Florian Frenken

A multivariate register perspective on Reddit: Exploring lexicogrammatical variation in online communities

Abstract: Even though social media has become a growing topic of linguistic research in recent years, the internal variation associated with emergent register contexts engendered by its various forms remains largely unknown. To address this gap, the present study evaluates a geometric multivariate approach for the domain of online interactions by investigating patterns in visualisations of forty-two lexicogrammatical features derived from systemic functional theory for individual texts from thirty-three communities on Reddit. Their analysis reveals that these so-called subreddits form overlapping clusters in multidimensional feature space that align with their contextual and thus functional (dis)similarities. It therefore becomes possible to interpret them as subregisters with continuous variation inside a hypothesised macro-register of the website at large. As such, this study argues that Reddit's communities reflect the wider internet landscape on a smaller scale, being an online microcosm of sorts that provides convenient descriptive labels for aggregated content hubs that would otherwise be distributed and hard to find. Subreddits generally fulfil varied purposes, often showing similar features to offline registers on top of more general indicators of user interaction. This attests to their status as independent hybrid registers. At the same time, computer-assisted content moderation emerged as a potentially significant factor in shaping the social context of community interaction – an issue whose linguistic implications demand further attention. Overall, investigating text-level rather than averaged feature correlation patterns appears well-suited for multidimensional analyses of subregisters in general and the web specifically as the sensitivity of the geometric approach helps to operationalise linguistic variation at lower levels of instantiation. This contribution therefore hopes to incentivise further research on platform-internal register differences, potentially with even greater granularity, not least for its practical implications in context-informed automatic classifications of web documents.

Keywords: systemic functional linguistics, (sub)register variation, internet language, online communication, geometric multivariate analysis, social media, reddit

Florian Frenken, RWTH Aachen University, e-mail: florian.frenken@ifaar.rwth-aachen.de

https://doi.org/10.1515/9783111434018-005

1 Introduction

The internet has undoubtedly fundamentally changed everyday communication, seeing as how a significant portion of human-human interactions now occur via computer-mediated means. As users learn to navigate this new environment, they face unique communicative contexts to which they must adapt, socially and linguistically. Online forums, for instance, show certain "transient" characteristics of spoken face-to-face conversations in that responses can be edited and deleted or lost to attention despite being expressed in writing (Berber Sardinha 2018: 131). Naturally then, Biber and Conrad (2009: 177) contend, "anyone interested in register variation will wonder how language is used in these new registers." Interestingly, however, knowledge about the linguistic characteristics associated with these innovative forms is rather limited as of yet despite technological advancements that would permit accounting for the wealth of data available (Titak and Roberson 2013: 236).

Surely one of the most prominent new variants of computer-mediated communication (CMC) is social media. As with established "offline" registers like research articles and newspapers, which are not native to the internet, Facebook and Twitter represent the salient "online" registers of today "whose labels are instantly recognized" (Berber Sardinha 2018: 126). By mere exposure, these platforms play a disproportionately high role as communicative contexts, so their relevance should not be understated (cf. Biber and Egbert 2018: 3). This becomes especially pertinent considering that such "registers continue to emerge and evolve quickly, illustrating the dynamic nature of language development online" (Titak and Roberson 2013: 236). As the medium matures and the broader register landscape settles, these innovations often arise not only from new platforms but as more subtle variations of already existing forms. One of the most productive examples of this phenomenon is Reddit. It principally differs from other social media in that it seems to foreground not individual users but the specialised communities they form and the content they produce. The website realises this concept through its organisation into subreddits, which show a complex interplay of explicit self-imposed community-specific rules and implicit communicative norms.

In light of these rules governing each subreddit, enforced by self-appointed moderators who can ban users and remove contributions deemed inappropriate, this study argues that Reddit's communities represent unique contextual variants of the website at large that may engender characteristic linguistic variation. It therefore explores whether subreddits, as user-curated categorisations of web content, are linguistically meaningful, i.e., sufficiently contextually and therefore functionally different that they constitute subregisters of Reddit, as identifiable by systematic clusters in the distribution of lexicogrammatical features like pronoun use or

mood choice. This study therefore treads new ground by investigating platform-internal register variation rather than comparing overt differences across conventionally recognised groups of web documents. This perspective can not only improve the current understanding of linguistic differences at lower levels of instantiation but also further ongoing efforts of "anatomizing" (Kilgarriff and Grefenstette 2003: 345) the web since Reddit demonstrates the benefits of functional categorisation at a smaller scale, allowing users to find communities and types of content matching their interests. For information retrieval purposes, then, it may become possible to filter web searches not only semantically by content, but according to exceedingly specific functional purposes and contexts.

The next section will first provide the necessary background information on how Reddit works as a website to justify the assumption that it inherently promotes the emergence of new and diverse situational contexts. This then leads into the motivation for using a multidimensional approach grounded in systemic functional register theory against the background of previous research on online registers. Afterwards, the preprocessing steps and method at the heart of this investigation, namely Geometric Multivariate Analysis (GMA), will be described. This includes the operationalisation of text as threads used in this study as well as the selection criteria for subreddits and lexicogrammatical features. The results section then analyses a sample of ten (out of thirty-three) subreddits and relates their contextual descriptions to textual patterns in a two-dimensional scatterplot visualisation of the linguistic feature space provided by the GMA. This is accompanied by examples from concrete texts illustrating salient register characteristics based on their clustering in this plot. Lastly, the discussion reflects on these results with respect to implications for the status of subreddits as (sub)registers proper and calls attention to the impact of content moderation (especially by bots) and Reddit's social systems (especially voting) as promising avenues for future register research.

