Proceedings of the 10th International Conference on CMC and Social Media Corpora for the Humanities

14–15 September 2023, University of Mannheim, Germany

Editors: Louis Cotgrove, Laura Herzberg, Harald Lüngen, Ines Pisetta

Published by: Leibniz-Institut für Deutsche Sprache
Mannheim, 2023

DOI: *https://doi.org/10.14618/1z5k-pb25*
ISBN: 978-3-937241-95-1

Conference website: https://www.uni-mannheim.de/cmc-corpora2023

The CMC-Corpora 2023 conference is funded by *Deutsche Forschungsgemeinschaft* under the project number 524949653.

A Multivariate Register Perspective on Reddit: Exploring Lexicogrammatical Variation in Online Communities

Frenken, Florian
RWTH Aachen University
Kármánstr. 17/19, 52062 Aachen, Germany
florian.frenken@ifaar.rwth-aachen.de

Abstract

Even though social media have shaped day-to-day communication for years, the internal linguistic variation associated with these emergent register contexts still remains largely unknown. To fill this gap, the present study evaluates a geometric multivariate approach for this domain by investigating patterns in the visualisation of forty-two lexicogrammatical features derived from systemic functional theory for thirty-three communities on Reddit. The results successfully demonstrate that these subreddits can be interpreted as subregisters of a yet hypothetical macro-register that align with contextual and thus functional differences. Accordingly, this study argues that investigating individual texts rather than broad feature correlation patterns would improve multidimensional analyses of subregisters in general and hybrid web registers specifically. This perspective can not only improve our understanding of language variation at lower levels of instantiation but also hopes to incentivise further research on platform-internal register variation in light of practical implications for context-informed automatic classifications of web documents in functional terms.

Keywords: register variation, geometric multivariate analysis, reddit

1. Introduction

Over the years, computer-mediated communication (CMC) has continued to play a central role in everyday discourse (Crystal, 2001, 17). As users learn to navigate this new environment, they face continuously evolving communicative contexts to which they must adapt. Having no obvious offline counterpart, social media, in particular, represent the emergent registers of today whose labels are "instantly recognized" (Berber Sardinha, 2018, 126), yet surprisingly, their linguistic characteristics are not well understood despite the wealth of data available (Titak and Roberson, 2013, 235). A particularly salient perspective in this regard is the growing internal variation within these online platforms. One of the most productive examples of this is Reddit, an American social news website where users can submit, rate, and discuss content on various user-created boards called subreddits. With their hierarchical text structure, the resulting threads have clearly conversational character but the interactions are asynchronous and not necessarily linear (Crystal, 2001, 130–151).

In light of the specific rules governing each subreddit, enforced by self-appointed moderators who can ban users and remove contributions, this study argues that these communities represent unique contextual variants of the site. 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 their lexicogrammatical features. This perspective can not only improve our understanding of linguistic differences at lower levels of instantiation but also further yet ongoing efforts of "anatomizing" the web (Kilgarriff and Grefenstette, 2003, 345) since Reddit demonstrates the benefits of functional categorisation at a smaller scale, allowing users to find content and communities matching their particular interests.

In general, the notion of register refers to groups of texts showing systematic linguistic patterns that are functionally associated with specific situational contexts (Matthiessen, 2019). Previous research on internet registers has so far largely followed the multidimensional approach (MDA) by Biber (1988), consistently identifying dimensions of variation that demonstrate significant overlaps with Biber's original factors as well as offline registers (Titak and Roberson, 2013; Berber Sardinha, 2014). However, they tend to keep the crucial step from variable contexts to systematic differences in language use rather vague, often operating on face validity of labels, which says little about the actual text types, especially online (Kilgarriff and Grefenstette, 2003, 343), and thus hinders generalisation due to the web's fluidity (Crystal, 2001, 14).

Against this background, Biber and Egbert (2018) enlisted coders to categorise texts based on predefined, perceptually salient, situational characteristics with only minor success due to the inherent hybridity of web registers they encountered. As such, the present work argues that this gap can be best addressed by conceptualising internet registers from a systemic functional perspective (Halliday, 1978), which combines top-down and bottom-up approaches by deriving features for the linguistic analysis from the contextual parameters of a text, independent of the current register landscape. In doing so, this study theorises Reddit as an example of a macro-register according to Matthiessen (2019), comprised of more specific instantiations in a continuum of subregister variation; after all, this notion is already implied in its organisation into subreddits.

