Persuasive language and features of formality on the R/CHANGEMYVIEW subreddit.

Abstract

The paper investigates formal language in persuasive discourse on the R/CHANGEMYVIEW subreddit. We collected a corpus of 100 million messages, split into subcorpora based on the user-awarded marker *delta*, which rewards changing an original poster's view. Assuming that formality/informality is potentially an important factor in the persuasiveness of a message, we examine the two subcorpora with respect to formality markers.

The results indicate no systematic variation along the formality/informality continuum between persuasive and non-persuasive posts on R/CHANGEMYVIEW. The posters use personal pronouns, suasive verbs, emphatics, imperatives, elaborate connectors and WH-questions with similar frequency, and express themselves using vocabulary and syntax of similar complexity. Moreover, keyword lists and n-gram rankings indicate no register difference. A qualitative analysis of concordance lines for *persuade* and *change PRONOUN view* paints a picture of a community that values factual, evidence-based discourse and openness to logical persuasion, with a linguistic norm of relatively formal, sophisticated register.

Keywords: corpus pragmatics; change my view; reddit; register variation; persuasive language

1. Introduction

The community of users of the subreddit R/CHANGEMYVIEW (CMV), a forum where users come to post an opinion and invite others to change their view, takes a stance on online discussion that is different from much of social media. The CMV *About* blurb reads: "A place to post an opinion you accept may be flawed, in an effort to understand other perspectives on the issue. Enter with a mindset for conversation, not debate." This focus on cooperative engagement stands in sharp contrast to the 'outragefication' of social media, the tendency to play on moral outrage and polarizing topics to maximise reposts and likes.

Social media platforms have done more to influence communication than giving us a new means of reaching across the physical space. The information environment we exist in, the so-called *like economy* (Gerlitz and Helmond 2013), measures success in terms of views, clicks, retweets and other metrics of engagement with content. The importance of language to achieve user engagement online provides a link to the topic of the special issue, *formality and informality*. The continuum from formal (ritual, complex, prescriptively correct) to informal (casual, relaxed, non-standard) register is understood here in the Hallidayan sense as a variety of speech usage, with different registers available to the person dependent on the speech situation (Halliday 1978). In this article, we will investigate whether linguistic features of a persuasive message, characteristic of a certain register,

reflect the success or failure to persuade and ultimately to achieve personal validation through social media metrics.

This environment, where "user interactions are instantly transformed into comparable forms of data and presented to other users in a way that generates more traffic and engagement" (Gerlitz and Helmond 2013: 1349), inevitably changes how we talk and how we discursively construct our identity. Against the background of sensational headlines and microcelebrity (see Senft 2013), R/CHANGEMYVIEW stands out for its ostensibly different focus. Reddit users, including those on the CMV subreddit, may also aim for social validation in the form of *karma,* the platform-inherent community evaluation similar to likes (e.g. Facebook), kudos (e.g. LinkedIn) or upvotes (e.g. YouTube) on other networks. Importantly, however, the main persuasion capital for comments on CMV is another validation system, called *delta*, which can only be obtained if the commenter successfully achieves a 'change of view' of the original poster. Comments which achieve this goal are called DACs (delta awarded comments), whereas all other comments are non-DACs. 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, and a CMV corpus thus allows the comparative study of persuasive discourse that has persuaded, or has failed to persuade.

This paper examines one linguistic dimension that has been emphasised in the literature as relevant to successful audience design in persuasion: formal vs. informal register. Using a 148-million-words corpus of R/CHANGEMYVIEW, we address the following research questions:

1. What metadiscourses of persuasion can be identified in both subcorpora?
2. Do the Delta and non-Delta subcorpora exhibit linguistic variation in lexical and grammatical complexity along the formal/informal dimension?
3. Can a difference in genre norms between the Delta and non-Delta subcorpora be detected?