2 Background

Reddit is a social news aggregation website where registered users, or redditors, can submit and rate content posted in discussion forums they create. Unlike most other social media, which typically foreground the social networks of its users, Reddit is therefore, by design, not a monolithic platform, but consists of millions of smaller communities specialised for certain topics and purposes. Moreover, Reddit's content policy – platform-wide guidelines establishing a basic code of conduct for everyone – is actually situated "[b]elow the rules governing each community"

(Reddit Inc. 2022). These rules, displayed openly on the respective subreddit's sidebar, are enforced by moderators appointed from the community by the existing team of moderators (and initially its creator) who have the power to ban users, remove contributions in case of rule violations, and close threads if they no longer deem them conducive to the ongoing discourse. In other words, the "culture of each community is shaped explicitly, by the community rules …, and implicitly, by the upvotes, downvotes, and discussions of its community members" (Reddit Inc. 2022), which naturally filter out contributions that do not follow the subreddit's conventions by reducing their visibility.

What also factors into Reddit's system of self-governance is the so-called Reddiquette. This portmanteau of Reddit and etiquette describes "an informal expression of the values of many redditors, as written by redditors themselves" (Reddit Help 2021), which is endorsed by the administrators, employees of the company Reddit, Inc. who maintain the website and monitor its content. In doing so, they lay the groundwork for respectful interactions, for example by moving against harassment and spam, but generally only involve themselves with specific communities if their rules violate these basic terms. Besides reiterating general communication guidelines, the Reddiquette also recommends more concrete behaviours with implications for the nature of Reddit's content. Most importantly, the users seem to strive for high integrity and thus set an unusually high linguistic standard for themselves. Among other things, they encourage using "proper" grammar and spelling (even encouraging corrections, though no specific standard or guidelines are given) as well as proof-reading submissions, remaining factual, referencing original sources, avoiding redundancy, and providing constructive criticism where appropriate.

In practice, however, subreddits seem to differ in terms of the extent to which they adhere to these ideals. While communities with stricter rule enforcement do exist (e.g., r/AskHistorians, which requires all answers be comprehensive and well-sourced), most subreddits, especially those in the spirit of more casual discussion forums, simply do not provide a reasonable context for such demands. As such, Reddit still offers ample opportunity for variation, not least because its community-driven design encourages users to create subreddits with their own rules if the available options do not agree with them. This freedom will hence engender distinct communicative contexts that cater to the specific needs of groups of people that have not yet found "their community" elsewhere. In combination with Reddit's voting system and the claim to "[m]oderate based on quality, not opinion" (Reddit Help 2021), then, it seems likely that moderation rather reinforces what these communities consider (contextually, i.e., for the respective community) appropriate language use instead of determining standards a priori. The existence of a karma system (named after the religious concept), i.e., points earned by receiving upvotes and paid awards, further amplifies this effect.

Against this background, it should become apparent that Reddit encourages the continuous creation of situationally and hence functionally distinct contexts of language production. In view of this dynamicity, Titak and Roberson (2013: 236) argue convincingly that internet linguistics is in desperate need of a theoretical model that "provides focus for linguistic research across web texts" such that rigorous comparisons of the interplay between sociocultural factors and their linguistic consequences in CMC contexts become possible. One concept that has been repeatedly employed in hopes of fulfilling this purpose is register, a cover term for any "variety associated with a particular situation of use" (Biber and Conrad 2009: 6). Previous research on internet registers has so far largely followed the multidimensional approach (MDA) by Biber (1988), identifying significant linguistic overlaps with offline counterparts (see e.g., Titak and Roberson 2013; Berber Sardinha 2014). Notably, the most prominent dimension of variation online also seems to be the distinction between involved vs. informational production, which contrasts pronoun-heavy personal involvement with content-focused nominal features. However, these studies tend to keep the crucial step from variable contexts to such systematic differences in language use rather vague, selecting features indiscriminately with little theoretical motivation for the registers at hand.

In this context, Biber and Egbert (2018: 26) further criticise that many corpus-based studies of register variation identify registers of interest based on perceptual salience, evaluating texts against the face validity of labels assigned to them. However, their framework, which relies on human coders to categorise texts based on predefined situational characteristics is not decidedly different from relying on perceptual salience as it likewise artificially restricts the types of text deemed linguistically meaningful. Indeed, the fact that raters often disagreed at the lowest granularity due to the seemingly inherent hybridity of web texts, which often combine features from different registers, advocates for an approach that focuses solely on characterising documents by their contextual parameters and connecting them to concrete linguistic correlates (i.e., textual features). Such an approach should form the basis of analysis because "linguistic differences among registers can be derived from situational differences", but "patterns of behaviour cannot be derived from any linguistic phenomena" (Biber and Conrad 2009: 9). As such, though these studies may provide convincing evidence for the existence of variation, they fall short in terms of systematically describing this crucial underlying relationship, which hinders comparable generalisations, especially online.

The present work argues that this gap can be best addressed by conceptualising online registers the same way as offline ones by connecting both perspectives through the concept of instantiation in systemic functional theory (Halliday 1978; see Halliday and Matthiessen 2014 for an introduction). Instantiation describes the gradual process of realisation whereby language users collapse the overall mean-

ing potential of the language system (i.e., all available linguistic choices) into concrete instances (i.e., texts with specific features) according to the particular "scenario … from which the things which are said derive their meaning" (Halliday 1978: 28). Naturally, this context of situation frequently recurs probabilistically in particular configurations that can be defined in terms of the continuous variables field, tenor and mode of discourse, which broadly correspond to topic/purpose, participant relationships like social distance, and aspects of text construction such as medium or preparedness. Together, these three dimensions enable consistent groupings of web documents because they resonate with the basic metafunctions of language and therefore have immediate functional correspondences (Halliday and Matthiessen 2014: 34). For example, on a subreddit for providing expert answers to historical questions like r/AskHistorians, texts will likely have a high lexical density to provide information and use verb phrases in past tense.

