them is not necessarily that broad, and the number of texts of any specific length in them is not that large.

On the other hand, textual data from Reddit is particularly well-suited for this kind of analysis. While most Reddit comments, like most social media comments overall, tend to be extremely short, Reddit comments cover a wide range of lengths, from single word responses to extremely long and detailed postings. Moreover, Reddit data is plentiful. For instance, the Reddit dataset Liimatta (2022a, 2022b) makes use of covers all publicly available Reddit comments from only a period of one month, but the cleaned dataset of this relatively short time period alone comprises close to 80 million individual comments. Furthermore, Reddit data has been relatively easily available compared to many other social media platforms, does not have artificial length restrictions like Twitter, and is largely public unlike e.g. Facebook postings.

Liimatta performs the analysis by comparing the average frequency of over sixty functional linguistic features, commonly used in register studies, across comments of every individual length on Reddit.<sup>4</sup> In Liimatta (2022a), this analysis is performed on all of Reddit as a whole. The analysis shows that there indeed appear to be Reddit-wide functional differences between texts of different lengths. For instance, longer comments contain more informational features such as nominalizations and fewer colloquial features such as contractions, suggesting that longer comments tend to be more informational and edited, whereas shorter comments tend to be more casual and unedited. Longer comments are also found to include more narrative features such as past tense verb forms and third-person pronouns, showing that on Reddit overall, narrative content tends to favour longer comments.

In a follow-up study, Liimatta (2022b) then zooms in on a handful of subreddits. The study shows that not only does text length have a functional role, but that the role of text length can differ in different registers and contexts of use. For instance, in most of the subreddits analysed in the study, first-person pronouns are most common in the very shortest comments, suggesting that content involving the commenter tends to be relatively short, but in e.g. *r/AskReddit* the frequency of first-person pronouns stays at a high level even in longer comments, reflecting the fact that the commenters on the subreddit are often personally involved even in their longer answers somehow, such as relating a past experience or a personal opinion. Similarly, past tense forms are found to either stay level or even decrease slightly in most subreddits as the comments become longer, but in *r/AskReddit*, their frequency increases with text length, again reflecting the tendency of the longer comments on the subreddit to be narrative comments with various kinds of anecdotes.

Perhaps the most strikingly different subreddit included in the analysis in Liimatta (2022b) is

*r/RocketLeagueExchange*, a subreddit dedicated to making item trade deals in the multiplayer game Rocket League. Based on the analysis, this subreddit has a dual nature, with at least two clearly different registers which are also differentiated by comment length. The shorter comments on *r/RocketLeagueExchange* tend to be trade negotiations. This is seen in, for example, the fact that the shorter comments on the subreddit have a much higher frequency of the conditional adverbial subordinators *if* and *unless* compared to all of the other subreddits studied, but their frequency quickly decreases as the comments become longer. In the longer comments, they are replaced by, for example, nouns, reflecting the fact that longer comments on the subreddit are commonly either advertisements of items one wants to get or is ready to part with, or listings of completed trades.

Liimatta (2022a, 2022b) shows that text length appears to be linked to the function of the text and, more generally, that text length can be an important factor in linguistic variation. While Reddit data could be expected to show the connection between the length of a text and its function particularly well, the effect could be expected to also be present in most other kinds of texts as well.

## 6. Conclusion

The four studies discussed above represent a small slice of communication-focused research on digital debating cultures: they range in their focus from research method to linguistic resources available to digital debaters, and in their choice of subreddits from explicitly debate-centred to general interest or even Reddit as a whole. Despite such variety, these studies, and other digital humanities work on Redditesque communication, can be seen as united by exploring posts and comments to subreddits as situated language use in a particular multimodal setting and as a type of CMC. The communicative acts on this platform are further understood as user-generated content, a part of social media discourse, the result of what users individually and collectively make of the particular technological affordances Reddit offers, and as communication subject to the norms that the Reddit community and the subreddit communities jointly negotiate, set and codify.