In section 2 below, we review the main approaches to linguistic persuasion and then focus on the literature that investigated the formality/informality dimension, and its mapping to lexico-grammatical complexity, in persuasive discourse. The final subsection 2.3 summarises the existing studies of R/CHANGEMYVIEW, which fall under the umbrella of Natural Language Processing. In section 3 we describe the corpus and outline the methodological procedures used in the analysis, corpus-driven as well as corpus-assisted. Section 4 is dedicated to the qualitative analysis of suasive expressions in DACs. This analysis is intended to check whether the formality/informality continuum, which emerged in existing literature as relevant to the success of persuasion, is also salient to the CMV community members themselves. Section 5 presents the results of the quantitative analysis of formality markers in both corpora and of the keyword and n-gram analyses.

Our quantitative findings indicate that there is no systematic variation between delta and non-delta CMV posts when it comes to the simple measures of lexical and grammatical complexity. This validates the approach of more pragmatically- and discursively-based work, which takes into account the importance of context, genre and other factors related to *appropriateness*.

2. Setting the scene: language and persuasion

2.1 Defining persuasion

The topic of persuasion has received attention from every field of human sciences (for a broad review from a linguistic perspective with a special focus on pragmatics, see Rudolf Von Rohr [2018]). In this section, we will focus on the literature from pragmatics and discourse analysis strands of linguistics and from register studies. The general framework of the paper is corpus linguistics and corpus-assisted discourse analysis. We aim to establish on the basis of existing literature whether formality/informality dimension is relevant to persuasion, to identify microlinguistic features relevant to formal/informal register and to study variation in these features across subcorpora and to identify discourse strands using keyword and collocation techniques.

The distinction between (1) persuasion as a process, i.e. the attempt to be persuasive, and (2) persuasion as an effect, i.e. having persuaded someone, is sometimes rendered as persuasion and influence (Segal 2005). We follow this distinction and measure influence in terms of the CMV-inherent delta system, but focus our analysis on the linguistic realisation of persuasion (process). Following Lakoff (1982:28), persuasion is understood then as "an attempt or intention of one participant to change the behaviour, feelings, intentions, or viewpoint of another by [linguistic] communicative means."

Lakoff's further specification that persuasive discourse consists in non-reciprocal attempts to affect others enables the study of persuasion from the angle of social power, which has often been taken by Critical Discourse Analysis studies (Fairclough 2001, van Dijk 2008). Van Dijk (2008: 15) points out that institutions have a privileged position to persuade and imprint ideologies since they hold preferential access to public discourse. Participatory Web and social media have been changing this balance of power, however, and now individual speakers can potentially reach millions of eyes and ears – a prerogative previously reserved for traditional mass media. Looking ahead to our own study, it is important to note, however, that participatory Web's's expansion of the individual's communicative reach does not necessarily cause a shift away from dyadic engagement towards a one-to-many model of traditional mass media. One of the interesting aspects of CMV's delta system is precisely that it hinges on the subjective perception and assessment of persuasiveness of the individual comment by an individual poster.

Since Aristotle, persuasion has been theorised as containing three aspects: logos (the use of arguments), pathos (emotional involvement of the audience) and ethos (credibility of the speaker) (Cockroft and Cockroft 2005). The first component lends itself well to the study by means of testing the change in performance by using two alternative versions of the same stimulus to investigate the effects of changing variables on the recipients, as has been done in the tradition of psychology and communication science. The other two have to do with persuasion as an interpersonal process that relies on relational elements to change the addressee's behaviour – an observation which led Rudolf Von Rohr (2018: 4) to identify the interpersonal pragmatics framework (Locher and Watts 2008) as a useful way to study the linguistic realization of persuasion.

Adopting this framework means recognising that interpersonal aspects of a message – including the use of a more formal or informal register – are as important as transactional ones to achieve a persuasive effect. This perspective is supported by existing pragmatics research. Langlotz and Locher (2013), for instance, in their work on online disagreements underscore the role of emotion in relational work, with linguistically signalled emotions marking the shift from persuasive to confrontational communication. Other sources cite proximity and similarity between the

communicators as a factor that increases persuasiveness of messages (Wright 2015, Thurnherr et al. 2016). In the section below, we will review the findings of linguistic studies that identify specific markers of persuasive discourse, with a reference to the formality/informality continuum.