On the so-called cline of instantiation, registers occupy the mid region between the two outer system and instance poles because they exist at "varying degrees of specificity" (Halliday 1978: 111) so that the differences between them can be interpreted as a multidimensional "continuum of variation" (Biber and Conrad 2009: 33). As such, linguistic variation according to context of use should be observed both top-down, i.e., from above at the system pole, as registers defined by situational aspects, and bottom-up, i.e., from below at the instance pole, as text types defined linguistically, since these perspectives provide complementary information, like Biber and Egbert (2018: 213) also conclude. By the same token, register analyses typically focus on different ends of this continuum to foreground one area within the "range from a macro-register to the micro-registers that it consists of" (Matthiessen 2019: 20). Transferred to the present work, Reddit can be theorised as one example of a macro register, comprised of more specific instantiations in the form of specialised communities; after all, the notion of a hierarchy is already implied in its organisation into subreddits. Assuming such a structured perspective on categorisation largely obviates the need for subjective judgements and therefore ensures meaningful comparisons even as the register landscape changes.

For Reddit, Liimatta (2019) found evidence of systematic groupings by community along linguistic dimensions comparable to other internet registers (despite a noticeable personal bias in the corpus design), yet the comparatively low variance explained nicely demonstrates that comparing average frequency scores hides more nuanced differences between these presumably more specific texts (see Matthiessen 2019: 20). To combat this shortcoming, Diwersy et al. (2014) developed GMA, a pipeline for visualising linguistic differences between individual texts in multivariate register space. To do so, GMA uses Principal Component Analysis

(PCA) to identify latent dimensions of variation in the data that combine as much of the features' shared variance as possible. The resulting smaller subspace is chosen such that distances between data points remain meaningful with respect to the linguistic (dis)similarities of the original feature vectors, thereby revealing more delicate distribution patterns than aggregated group centroids or broad feature correlations patterns could (Neumann and Evert 2021: 146). Compared to Biber's (1988) MDA, GMA encourages exploring visualisations based on theoretical considerations, which is particularly helpful in online contexts where the significant functional and linguistic variables may not always be intuitively obvious (see Biber and Egbert 2018).

3 Method

The corpus for this study was compiled as a subset of the ConvoKit (Chang et al. 2020) datasets based on the r/ListOfSubreddits (2018) wiki, which contains a user-sourced list of communities grouped by categories such as discussion, education, or entertainment. Of course, one community is not representative of the entire user-base of Reddit; still, this list naturally emerged as a community effort and was not elicited according to a predefined schema (cf. Biber and Egbert 2018). The categories can therefore be considered authentic, albeit removed from linguistic theory, which is, in fact, desirable for the purposes of this study because it becomes possible to test whether these groupings are not only socially but also linguistically "real" from a register standpoint. To enable comparability, the subreddits were functionally characterised in terms of the field, tenor, and mode parameters above, and selected to cover a wide range of contextual variation. Where multiple options were feasible, the largest one deemed most representative of its category took precedence, assuming a lower specificity of features that is more amenable to this first exploration of linguistic distinctiveness (see Matthiessen 2019: 30). For each one, only the first 5000 threads before May 4th, 2018, were analysed because they had been archived and could therefore be considered complete texts that are fully realised linguistically. To reduce noise in the visualisation, this paper reports on only ten of the thirty-three selected subreddits as illustrative examples (see Table 1).

Table 1: Summary of selected subreddits with descriptions.

Subreddit	Description
AskHistorians	answers to questions about history
DIY	talking about homemade projects
GifRecipes	recipes in short video format
history	general discussions about history
recipes	sharing different kinds of recipes
science	discussing new scientific research
talesfromtechsupport	stories about working in tech support
techsupport	troubleshooting technical issues
UnsentLetters	sharing unsent personal letters
WritingPrompts	prompts for creative writing

Since GMA regards each text individually, sampling issues do not run as high a risk of under-representing registers with more internal variation, like Berber Sardinha (2014: 86) cautions for MDA. At the same time, this means defining what exactly constitutes one text is a crucial theoretical consideration. On Reddit, posts are best viewed as initial turns in a conversation continued by recursive comments from other users, leading to a hierarchical tree-like structure. The present study considers each of these emergent threads as one text for two main reasons. Firstly, the immediate context of situation pertains to the entire thread, so regarding comments separately, like Titak and Roberson (2013: 242) do for blogs, seems arbitrary here since they are not merely about a text but actively co-create it and must therefore be considered a component part that usually cannot stand alone. In a similar vein, the producer-user distinction proposed by Berber Sardinha (2014: 83) appears unfounded, considering that any user actively participating in a thread, by definition, simultaneously produces and consumes it. Additionally, there's the more practical consideration of statistical validity, which demands a certain minimum text length to achieve meaningful quantitative results. For GMA, this threshold lies around 100 words, or 10 sentences (Neumann and Evert 2021: 149), which even threads often fail to reach, so using shorter individual contributions would be unfeasible. Ultimately, this approach resulted in a sample of 74,960 texts.