Indeed, Liimatta (2019) found evidence of systematic linguistic differences between communities on Reddit despite a noticeable personal bias in the corpus design. However, the low factor loadings as well as variance explained nicely demonstrate that comparing average frequency scores hides more nuanced differences between these presumably more specific texts (Matthiessen, 2019, 20). To combat this

shortcoming and provide an alternative to Biber's approach, Diwersy et al. (2014) developed the Geometric Multivariate Analysis (GMA) pipeline, a procedure for visualising differences between individual texts rather than broad feature correlation patterns. To interpret the higher-dimensional space of register variation, GMA uses Principal Component Analysis (PCA), which projects the data onto latent dimensions that capture as much of its original variance as possible. The distances in this subspace are meaningful with respect to the linguistic (dis)similarities of the initial feature vectors, revealing more delicate distribution patterns than aggregated group centroids (Neumann and Evert, 2021, 146).

2. Method

The corpus for this study was compiled as a subset of the ConvoKit (Chang et al., 2020) subreddit datasets based on categorisations in the r/ListOfSubreddits (2018) wiki. Of course, one community cannot realistically represent the entire userbase 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), so the categories can be considered authentic, albeit removed from linguistic theory. Due to the transience of web registers, field, tenor, and mode parameters were used to decide which communities from the general content category to include (Halliday, 1978, 62). Where multiple options were feasible, the most popular and basic one took precedence, assuming a lower specificity of its features (Matthiessen, 2019, 30). Only the first 5000 posts before May 4th, 2018, were analysed because they got archived and could therefore be considered finished. To reduce noise, this paper reports on only five of the thirty-three selected subreddits as an example (see Figure 1).[1]

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, especially for the web. The present study equates the notions of thread and text for three reasons. Firstly, the context of situation pertains to the entire thread, not only individual comments, so regarding them separately, like Titak and Roberson (2013, 242), seems arbitrary as they are not merely about a text (like for blogs), but actively co-create the thread and must therefore be considered part of it. Secondly, the producer-user distinction proposed by Berber Sardinha (2014, 83) appears unfounded, considering that every producer is, by definition, also a user (though not vice versa). Thirdly, the statistical measures of per-text frequencies for GMA have been shown to require a minimum of 100 words, or 10 sentences (Neumann and Evert, 2021, 149) – a threshold most threads fail to reach. Ultimately, following this approach resulted in a sample of 74,960 texts, with fewer in subreddits focusing on ancillary language use (e.g., r/DIY).

To extract even complex lexicogrammatical features, all texts were normalised in terms of formatting and tokens tagged for their part of speech using the CLAWS tagger and C7 tagset (Garside and Smith, 1997). Though not specifically trained on "dirty" web data (Kilgarriff and Grefenstette, 2003, 342), a cursory inspection showed no systematic errors that would disproportionately affect its accuracy, not least because Reddit seems unusually concerned with correctness compared to other social media (Crystal, 2001, 45). The ConvoKit corpora were indexed in verticalised format for automatic feature extraction with the CWB platform (Evert and Hardie, 2011) using a query script by Neumann and Evert (2021) whose linguistic operationalisations based on situational parameters enable generalisability. The feature catalogue was slightly adapted to count usernames as proper nouns to replace salutations. Due to high correlations, which may exaggerate effect sizes by measuring the same underlying structures, aggregate adjective counts were removed from consideration. Additionally, titles were disregarded in favour of contractions and URLs as characteristic features on the web. Three other additional features measuring emojis, edits and forms of address were too sparse to include. As a result, the final table consisted of 42 features (see Figure 2), all normalised as relative frequencies with respect to sensible units of measurement (Neumann and Evert, 2021, 150).

PCA relies on correlations to project these features onto new axes that capture their combined variance. Compared to the rather opaque semantic relationships modelled by embeddings (inaccessible here), 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 were used to deskew the distributions (Neumann and Evert, 2021, 151). Since higher-dimensional visualisations become increasingly harder to grasp and each PC explains significantly less variance, only the first four components were analysed. Together, they already explained 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 tentative exploration of a Linear Discriminant Analysis, which can reveal more subtle variation (Neumann and Evert, 2021, 46), hid pronounced differences that emerged quite clearly in the PCA, so the study omitted this step of the GMA procedure. All calculations and visualisations were performed in the statistical programming language R (R Core Team, 2021).