2.2 Linguistic aspects of the persuasive message

Although persuasive discourse can occur in ordinary conversation, Lakoff (1982) points out that some genres fall into the persuasive bin more readily than others. As examples, she names advertising, political rhetoric and religious sermons. It would come as no surprise then that these genres have garnered much attention from linguists. The literature cited below provides evidence that the formality/informality dimension is relevant to emotional involvement of the audience and to constructing credibility of the speaker – two of the cornerstones of successful persuasion.

The need to appeal to both aspects of desirable identity – rational and relational – has been captured in advertising and sales discourse as *reason* vs. *tickle* dichotomy (first introduced for copywriting strategies by Bernstein 1974, qtd. in Simpson 2001). *Reason* advertisements suggest a motive or reason for purchase; *tickle* advertisements appeal to humour, emotion and mood. From a classic pragmatics perspective, these strategies can be associated with different inferential paths the addressee has to follow: the *reason* message is simply and directly communicated, it is bald-on-record and maximally efficient in its presentation of the qualities of the product, the needs it satisfies and the grounds for buying it. The *tickle*'s best route is an indirect one, appealing to feeling via emotion, imagination and poetic truth. *Tickle* is indirect, it relies on implicature and off-record formulations (Simpson 2001).

The dichotomy reminds a linguist of the concept of involvement used by Chafe (1982) and Tannen (1985) to capture the difference between 'oral-like' and 'literate-like' language: the more involved a speaker (or writer) is, the more attention they pay to the act of communication, to the needs of interlocutors and to the verbal richness of the output. In literature, the linguistic features characteristic of complex, literate-like language have been mapped to formal register, and the features of oral-like simple language to informal register (see Ochs 1979, Chafe 1982, Tannen 1982). Biber's (1988) multidimensional, multifactorial analysis of spoken and written language further revealed that the formality dimension of variation includes microlinguistic features of lexical complexity such as word length, sentence length, and nominalizations.

In political discourse, the *reason* vs. *tickle* distinction has also been mapped onto complex vs. simple linguistic features. Dedaíc (2006: 704), for example, in her long list of linguistic markers studied in connection with persuasive argumentation in politics, names vocabulary choices, vague or imprecise words, syntactic structures such as complexity and passivization and textual complexity, which have all been tied to the formality/informality cline.

Simple persuasive language has been well researched on the example of Donald Trump. Viewed in the light of findings of research on advertising discourse, Trump's speech appeals to feeling and emotion for persuasion, or *tickles*. Trump realises the *tickle* strategy through high sentiment, either negative and aimed at inducing fear, or positive and aimed at self (Liu and Lei 2018, Yaqub et al. 2017). In addition to highly emotionally charged vocabulary, his persuasive tactics rely on informal language to create a feeling of connection with the audience (Ali 2019, Kreis 2017, Quam and

Ryshina-Pankova 2016). This 'simple language' includes short, high-frequency words, short sentences, simple sentence structure, and accomplishes positive self-presentation with simple emotive vocabulary.

Along with political speech, specific linguistic devices that carry *reason* and *tickle* strategies have been identified for advertising discourse. The power of simple language to *tickle* is exploited in clickbait, which also frequently contains features such as vivid concrete nouns, numbers, short words, personal pronouns, sentiment words, orality punctuation, vocatives, deictics and questions (Kuiken et al 2017, Blom et al. 2014). Literature on advertising (Boyland et al 2011, Glinert 2005, Popova 2018, Labrador et al. 2014, Vestergaard and Schroder 1985) usually links *tickle* to spoken-style language that minimises the distance between advertiser and customer and appeals to emotion. Such language includes the present continuous tense, imperative constructions, modal verbs, the passive voice and impersonal sentences, hyperbole, aesthetic linguistic innovation ("minty good"), vague quantifying expressions, second person pronouns, emphatic enumeratives ("another plus is…"), multiple modification (e.g. "simple, versatile software"). *Reason* is linked to formal, planned language and strategies of justified argumentation, for example, citing of facts and statistics through numerals, appeal to the authority of science or medical profession (sometimes reinforced by enactment of consumer ignorance and embarrassment) and remote (uninvolved) tone of the narrative.