All texts were normalised in terms of formatting and tokens tagged for their part of speech using the CLAWS C7 tagset (Garside and Smith 1997) whose granularity allows querying more complex lexicogrammatical features. Though not specifi-

cally trained on "dirty" web data (Kilgarriff and Grefenstette 2003: 342), a cursory inspection of the results showed no systematic errors that would disproportionately affect certain communities in the statistical analysis, not least because Reddit appears unusually concerned with correctness, as previously discussed. The selected subcorpora were transformed into a verticalised format (one token per line) and indexed for automatic feature extraction with the CWB platform (Evert and Hardie 2011). The query script by Neumann and Evert (2021) was used as the starting point for linguistic operationalisation of their contextual differences. As intended for GMA, the feature catalogue was adapted to count usernames as proper nouns. Additionally, titles were disregarded in favour of contractions and hyperlinks as characteristic features online. Three other features measuring emojis, edits, and forms of address, intended to replace salutations, ended up being too sparse to include. Due to high correlations, which may exaggerate effect sizes by measuring the same underlying structures, aggregate adjective counts were also removed. The input table therefore consisted of 42 features (see Table 2), all normalised as relative frequencies with respect to sensible units of measurement (see Neumann and Evert 2021: 150).

PCA relies on correlations between these features to project them onto new axes, which are chosen such that their combined variance is maximised. Compared to the rather opaque semantic relationships modelled by embeddings, its deterministic visualisation enables systematic interpretations grounded in register theory at the cost of being sensitive to scaling differences. The raw feature scores showed extreme variation and outliers, so log-transformed z-scores are used to deskew the distributions (see Neumann and Evert 2021: 151). Since higher-dimensional visualisations become increasingly harder to grasp and each Principal Component (PC) explains significantly less variance, only the first four components were analysed. Together, they already account for 42.9% of the original data, comparable to Biber and Egbert (2018) and a significant improvement over 17%, achieved by Liimatta (2019) using MDA. Here, only the first two, still accounting for over 30% variance, are described. Due to its overly optimistic group-awareness, a complementary Linear Discriminant Analysis (LDA), which can be used to reveal more subtle variation (Neumann and Evert 2021: 46), hides pronounced differences that emerge quite clearly in the PCA, so this step was omitted for the purposes of this study. All calculations and visualisations were performed in the statistical programming language R (R Core Team 2021) using the GMA utilities provided by Neumann and Evert (2021).[1]

1 The compilation and analysis scripts for the full corpus are available at https://osf.io/a7m9d/ (last accessed 14 February 2025).

Table 2: Summary of selected features with descriptions.

Feature	Description
adv_initial_S	sentence-initial adverbs per sentence
atadj_W	attributive adjectives per word
contr_W	contractions per word
coordination_F	coordinating conjunctions per finite
disc_initial_S	sentence-initial discourse markers per word
finite_S	finites per sentence
imperative_S	imperatives per sentence
infinitive_F	to-infinitives per finite
interrogative_S	interrogatives per sentence
it_P	it-pronouns per pronoun
lexical_density	lexical density (proportion of content words)
modal_verb_V	modal verbs per verb
neoclass_W	neoclassical compounds per word
nn_W	common nouns per word
nom_initial_S	sentence-initial nominal elements per sentence
nominal_W	nominalisations per word
nonfin_initial_S	sentence-initial infinitive clauses per sentence
np_W	proper nouns per word
p1_perspron_P	1st person personal pronouns per pronoun
p2_perspron_P	2nd person personal pronoun per pronoun
p3_perspron_P	3rd person personal pronouns per pronoun
passive_F	passives per finite
past_tense_F	past tense verbs per finite
place_adv_W	adverbs of place per word
pospers1_W	1st person pronouns per word
pospers2_W	2nd person pronouns per word
pospers3_W	3rd person pronouns per word
poss_pronoun_W	possessive pronouns per word

Feature	Description
predadj_W	predicative adjectives per word
prep_initial_S	sentence-initial prepositional phrases per sentence
prep_W	prepositions per word
pronoun_all_W	all pronouns per word
subord_initial_S	sentence-initial subordinate clauses per sentence
subordination_F	subordinating conjunctions per finite
text_initial_S	sentence-initial discourse markers per sentence
time_adv_W	adverbs of time per word
url_W	hyperlinks per word
verb_initial_S	sentence-initial verbal elements per sentence
verb_W	verbs per word
wh_initial_S	sentence-initial wh-elements per sentence
will_F	will futures per finite
word_S	words per sentence

4 Results

Figure 1 shows a grouped scatter plot of the first two PCs for the ten exemplary subreddits analysed in this study, with PC1 on the y-axis and PC2 on the x-axis. The scatter plot is split into two faceting groups for better readability. All axes are scaled equally so as to be understood as different perspectives on the same underlying space comprised by the first four PCA dimensions. Within this space, each dot, colour-coded for subreddit, represents one text whose position is determined by its score for the respective PCs such that their potential clustering can be analysed based on a dumbbell plot of the loadings for PC1 and PC2 (Figure 2). These loadings indicate the relative prominence of each linguistic indicator after reducing the dimensions of the original data vectors by capturing their combined variance as linear combinations of the input features. In other words, the score (and thus the position) of each text on a given PC depends on how strongly its most frequent features are represented by that dimension. The quantitative focus of the results is enriched with selected qualitative analyses to help ground abstract feature frequencies in their functional expression within concrete texts. To protect

the pseudonymity of users, examples reference the unique ID of the post they belong to, which enables replicability but hopefully hinders immediate user identification.