We argue that the Reddit language research space proposed in Section 1 of this article can help digital humanities researchers to make sense of this extreme variety of theoretical vantage points and methodologies. Although, as any model of this kind, it is necessarily very reductive, it can serve as a helpful guideline in choosing the focus for one's study and in reviewing existing literature, as the space becomes more populated with relevant citations. As members of the CopRe research group (Communicative Practices on Reddit, <https://copre.org/>), we call for mutual

engagement among different disciplinary strands of Reddit language research to avoid compartmentalization into the siloes of social studies, communication studies, linguistics etc. In its self-proclaimed status as ‘the front page of the internet’, Reddit constitutes a treasure trove of easily accessible, naturally occurring data in forty-four languages<sup>5</sup> on thousands of topics and illustrates a multitude of communicative practices. There are also important issues that the research community would do well to address together, for example, the questions of research ethics related to collection, publication, and storage and reuse of Reddit corpora. Engagements with these questions is often absent from the computational research; while more qualitative strands of research struggle with replicability or scaling their analysis to larger samples. We believe that creating more synergies in Reddit language research in neighbouring disciplines would benefit all types of research by making it more robust, ethical, efficient, and insightful.

## Author contributions

Thomas Messerli (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision), Daria Dayter (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Visualization), Sven Leuckert (Formal analysis, Investigation, Methodology), Aatu Liimatta (Formal analysis, Investigation, Methodology), Hanna Mahler (Formal analysis, Investigation, Methodology), Axel Bohmann (Formal analysis, Investigation, Methodology), Gustavo Kozma (Formal analysis, Investigation, Methodology), Rafaela Tosin (Formal analysis, Investigation, Methodology).

## Notes

1. Since online, digital communication is of course not only mediated through prototypical computers, other terms like technology-mediated communication (TMC, see [Chovanec and Dynel 2015](#)) or keyboard-to-screen communication ([Jucker and Dürscheid 2012](#)) have been suggested, but we follow here the more widely used term CMC, adopting a broad understanding of ‘computer’ as including any stationary or mobile devices with which—in this case—Reddit can be accessed as a site of communication. Typically, computers will in this case thus be devices that allow communication on Reddit via either a dedicated application (e.g. smartphones or tablets running Android, iOS, iPadOs, etc.) or a web browser (see also [Herring 2019](#)).
2. The dispersion, that is, the relative frequency of moves in the sample, is not of great importance for the purposes of this study. However, it allows a distinction between obligatory, frequent, and rare moves.
3. **Submissions:**  $\chi^2(2) = 3909.3$ ,  $SE = 0.02$ ,  $P < .001$ ; **Delta-awarded comments:**  $\chi^2(2) = 729.81$ ,  $SE = 0.03$ ,  $P < .001$ ; **Non-delta comments:**  $\chi^2(2) = 8180.04$ ,  $SE = 0.01$ ,  $P < .001$ .

4. In practice, Reddit data is so heavily skewed towards the shorter comment lengths that even with such a large dataset it was necessary to combine some of the longer text lengths together. However, combining longer comment lengths like this is unlikely to affect the results much as the relative difference in text length in such cases is relatively minor.
5. The Most Popular Languages on Reddit after Analyzing ~1M Comments, 28.6.2023 <https://towardsdatascience.com/the-most-popular-languages-on-reddit-analyzed-with-snowflake-and-a-java-udtf-4e58c8ba473c>

## Funding

Funding support for this article was provided by the University of Basel (Open access funding).

## Data availability

Study 1: The data underlying this article cannot be shared publicly due to the terms of Reddit User Agreement.

Study 2: The data underlying this article will be shared on reasonable request to the study author.

Study 3: The data underlying this article were accessed from <https://pushshift.io>. The derived data generated in this research will be shared on reasonable request to the study authors.

Study 4: No new data was generated or analyzed in support of this research.