The choice of linguistic structures in addressing readers is crucial with respect to the involvement aspect of persuasion (Durant and Lambrou [2009], Rudolf Von Rohr 2018: 55). According to Simpson and Mayr (2010: 115), for example, the audience is more likely to align with an expert's point of view if they accommodate to the audience linguistically. Communicative integration of the audience is a key persuasive strategy in advertising discourse on- and offline. This gives rise to the strategy of "synthetic personalization" in one-to-many discourse, defined as communication that is seemingly directed at individuals instead of a large group of people (Fairclough 2001: 52). High-involvement and proximity linguistic strategies include frequent use of personal and possessive pronouns and determiners (Janoschka 2004, Koteyko 2009), specifically those in the first and second person (Benwell and Stokoe 2006, Ng and Bradac 1993). Apart from pronominal reference, questions to the reader/hearer and deictic terms are supposed to create a feeling of immediacy and co-presence (Benwell and Stokoe 2006, Koteyko 2009). All of these – personal pronouns, direct questions and deixis – have been linked to informal register (see Giménez-Moreno 2011). The paper by Giménez-Moreno (2011), along with much subsequent work on quantifiable linguistic features of involved vs. informational style, are based on the lists compiled by Biber (1988) for his seminal investigation of register in English. We will rely on his slightly adapted list for our investigation (see section 3.2).

The importance of audience design and the corresponding adaptation of traditionally formal genres to more informal language has been observed for legal and courtroom discourse. Advice on using simple, persuasive language to juries includes use of familiar language, simple words with few syllables, short and linear sentences and being careful with figurative expressions. Findley and Sales (2012) also recommend avoiding legalese and using so-called "memorable impact words" (reminiscent of Trump's reliance on emotive words, see above). Durant and Leung (2017:76) summarise the research on the topic as follows:

> "Jurors may fail to attend to or understand – or may even appear bored by – evidence presented to them. One requirement of verbal communication to jurors is therefore simplicity, despite the complexity of the subject matter and legal framework governing what has to be communicated."

To sum up, there appears to be a clear link between linguistic simplicity characteristic of informal register and effective persuasion. Linguistic complexity characteristic of formal register appears to be less successful in achieving emotional engagement.

However, the current thinking in the vein of interpersonal pragmatics recognises that a specific discursive effect is not achieved by resorting to a static set of linguistic resources. In contrast, the question is whether a particular linguistic rendition conforms to the norms of a particular practice in a particular context (Locher 2013). That is, whether a linguistic action is successful depends on whether it is appropriate. Simple quantifications of lexico-grammatical features are not sufficient to establish appropriateness of a persuasive message. To form an understanding of the CMV subreddit as a genre and of its linguistic norms and context, we review the existing studies of the subreddit from neighbouring disciplines in section 2.3 and conduct a qualitative analysis of our data in section 4.

2.3 Extant studies of CMV

To our knowledge, no research of CMV has been published within corpus linguistics or linguistics more generally. However, there have been studies based on text/data mining research paradigms, which we will briefly summarise in this section. Our focus is on the linguistic features that have been included and identified as predictors of persuasiveness in these studies.