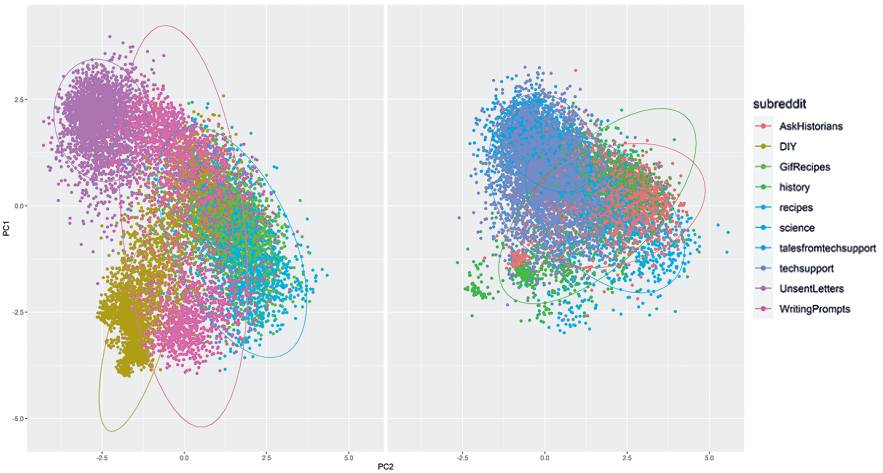

Figure 1: Scatter plot of text scores for PC1 and PC2.

Regarding the loadings of PC1, imperatives, hyperlinks, common nouns, and second person personal pronouns have the strongest negative contributions, being slightly below -0.2. Both imperatives and pronouns point towards an instructional goal orientation and interpersonal albeit authoritative involvement with an addressee for the tenor of discourse. This is also supported by the significant contributions of discourse markers in theme position and, to a lesser extent, initial verbal elements. At the same time, common nouns and, relatedly, lexical density indicate a rather informational character in terms of purpose. That hyperlinks load as strongly as imperatives suggests that language takes on an ancillary role in texts on this part of the first dimension, accompanying content outside of the thread. The positive side of PC1 loadings, on the other hand, is dominated by features typical of a spoken medium with narrative purpose, as evidenced by five out of the seven strongest indicators relating to the use of first- and third-person personal pronouns, which presumably also explains the weights above +0.2 for initial nominal elements. The loading just under +0.2 for past tense in combination with contractions indicate rather informal narration, presumably in the form of personal stories.

For the second PC, the strongest indicators on the negative side resemble the positive side of PC1 in that they largely pertain to pronoun use; however, instead of

the third person, second-person pronouns have the highest value, even surpassing said indicator with a value of under -0.3. Since five out of six strong features are again pronoun-related, this is closely followed by a general indicator for all pronouns as well as possessive pronouns slightly below -0.3. Rather than narration, this dimension therefore seems to capture interpersonal involvement more generally, positing conversation as a communicative purpose in itself. The apparent correlation with a high frequency of verbs, which are typical of spoken discourse, supports this notion. On the opposite end, attributive adjectives, which are generally associated with more informational writing, have the highest loading score. Appropriate indicators of nominal style also make considerable contributions to the positive side of PC2, especially lexical density coming in fourth but also common and proper nouns with loadings around +0.2. This is complemented by third person personal pronouns and the pronoun *it*, both above +0.2, which hint at an expository goal orientation. The rest of this section will now relate patterns in the distribution of texts from different subreddits to these feature combinations.

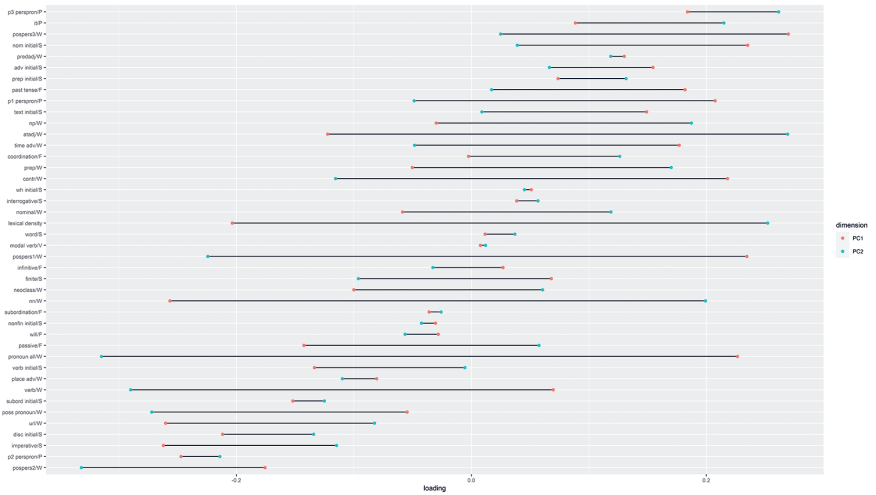

Figure 2: Dumbbell plot of feature loadings for PC1 and PC2.

r/GifRecipes is a media-sharing community, so its texts would be prototypically expected to contain features of ancillary language use, such as hyperlinks and imperatives. Indeed, they strongly favour negative scores on PC1, which is associated with these indicators. Looking at concrete examples reveals that this results from posters including written versions of the recipes shown, which link to the source video and use imperatives to provide step-by-step instructions like “Pat dry

the duck breasts with a paper towel." (7jwqig). The list of ingredients, then, also explains the apparent prominence of common nouns, engendering a high lexical density that moves the texts towards positive PC2 scores. Consequently, they often show strong similarities to their printed counterparts in terms of form and content, as seen in Biber and Egbert (2018: 138). Perhaps unsurprisingly, the functionally similar but less specific r/recipes also tends towards negative scores on PC1 and the positive side of PC2 but shows more variation overall, suggesting that one encompasses the other. Outliers are readily explained by posts that only link to a recipe (see e.g., 7s1xcv), or questions (see e.g., 7sir6i). In both cases, personal deixis from the comment section will start to dominate, pushing texts towards the lower right quadrant of higher interpersonal involvement.