3. Results

Figure 1 shows a scatter plot of the first two PCs for the five exemplary subreddits, grouped pairwise with PC1 on the y-axis and PC2 on the x-axis, scaled equally so as to be understood as different perspectives on the same three-dimensional space. Within this space, each dot, colour-coded for subreddit, represents one text whose position is determined by its score for the respective PCs. Their potential clustering is analysed based on a dumbbell plot of the feature loadings for PC1 and PC2 (Figure 2), which indicate their relative prominence. The quantitative focus of

[1] The compilation and analysis scripts for the full corpus are available at https://osf.io/a7m9d/.

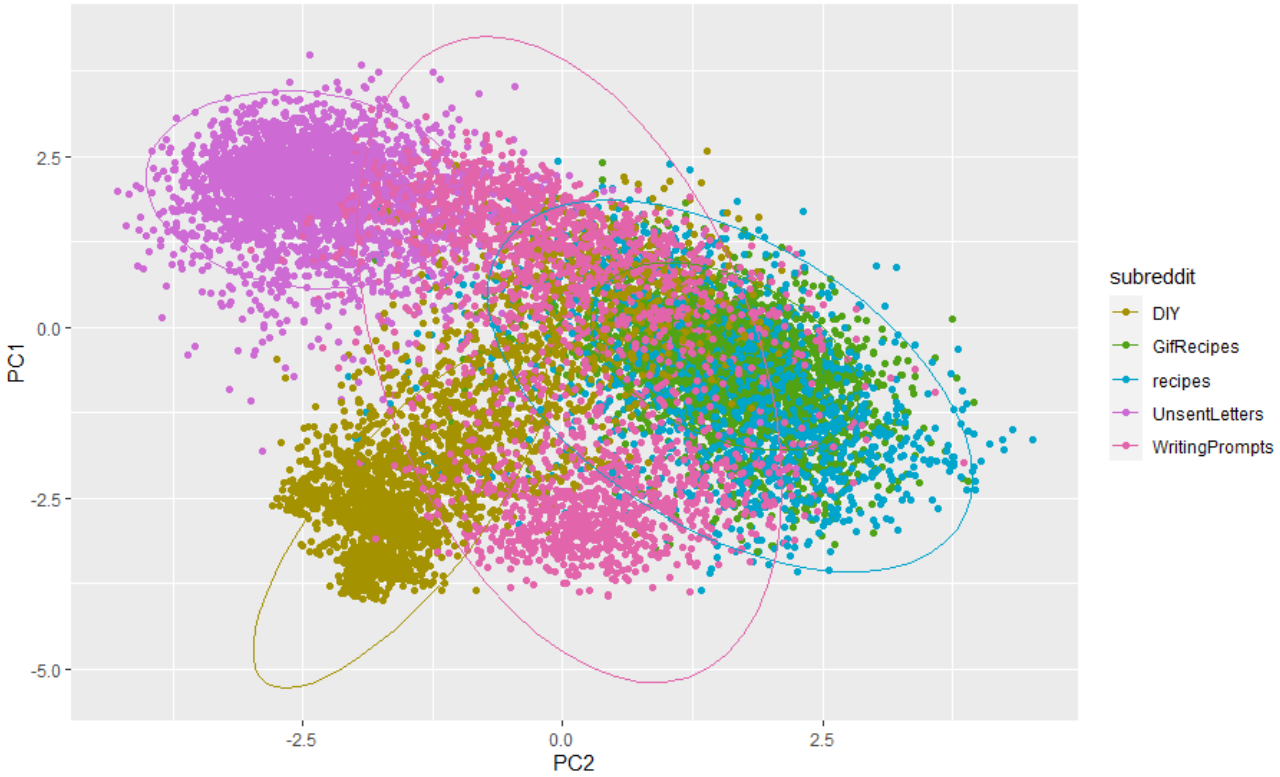

Figure 1: Scatter plot of text scores for PC1 and PC2

this study notwithstanding, the results are enriched with selected qualitative analyses to help ground abstract feature frequencies in their functional expression in concrete texts. To protect the pseudonymity of users, examples reference the unique ID of the post they belong to, which enables replicability but hopefully exacerbates user identification.

r/GifRecipes is a media-sharing community, so its texts should contain features of ancillary language use, such as URLs and imperatives. Indeed, they strongly favour negative scores on PC1, which are 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 (e.g., 7jwqig). The list of ingredients, then, also explains the prominence of common nouns, engendering a high lexical density that moves the texts towards positive PC2 scores. As such, they often show strong similarities to their printed counterparts in terms of form and content, as Biber and Egbert (2018, 138) also find. Perhaps unsurprisingly, the functionally similar r/recipes also tends towards negative scores on PC1 and the positive side of PC2. Outliers are readily explained by posts that only link to a recipe (e.g., 7s1xcv), or questions (e.g., 7sir6i). In both cases, personal deixis from the comment section will start to dominate, indicating a higher involvement of users.