First insights into CMV discussions were provided by Tan et al. (2016), who examined data posted in the first two years of CMV's existence (2013–2015). Their study finds that the earlier a commenter joins the discussion, the higher is the likelihood for her/him to win a delta. They also observe a significant correlation between the number of users in the thread and the likelihood of the original poster (OP) changing their view. Interestingly, an extensive back-and-forth between the OP and the commenter is unlikely to result in a change of view, which suggests that the OP decides early on whether they find an argument persuasive. Testing the similarity of linguistic features, Tan et al. (2016: 618) establish that successful arguments are less similar to the original post in content words, but more similar in 'stopwords', i.e. those function words often excluded from language models in NLP research (such as *the* or *of*). Longer replies are strongly correlated with success. The authors also examined which linguistic features of the original post correlate with delta awards, a characteristic they label "malleability of opinion". The finding is that first person singular pronouns indicate malleability, while first person plural pronouns indicate resistance – a result that Tan et al. (2016: 621) tie to psychological research linking self-affirmation with open-mindedness, and that appears to us somewhat forced. Although the report by Tan and colleagues is undoubtedly interesting as a first description of CMV discourse, it is uninformed by linguistic theory and leaves ample space for more in-depth study.

Wei et al. (2016) study popular CMV threads (containing more than 100 posts) with data posted during one year and argue that *karma* is a more promising metric of persuasiveness than deltas – mainly because of the relative rarity of DACs (in their data, only 0.5% of comments received a delta).

Their approach to persuasion is realised as a ranking task which includes social interaction features (related to the comment tree and the position of the individual comment within the tree) as well as linguistic (surface text and argumentation related) features as potential predictors for *karma* and thus persuasiveness. Their Machine Learning algorithm proved to work best when all social and linguistic features were combined. However, in terms of individual predictors, they confirm Tan et al.'s (2016) finding that early posts are more likely to receive tangible positive feedback. Wei et al. (2016) also find that surface text features (number of words, POS tags, URLs, punctuation) are unimportant, whereas argumentation related features result in the best performance. These features consist of two sets of metrics, a local one, which is based on sentences identified as argumentative or non-argumentative by a Machine Learning based classifier[1], and an interactive one, which is based on the similarity (regarding term frequency) of comments with original posts and previous comments. Of these two, interactive features turned out to be more effective (Wei et al. 2016: 198).

Hidey et al. (2017) approach persuasion as part of argumentation. They use the dataset from Tan et al. (2016), but include (1) manual annotation by experts, who distinguished claims from premises; (2) crowdsourcing to classify claims according to Aristotle's modes of persuasion, interpretations and evaluations, and agreement and disagreement; and (3) quantitative analysis of correlations between annotated claims and premises. Looking at the dataset in general, they find certain patterns regarding the distribution of *logos, pathos* and *ethos,* including that premises in posts frequently orient to *logos* and *pathos* and only rarely to *ethos*. Furthermore, arguments of the same Aristotelian type often occur in sequence (e.g. *logos* followed by *logos*). When it comes to the comparison between DACs and comments without award, Hidey et al. (2017) find small, but significant differences. Their results show that DACs rely less on rational evaluations and more often combine *pathos* and *logos* arguments. DACs also contain more agreements early in the post, which the authors interpret in terms of the rhetoric strategy of introducing a counterpoint with agreement.

In addition to the three summarised studies, two further studies with more specific interests have been conducted on CMV datasets. The first one, by Jhaver et al. (2017) is interested in "design mechanisms and social norms" at play in communication on the subreddit. Based on participant observation and structured interviews the authors find the main motivations for contributing to CMV to be earning deltas, changing other people's opinions, engaging in threads of interest to the posters and learning techniques of persuasiveness. Original posters, on the other hand, seem to be motivated by crowdsourcing information, a lack of satisfying face-to-face conversations on the topic at hand, the unsuitability of other online platforms and the active CMV community. Users, when asked whether CMV really changes minds, responded that many original submissions get insufficient responses to be of use to the OP. They also noted that many original posters are not truly interested in changing their views, but use the platform to simply post their opinion and that the changeability of opinions is tied to the topic. Finally, the study raises the question whether CMV will inevitably only cater to those who are already open minded, whereas those who are set in their views will simply not post on this subreddit.

[1] Wei et al. (2016) do not specify what features they used to identify argumentative sentences. They only mention that they used features proposed by Stab and Gurevych (2014). We can thus only infer that the classifier Wei and colleagues used items from the following list of features: "*major claims*, *claims* and *premises*, which are connected with argumentative *support* and *attack* relations" (Stab and Gurevych 2014:49, their emphasis).