In line with formal expectations, r/DIY should likewise be characterised by imperatives and, accordingly, verbs in sentence-initial position since the subreddit requires submissions to include detailed instructions. However, given that the positive indicators for PC1, where these texts primarily cluster, strongly weigh personal pronouns, this subreddit does not seem to be prototypically instructional but rather narrative. This is because, if present at all, specific instructions are usually only given as links to image albums (see e.g., 8cr74f). Unlike the skills and hobbies category from the International Corpus of English (ICE), then, located on the side of conceptual writing in Neumann and Evert (2021), pronouns commonly occur as theme here because users seem to frame their projects as personal stories rather than formal manuals intended for replication, as this example illustrates:

> After I had my plan all drawn up and the rough dimensions in my pocket, I went out to Home Depot to get the wood products I needed. I already had the addressable LED strips from a previous project that I might post later, and I had the electronics from kits. With all the supplies it was time to get working. (8ebeqx)

That contractions also contribute positively to the first dimension supports this notion. Help requests, the other type of permissible content on r/DIY, contain first and second person pronouns, too, due to being more advisory rather than instructional, again indicating that users employ a more involved style (cf. Biber and Egbert 2018: 57).

In contrast, r/WritingPrompts generally favours pronoun usage across PCs where they attest a narrative goal orientation, which is consistent with Neumann and Evert's (2021: 11) findings for the creative writing category in the ICE (cf. Biber and Egbert 2018: 102). For PC2, there are texts that demonstrate indicators more in line with the literate-nominal dimension from Biber and Egbert (2018) as well, however. In the following excerpt, a prompt about Canada dropping an atomic bomb in the First World War was addressed with a historical speech, for instance: "1917, the

year everything changed. On April 6 the US declared war on the Allied Powers after an American ship was accidentally sunk by British torpedoes and a French message to Mexico was intercepted by the Germans." (8eg6zf). It seems, therefore, that this variation is attributable to differences in the register targeted by the prompt and its realisation, which may be predictable from the position in the feature space. Still, neither of these functional characterisations explain the joint clustering of r/DIY and r/WritingPrompts at the negative end of the first PC. Taking a closer look at the text with the lowest score on PC1, located at -4, reveals the following comment:

> Your submission has been removed for one or more of the following reason(s): Your question might be answered with a few minutes of basic research of this subject. … Please read our guidelines before resubmitting. If you believe this was a mistake, please message the moderators. (r/DIY: 8gmgnw)

This response was sent by a moderator of r/DIY because they deemed the poster failed to do their own research before asking a question as is required by the subreddit's guidelines. Since the original submission was removed due to this rule violation, the thread consists solely of the above cited comment. This, then, explains the feature combination that seems to predominantly characterise the negative side of PC1. Across the text, there are several exophoric references in the form of hyperlinks to helpful resources, using imperatives with the discourse marker *please* as theme to point out aspects of their submission (hence also the comparatively high frequency of second person possessive pronouns) and instruct them how to avoid having their posts deleted in the future. Additionally, moderation messages are presumably not produced spontaneously but prepared beforehand and constantly refined, striving for conciseness and intelligibility, which would explain the somewhat nominal style evident in the feature loadings. While these messages could have been removed to bring out the characteristics of user contributions more, the apparent linguistic impact and curating function warrant their inclusion as a prominent register feature, at least for an explorative study such as this.

To provide another example, a nearby text from r/WritingPrompts (8efxuk) has no responses because the user deleted their post and only contains an automated message by a bot that reads: "Off-Topic Discussion: All top-level comments must be a story or poem. Reply here for other comments." Once again, the comment contains imperatives and hyperlinks, but because it is a more general message not directed to a specific user, it lacks second person personal pronouns, so the text is more centred on the second PC. For the sake of thread organisation, every post on r/WritingPrompts automatically receives such a notification. Given this central function, they must likewise be considered part of this community's register despite (or rather because of) their impact on the feature distribution. As texts move

towards the positive end of PC1, these kinds of messages become less prominent linguistically. The difference between the lower and upper cluster of texts, then, lies in the prevalence of responses to the prompt. In that case, the distinctive features of moderation will be gradually overshadowed by narrative indicators, which are characteristic of the subreddit. In other words, the higher the proportion of text produced by actual users, the further the data points are pushed along PC1. The writing prompt with the highest positive score (8h2wyr), for example, received only one story, but due to its length, the bot comment becomes negligible in comparison.

Looking at other selected texts in the bottom cluster of the first PC reveals that this phenomenon roughly occurs below scores of -2 and remains consistent across subreddits, so PC1 seems to separate user comments from moderation messages quite well. The respective subgroupings can be traced back to different kinds of rule violations and comparatively minor variation in the posts' titles, which suggests that the underlying causes for moderation action could be reliably derived from linguistic indicators alone. The prominent group of r/DIY texts around -3, for instance, predominantly seem to have been moderated because they consisted of only a single image (see e.g., 8ajadx). This, then, also explains why only a few hundred texts from r/DIY were too short to include in the analysis compared to over half for most others. Instead of potentially indicating the type of content found in a community, or even specific rules (providing detailed instructions for a project presumably requires a certain number of words, after all), text length may therefore hint at how actively the community is moderated. In any case, the presence of such messages adds another layer to the already somewhat opaque social relationships online as interactions need not occur exclusively between humans.