One would likewise expect r/DIY to be characterised by imperatives since submissions should include detailed instructions. However, this subreddit does not seem to be prototypically instructional but rather narrative. Unlike skills and hobbies, located on the side of conceptual writing in Neumann and Evert (2021), where pronouns are infrequent, their (primarily possessive) forms frequently occur in theme position here because users are discussing their personal projects rather than writing a formal manual. 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 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 creative writing from the International Corpus of English (Neumann and Evert, 2021, 153). For PC2 especially, there are also texts that demonstrate indicators more in line with Biber and Egbert's (2018) literate-nominal dimension. It stands to reason that this variation is mainly attributable to differences in the field of discourse.

Still, this does not explain the prominent clusters of texts at the negative end of the first PC. Taking a closer look at concrete texts from r/WritingPrompts (e.g., 8efxuk), reveals that they contain moderation messages, often by bots, linking to helpful resources and using repeated imperatives with the thematic discourse marker please to instruct users how to avoid future rule violations. They are presumably constantly refined, striving for conciseness and intelligibility, which would explain their somewhat nominal style. As texts move along PC1 towards the positive end, these messages become less frequent, while the number of comments by actual users increases. The difference between the lower and upper cluster of texts, then, is the presence of contribu-

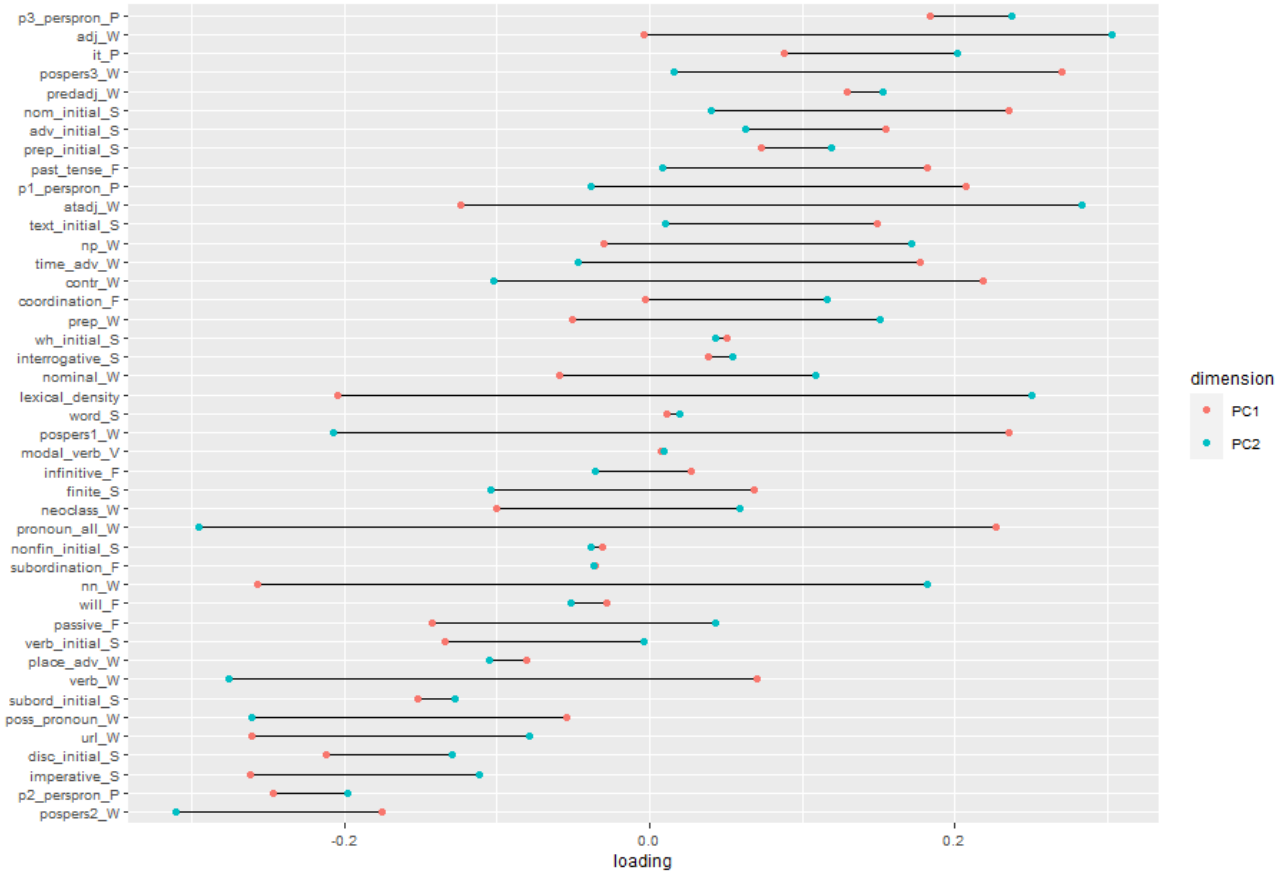

Figure 2: Dumbbell plot of feature loadings for PC1 and PC2

tions by humans. In that case, its distinct features will be gradually overshadowed by the narrative indicators, which are characteristic of the subreddit. In other words, the more interactions by actual users, the further the data points are pushed towards positive feature scores on PC1.

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. The sub-groupings can be traced back to different kinds of rule violations, which suggests that they can 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 (e.g., 8ajadx). This, then, also explains why only a few hundred texts from this subreddit 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 also 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. 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 indicators of problem solving 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 (e.g., 8b1vmy). The premise of unsent letters – personal messages that users were too afraid to post – explains this 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, past tense, and time adverbs. For example, in the text 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.

4. Discussion

The results reveal that subreddits systematically cluster in terms of their linguistic features, suggesting that they can indeed be considered subregisters of Reddit. Conceptually related communities generally cover similar areas, attesting to a continuous space of variation due to the hybridity of web registers (Neumann and Evert, 2021, 152). Specifically, it seems that the majority of subreddits display fea-