Priniski and Horne (2018), finally, examine the role of citing evidence on CMV. The authors define evidence as reference to urls or use of statistical language and model persuasiveness again by use of delta awards. In addition, Priniski and Horne were interested in comparing those topics that are sociomoral (e.g. gender identity) and thus of interest to large parts of society, to other non-sociomoral topics (e.g. films). The results of their analysis of 100'000 comments from 500 threads reveal 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 is more likely awarded when evidence was provided.

While not all the results of these studies can readily be transferred to our own project, the extant literature lends support to our premise that deltas can be and have been used as a metric for persuasiveness in CMV communication. However, we do not expect to find a straightforward correlation between any surface linguistic measure and a persuasiveness effect. It is our belief that to investigate persuasiveness, a deeper look at the genre and community norms using the pragmatic and discourse analysis methodology is required, since the same linguistic resource can potentially aid or interfere with an effective persuasive message.

3. Data and method

3.1 r/ChangeMyView: describing the corpus and the data source

ChangeMyView is a channel, or 'subreddit', on the online forum platform Reddit. With 430 monthly active users, Reddit has become a major locus for building relationships and communities. Within Reddit, the CMV subreddit is a self-described place "to post an opinion you accept may be flawed, in an effort to understand other perspectives on the issue" (https://www.reddit.com/r/changemyview/). Typically, an author interested in engaging in a discussion would create a new posting stating their opinion and inviting other reddit users to "change my view". Figure 1 below illustrates the basic structure of a CMV thread:

↑ 25.9k **Posted by OP 23 hours ago** **CMV: The average homeowner does not benefit from constantly rising house prices**	1. *Number of upvotes by other users* 2. *Concise statement of the original poster's opinion*
I often hear that consistently inflation beating rises in house prices are A Good Thing. People who own houses seem very happy that their house has increased in monetary value, despite the fact that the utility they get from it has not increased at all. [...]	3. *Detailed statement of the original poster's opinion*
B 21 hours ago	4. *Persuasive comment by user B*

All property does not increase at the same rate. [...]	
OP 23 hours ago This approach only works for people who are free to move easily and whose income is not linked to where they live, or for investors who do not live in the property. [...]	*5. OP states they are not convinced by B's comment*
C 20 hours ago You are not taking into account leverage (i.e. taking out a loan to buy a house) which almost 100% of home buyers do. [...]	*6. Persuasive comment by user C*
OP 19 hours ago That is a good illustration of why falling into negative equity is a problem. Maybe never having been in that situation I underestimate the severity of the issue. [...] !delta for highlighting the importance of avoiding negative equity.	*7. OP states their opinion has been changed by user C* *8. OP state they would like to award a delta to C's comment*
DeltaBot 19 hours ago Confirmed: 1 delta awarded to C	*9. DeltaBot awards a delta to C's comment*

Figure 1. Basic structure of a CMV thread.

In an attempt to change the OP's view, users post persuasive responses to the original opinion, sometimes purely deliberative, sometimes with links to outside sources and statistics (see Priniski and Horne 2018). If the OP finds a comment convincing, he/she may reward them with a delta, the persuasive currency provided by this subreddit. The delta award is then confirmed by the Delta bot, which checks that the delta-awarding comment contains at least 50 characters of text. More than one delta can be awarded within a single thread. Furthermore, human moderators make sure that original posters elaborate on their opinions and that the discussion remains civil and articulate.

For a researcher, this setup provides a unique opportunity to examine persuasive language. The combination of naturally occurring, publicly available language data and pre-existing annotation for successful persuasion by the users themselves enables corpus research into linguistic variation of effective persuasive language. Apart from deltas as a metric of successful persuasion, CMV also makes use of a more collective measure called *karma* (see Section 2.3; Wei et al. 2016). We solely focus on delta awards in this study (see Section 3.2), since we understand deltas as an index for changing the OP's view, whereas karma upvotes mean that someone in the community found something to like about the respective comment.