Moving on from r/WritingPrompts to another narrative subreddit, r/talesfromtechsupport expresses the same goal orientation predominantly via first and third person pronouns. The texts also show similarities to phone calls in this respect, as investigated by Neumann and Evert (2021: 10), because they tend to originate in help desk situations and therefore heavily feature quotes of participants. The subreddit seems to be a lot more homogeneous as a result and therefore appears quite concentrated around moderately positive scores on PC1 in the visualisation, similar to creative writing. The following excerpt is a typical example illustrating this overlap:

> I get an email from a user … to say that they can't send email via our SMTP server. … I look at the logs and I can see he's having problems authenticating so my guess is that … he's fiddled with his settings. I phone him up and ask. "No", he says "I've not changed my settings at all". (76068s)

Interestingly, no moderation subgroup pattern emerges for this subreddit despite its rather strict rules because inappropriate posts are filtered, viz. removed, before submission. r/techsupport provides an interesting contrast to r/talesfromtechsupport because it is essentially a more general version of this community without the narrative focus; that its cluster in Figure 1 appears overall less focused, comparable to the difference between r/GifRecipes and r/recipes, aligns with this observation. Even so, for a subreddit about asking for help, the contribution of interrogatives to their position are surprisingly negligible. Looking more closely at individual texts reveals that this is because users often formulate their concerns as statements, such as "Computer will not shutdown" (8ff8qw).

Regarding the more explicit question-answering subreddits, r/AskHistorians is marked for the same nominal features (positive PC2) as r/science (see e.g., 88c69h) and r/history (see e.g., 8852wy, also a question), especially attributive adjectives, lexical density, and common nouns. This, then, means that the Q&A aspect of these subreddits is not dominantly reflected in frequent use of interrogatives either since questions would typically only occur in the post itself. On r/AskHistorians, in particular, comments are expected to be quite detailed and supported by credible sources, leading to a content focus evidenced by its general position in the scatter plot. For instance, one user's explanation for why Italy did not join the Central Powers during WWI starts as follows:

> With the exception of a few people in the Italian High Command no one considered the possibility of Italy joining the side of the Central Empires as a serious eventuality. And that included the Germans and Austrians who had clear understanding of the situation of the Triple Alliance. (8agen9)

In the upper groupings, the cline of experiential specificity from science to history seems to emerge as greater variation in the more general r/science. This is particularly evident on PC2 where outliers are attributable to copies of abstracts (see e.g., 89cpuh) or paraphrases of journal articles (see e.g., 89dxqi), explaining their position on the nominal as opposed to the more characteristically spoken negative end of the dimension. Conversely, the cluster of r/history is overall more focused around moderately positive values for PC2 because all posts are first reviewed by human moderators. In contrast to other moderation messages, the outliers for r/history and r/AskHistorians are therefore considerably less marked for PC1 and rather favour negative PC2 scores because they try to achieve a more personable, conversational tone, as the following examples both illustrate:

> Your submission has been automatically removed because it triggered some filters since you are fairly new. This is nothing to worry about, if your post follows the r/history rules we can approve it for you once you message us. (r/history: 8g1ptc)

> Hi there - unfortunately we have had to remove your question, because /r/AskHistorians isn't here to do your homework for you. (r/AskHistorians: 8ga21p)

Lastly, the userbase of r/ListOfSubreddits (2018) categorises r/UnsentLetters as a support community, but the subreddit's rules expressly forbid unsolicited opinions or advice. Accordingly, its texts lack the typical indicators of problem-solving present in other advice documents, being characterised by first and third rather than second person pronouns (cf. Biber and Egbert 2018: 128). Outliers on the negative end of PC1 and the positive end of PC2 seem to be primarily letters in other languages (see e.g., 8b1vmy). The premise of personal letters that users were too afraid to send explains this linguistic overlap with social letters in Neumann and Evert (2021: 151). At the same time, users frequently narratively reflect on past events in an informal manner, leading to even stronger PC1 loadings due to contractions, verbs in the past tense, and time adverbs. For example, in the letter with the highest positive score, the poster laments, "I wish I didn't love him anymore. I wish I didn't care about him anymore. I wish I didn't need him" (8h2hxj). Presumably due to the aforementioned rule, comments seem to be rare on this subreddit, so the features of such posts become more pronounced (or rather less blurred), which explains why its texts have such a prominent position, even in the full feature space.

5 Discussion

The results reveal that subreddits systematically cluster in terms of their linguistic features, suggesting that they can indeed be considered subregisters of Reddit. Indeed, conceptually related communities generally cover similar areas, attesting to a continuous space of variation in this yet tentative macro register (see Neumann and Evert 2021: 152), perhaps brought out further by their hybrid functions. Specifically, it seems that the majority of subreddits display features of involvement alongside those specific to the respective subreddit, which is expected of a social media site for discussing specialised interests. The analysis has also demonstrated striking overlaps with offline registers, which valorises these communities as registers proper. Based on those salient similarities, one could argue, as Biber and Egbert (2018: 42) do for blogs, that Reddit represents a kind of microcosm of the web, viz. the web at large is reflected on a smaller scale within its communities. In a way, subreddits are dense accumulations of web content that also exists elsewhere, which attract interested users with easily understandable and searchable labels that would otherwise be hard to find. By demonstrating that they can be differentiated linguistically via computational means, these findings pave the way toward

automated functional web categorisation, for example for the purpose of informational retrieval.