tures of involvement, which is expected of a social media site for discussing specialised interests. The analysis has also demonstrated striking overlaps with offline registers, which valorises online registers as registers proper. Based on salient similarities with other web registers, 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 are otherwise hard to find. By demonstrating that they can be differentiated linguistically via computational means, these findings pave the way toward automated functional web categorisation in informational retrieval.

A significant variable unique to the internet, and particularly public discussion forums, that emerged in this study but has so far been ignored in research of register variation on the web is moderation, which shapes the context of online communication not only socially and 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 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 research.

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. Choosing the thread as the unit of analysis under the assumption that each of them constitutes a single asynchronous 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 comment threads, it seems that contextual parameters often operate at lower levels of instantiation, either in local branches or perhaps at the level of individual contributions. This was reflected in the fact that the effects of ratings and other user interactions could not be properly accounted for as part of the tenor of discourse. Any future investigation of the sociolinguistic dynamics on the internet in systemic functional terms 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 investigates text at some level below the thread.

5. References

Berber Sardinha, T. (2014). 25 years later: Comparing internet and pre-internet registers. In Tony Berber-Sardinha et al., editors, *Multi-Dimensional Analysis, 25 Years on: A Tribute to Douglas Biber*, pages 81–105. Benjamins, Amsterdam/Philadelphia.

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

Biber, D. and Egbert, J. (2018). *Register Variation Online*. Cambridge University Press.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Chang, J. P., Chiam, C., Fu, L., Wang, A., Zhang, J., and Danescu-Niculescu-Mizil, C. (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*, pages 57–60. Association for Computational Linguistics.

Crystal, D. (2001). *Language and the Internet*. Cambridge University Press.

Diwersy, S., Evert, S., and Neumann, S. (2014). A weakly supervised multivariate approach to the study of language variation. In Benedikt Szmrecsanyi et al., editors, *Aggregating Dialectology, Typology, and Register Analysis*, pages 174–204. Linguae & Litterae, Berlin.

Evert, S. and Hardie, A. (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics Conference 2011*, University of Birmingham.

Garside, R. and Smith, N. (1997). A hybrid grammatical tagger: Claws4. In Roger Garside, et al., editors, *Corpus Annotation: Linguistic Information From Computer Text Corpora*, pages 102–121. Longman, London.

Halliday, M. A. K. (1978). *Language As Social Semiotic: The Social Interpretation of Language and Meaning*. Arnold, London.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347.

Liimatta, A. (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, C. M. (2019). Register in systemic functional linguistics. *Register Studies*, 1(1):10–41.

Neumann, S. and Evert, S. (2021). A register variation perspective on varieties of english. In Elena Seoane et al., editors, *Corpus Based Approaches to Register Variation*, pages 143–178. de Gruyter, Berlin.

R Core Team, (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

r/ListOfSubreddits. (2018). List of subreddits. https://www.reddit.com/r/ListOfSubreddits/wiki/listofsubreddits. Accessed 2023-05-15.

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