For our larger research project on CMV, we collected a corpus (the 'CMV corpus') of all comments posted on CMV between May 2013, when the first content appears, and May 2020, when the data were collected. The data were accessed via a Python script and through the pushshift.io Reddit API. Subsequently, comments tagged as deleted on the subreddit were removed, as were those with a warning that they were in breach of subreddit rules. The current study is based on two subcorpora: the Delta subcorpus, which consists of all DACs at least 50 tokens in length within the CMV corpus; and the non-Delta subcorpus, which contains a sample of 500,000 non-DACs from the CMV corpus. The sample was obtained by first applying criteria in terms of comment length (124 or more and 2000 or fewer tokens[2]) and then randomly selecting 500'000 comments. Table 1 presents the distribution of comments and words in the corpus and the two subcorpora.

	CMV Corpus			**Delta subcorpus**			**non-Delta subcorpus**	
Comments	6,060,217			54,680			500,000	
Words	668,944,426			14,554,306			133,761,424	
avg. words per comment (wpc)	110.38			266.17			267.52	
Standard deviation wpc	143.16			238.95			187.66	
median wpc	65			193			206	
DACs (delta awarded comments)	61,904	1.0%		54,680	100.0%		0	0.0%
DAC words	14,770,367	2.2%		14,554,306	100.0%		0	0.0%
non-DACs	5,998,313	99.0%		0	0.0%		500,000	100.0%
non-DAC words	654,174,059	97.8%		0	0.0%		133,761,424	100.0%

Table 1: Overview CMV Corpus, Delta subcorpus, non-Delta subcorpus

Our main corpus and subcorpora were tokenised, lemmatised and POS-tagged with spaCy (Honnibal and Montani 2020) and encoded in CWB (Corpus Work Bench, Evert and Hardie 2011). Our quantitative work was done in R (R Core Team 2020), where we used the polmineR package (Blaette und Leonhard 2020) to access the CWB-encoded corpora. However, all analytical steps can also be reproduced by accessing the same corpora directly through CWB (using CQP syntax).

3.2 Method

We first approach the data in our two subcorpora (Delta, non-Delta) qualitatively. We examine a random sample of a 100 concordance lines for *persuade** and *change* PRONOUN view* in the Delta subcorpus to explore the metadiscourse about the subreddit's main activity. The aim of this analysis is to confirm that the formality/informality linguistic dimension that emerged as relevant in the

[2] As a starting point for our sampling, we set as the only selection criterion a length of at least 50 characters to match the Delta bot selection criteria. We then started excluding the longest and shortest messages until we reached a similar average comment length in both corpora (266.17 and 267.52 wpc, respectively).

literature review (or, indeed, any other linguistic dimension) is also salient to the CMV community members.

We then move on to the quantitative analysis. To operationalize our interest in the formal/informal dimension, we adapt the features identified by Biber (1988) for Dimension 1 (involved vs. informational) that have appeared as salient in the discussion of the literature above. Table 2 below lists the features according to whether the rise in their frequency is expected in a more informal, casual style (left column) or in the more formal, standard style (right column). For instance, we would expect a longer average word frequency, longer average sentence length and higher lexical complexity in a more formal style. It is important to remember, however, that formality/informality is not a binary characteristic of a text, but a continuum.

We additionally look at the category of suasive verbs (Quirk et al. 1985: 1182–1183). Suasive verbs have been named as an important semantic category for argumentative discourse because they introduce indirect directives and imply an intention to bring about some change in the future. Examples of such verbs include *agree, demand*, or *insist*. Appendix 1 provides the full list of all the features in Table 2.

Informal/casual	Formal/ritual
• First and second person pronouns • First and second person pronominal possessives • WH questions • general emphatics • imperatives	• average word length (AWL) • lexical complexity (type-token ratio [TTR] and lexical density metrics) • sentence complexity (average sentence length [ASL] in words per sentence) • general hedges • elaborate connectors (e.g. *furthermore, nevertheless*) • nominalizations • suasive verbs

Table 2. Linguistic parameters of formal and informal register for the present study.