## References

- Akbik, Alan, Blythe, Duncan, and Vollgraf, Roland (2018) Contextual string embeddings for sequence labelling. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1638–1649, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.
- Baumgartner, Jason *et al.* (2020) The Pushshift Reddit Dataset, 14th edn. *Proceedings of the International AAAI Conference on Web and Social Media*.
- Behzad, Shabnam, and Zeldes, Amir (2020) *A Cross-Genre Ensemble Approach to Robust Reddit Part of Speech Tagging*. <http://arxiv.org/pdf/2004.14312v1>.
- Berber Sardinha, Tony (2018) ‘Dimensions of Variation across Internet Registers’, *International Journal of Corpus Linguistics* 23: 125–157.
- Berber Sardinha, Tony, and Pinto, Marcia Veirano (eds) (2014) *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber (Studies in Corpus Linguistics 50)*. Amsterdam, Philadelphia: John Benjamins.
- Biber, Douglas (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511621024>
- Biber, Douglas (2014) ‘Using Multi-Dimensional Analysis to Explore Cross-Linguistic Universals of Register Variation’, *Languages in Contrast* 14: 7–34.
- Biber, Douglas, and Conrad, Susan (2009) *Register, Genre, and Style (Cambridge Textbooks in Linguistics)*. Cambridge: Cambridge University Press.



- Journal of Lexicography* 37: 29–49. <https://doi.org/10.1093/ijl/ecd032>
- Leuckert, Sven, and Leuckert, Martin (2020). 'Towards a Digital Sociolinguistics: Communities of Practice on Reddit', in Daria Dayter and Sofia Rüdiger (eds), *Corpus Approaches to Social Media*, pp. 15–40. Amsterdam: John Benjamins. <https://doi.org/10.1075/scl.98.01leu>
- Liimatta, Aatu (2016) *Exploring Register Variation on Reddit: A Multi-Dimensional Study*. Helsinki: University of Helsinki Master thesis.
- Liimatta, Aatu (2020) 'Using Lengthwise Scaling to Compare Feature Frequencies across Text Lengths on Reddit', in Sofia Rüdiger and Daria Dayter (eds), *Corpus Approaches to Social Media*, pp. 111–130. Amsterdam, Philadelphia: John Benjamins.
- Liimatta, A. (2022a) 'Register Variation across Text Lengths: Evidence from Social Media', *International Journal of Corpus Linguistics* 28: 202–31. <https://doi.org/10.1075/ijcl.20177.lii>
- Liimatta, A. (2022b). 'Do registers have Different Functions for Text Length? A Case Study of Reddit', *Register Studies*, 4: 263–87. <https://doi.org/10.1075/rs.22007.lii>
- Locher, Miriam A (2006) *Advice Online: Advice-Giving in an American Internet Health Column*. Amsterdam: John Benjamins.
- Mahler, Hanna (2020) 'Mission Impawssible?': *An Analysis of Community-Specific Orthographic Variation and Lexical Creativity Online*. Freiburg: Albert-Ludwigs-Universität Freiburg Master thesis.
- Massanari, A. (2015) *Participatory Culture, Community, and Play Learning from Reddit*. Lausanne: Peter Lang.
- Miller, John K., and Gergen, Kenneth J. (1998) 'Lifeon the Line: The Therapeutic Potentials of Computer-Mediated Conversation', *Journal of Marital and Family Therapy* 24: 189–202.
- Nückles, Matthias et al. (2006). 'How Do Experts Adapt Their Explanations to a Layperson's Knowledge in Asynchronous Communication? An Experimental Study', *User Modeling and User-Adapted Interaction* 16: 87–127.
- Priniski, John, and Horne, Zachary (2018) 'Attitude change on Reddit's Change My View'. *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* 40: 2276–2281.
- Rudolf Von Rohr, Maire-Therese (2018) *Persuasion in Smoking Cessation Online: An Interpersonal Pragmatics Perspective*. PhD Dissertation, University of Basel. Freiburg: NIHIN.
- Rudolf von Rohr, Marie-Thérèse, Thurnherr, Franziska, and Locher, Miriam A. (2019). 'Linguistic Expert Creation in Online Health Practices', in Patricia Bou-Franch and Pilar Garcés-Conejos Blitvich (eds), *Analyzing Digital Discourse: New Insights and Future Directions*, pp. 219–250. Cham: Palgrave Macmillan.
- Searle, John R., and Vanderveken, Daniel (1985). *Foundations of Illocutionary Logic*. Cambridge: Cambridge University Press.
- Tan, Chenhao et al. (2016) 'Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions'. *Proceedings of the 25th International Conference on World Wide Web*: 613–624. <https://doi.org/10.1145/2872427.2883081>
- Wei, Zhongyu, Liu, Yang, and Li, Yi (2016) 'Is this post persuasive? Ranking argumentative comments in the online forum'. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* 2, 195–200.