A significant variable that becomes quite salient on the internet, and particularly public discussion forums, but has so far been ignored in register variation research is moderation. It shapes the context of online communication not only socially but also linguistically since moderators represent the de facto authority over the kind of language permissible in a given community. This has become abundantly clear by the separation of multiple subreddits into moderated and unmoderated texts on the first, most significant PCA dimension. A comparative investigation into the extent to which moderation solely occurs based on violations of conventionalised social norms and formal properties of contributions or if such measures also have a linguistic basis could prove valuable. The fact that certain subreddits evaluate submissions based on goal orientations with well-defined linguistic indicators (e.g., whether they entail a narrative element) certainly suggests so. This is especially relevant considering that many subreddits off-load moderation work to bots, a trend that has become increasingly relevant on the internet in recent years. In general, the issue of bots has likewise not yet received due attention in the field of internet linguistics despite important implications for the representativeness of register corpora and opportunities for variation studies.

Another relevant aspect of Reddit not touched upon in this study was the potential impact that votes and awards (purchased with real money) may have on a communities' language use because they can be used to affect the visibility of contributions. Depending on the chosen sorting strategy, posts with a high score have a chance to land on the website's front page and the respective subreddit's feed, whereas popular comments appear earlier in the thread. Notably, unlike other social media, users are discouraged from downvoting posts based on opinion; rather than emotional reactions, votes are thus supposed to reflect whether a post is "contributing to the community dialogue" (Reddit Help 2021). A post's score, then, is, at least in theory, not as much a stamp of approval as it is a sign of quality. In other words, from a linguistic perspective, one could argue that these ratings may not only indicate a more general value-judgement but also how well a given post fits the userbase's implicit notion of that subreddit's register. Future studies may therefore want to explore the extent to which selecting threads with negative scores would affect the positioning of texts in the visualisation and how this is reflected in terms of changes in the most prominent linguistic features.

A detailed investigation of lexicogrammatical differences for selected subreddits is required to gain more systematic insights into the patterns of linguistic features engendered by community-specific rules revealed in this explorative study. Introducing additional information into the analysis via a weakly supervised LDA could prove fruitful in this context. In particular, LDA may be used to reduce the

impact that moderation messages have on the feature space by foregrounding community-specific variation without removing them entirely. Moreover, choosing the thread as the unit of analysis under the assumption that each of them constitutes a single, homogeneous conversation and, by extension, one text, has had significant implications not only in terms of methodological possibilities but potentially also the results overall. Due to the tree-like structure of threads, it seems that contextual parameters often operate at lower levels of instantiation, either in local branches or perhaps even at the level of individual contributions. This was reflected in the fact that the tenor-related effects of user interactions could not be properly accounted for. Any future investigation of the sociolinguistic dynamics on the internet in systemic functional terms therefore presupposes an extensive adaptation of the framework and its operationalisations by considering the characteristic features of CMC. At the heart of this endeavour lies a follow-up study that uses GMA to investigate texts at some level below the thread.

References

Berber Sardinha, Tony. 2014. 25 years later: Comparing internet and pre-internet registers. In Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-dimensional analysis, 25 years on: A tribute to Douglas Biber*, 81–105. Amsterdam: Benjamins.

Berber Sardinha, Tony. 2018. Dimensions of variation across internet registers. *International Journal of Corpus Linguistics* 23 (2). 125–157.

Biber, Douglas. 1988. *Variation across speech and writing*. New York: Cambridge University Press.

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. New York: Cambridge University Press.

Biber, Douglas & Jesse Egbert. 2018. *Register variation online*. New York: Cambridge University Press.

Chang, Jonathan P., Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, & Cristian Danescu-Niculescu-Mizil. 2020. *ConvoKit:* A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 57–60.

Diwersy, Sascha, Stefan Evert & Stella Neumann. 2014. A weakly supervised multivariate approach to the study of language variation. In Benedikt Szmrecsanyi & Bernhard Wälchli (eds.), *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech*, 174–204. Berlin & Bosten: De Gruyter.

Evert, Stefan & Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In: *Proceedings of the Corpus Linguistics Conference 2011*, 1–21.

Garside, Roger & Nicholas Smith. 1997. A hybrid grammatical tagger: CLAWS4. In Roger Garside, Geoffrey Leech & Anthony McEnery (eds*.), Corpus annotation: Linguistic information from computer text corpora*, 102–121. London: Longman.

Halliday, Michael Alexander Kirkwood. 1978. *Language as social semiotic: The social interpretation of language and meaning*. London: Arnold.

Halliday, Michael Alexander Kirkwood & Christian Matthias Ingemar Martin Matthiessen. 2014. *Introduction to functional grammar*, 4th edition. New York: Routledge.

Kilgarriff, Adam. & Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics* 29 (3). 333–347.

Liimatta, Aatu. 2019. Exploring register variation on reddit: A multi-dimensional study of language use on a social media website. *Register Studies* 1 (2). 269–295.

Matthiessen, Christian Matthias Ingemar Martin. 2019. Register in systemic functional linguistics. *Register Studies* 1 (1). 10–41.

Neumann, Stella & Stefan Evert. 2021. A register variation perspective on varieties of English. In Elena Seoane & Douglas Biber (eds.), *Corpus based approaches to register variation*, 143–178. Amsterdam: Benjamins.

R Core Team. 2021. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Reddit Help. 2021. Reddiquette. https://www.reddithelp.com/hc/en-us/articles/205926439 (last accessed 15 May 2023).

Reddit Inc. 2022. Content policy. https://www.redditinc.com/policies/content-policy (last accessed 15 May 2023).

r/ListOfSubreddits. 2018. List of subreddits. https://www.reddit.com/r/ListOfSubreddits/wiki/listofsubreddits (last modified 14 October 2018).

Titak, Ashley & Audrey Roberson. 2013. Dimensions of web registers: An exploratory multi-dimensional comparison. *Corpora* 8 (2). 235–260.