In a second step, we complement this register-based approach with a keyness comparison between the Delta and non-Delta subcorpora in the tradition of corpus-assisted discourse analysis (CADS, see e.g. Baker et al. [2008]). This allows us to gain insights into the discourses that participants rely on in their CMV contributions. We extracted the top positive keywords in DACs with the help of the Quanteda package for R (Benoit et al. 2018) using two different measures:

The first measure, %DIFF, measures effect size based on word frequency (Gabrielatos and Marchi 2011). It states how much more frequent a keyword is in a corpus compared to a reference corpus.

The second measure, text dispersion %DIFF (TD%DIFF), is our own measure of effect size based on the combination of %DIFF and "text dispersion keyness" (TD, proposed by Egbert and Biber 2019 as a more adequate keyness measure for very large corpora). TD%DIFF calculates effect size by comparing normalised text dispersion in a corpus of interest (here Delta) and a reference corpus

(here non-Delta). It answers the question, how much more or less dispersed a keyword is in either corpus.

For both measures, we calculate significance with log likelihood (LL). LL is a traditional keyword metric that identifies which words are more and less frequently used in the corpus in question; or, for TD%DIFF, which words are more and less widely dispersed in the corpus in question. The difference in both %DIFF and TD%DIFF was found to be significant at $p<0.01$ ($LL>6.8$) for all the examples in Section 5.2. We additionally tried to mitigate the effect of specific topics skewing keyness by setting the cut-off point of text frequency to 1% of total texts in each corpus (meaning that keywords that occur in less than 1% of texts will not appear in the list).

Apart from keywords, we are also interested in multi-word clusters. We extract the top 20 bi-, 3- and 4-grams from the Delta subcorpus and from a random sample of the non-Delta subcorpus of comparable size and compare their inventories in delta and non-delta comments. It has been demonstrated that n-grams are an important means of differentiation of registers or genres (Biber et al. 2004). Academic language, for example, contains distinctive high-frequency n-grams that characterise the academic discourse conventions, such as *the importance of* (Greaves and Warren 2010, Carter and McCarthy 2006). We therefore hypothesise that if there are detectable differences in register between the successful persuasive messages (Delta subcorpus) and the unsuccessful ones (non-Delta subcorpus), they will be manifest on the level of n-grams.

4. Discourses of persuasion in CMV

4.1 Talking about *persuading*

Interpersonal pragmatics research has demonstrated that persuasive discourse hinges on constructing the identity of a "successful persuader" through claims to credibility, expertise and authority (Rudolf Von Rohr 2018). This can be done strategically, with the conscious effort of the speaker, or unconsciously (best understood in the Goffmanian terms of signals *given* and *given off,* Goffman 1956). Although studying linguistic choices usually means focussing on the signals *given off* (people seldom consciously decide to increase the relative frequency of nominalizations in their speech or decrease sentence to clause ratio), looking at the *given* signals is no less interesting.

To see what meta-discourses exist around the ostensive interactional aim in the CMV subreddit – the aim to change the OP's view and to win a delta – we examine a random sample of 100 concordances for the lemma *persuade* and the cluster *change* PRONOUN *view* in the corpus of successful comments, the Delta subcorpus. This analysis is interesting from the angle of formality/informality as the Latinate verb "persuade" sits firmly among the markers of more formal style, whereas "change someone's mind/view" is a more casual synonym. It also allows us to ascertain that community members see language as relevant to the aim of persuasion and do not place value solely on the factual aspect of argumentation.

Discarding those instances of *persuade* and *change my view* that do not refer to the CMV activity itself but rather to whatever topic is being discussed, we form the following tentative picture of the community's own view of persuasive activities.

Three types of usage emerge for the lemma *persuade*: negatively connoted, neutral and positively connoted (see Table 3). These connotations are recognised in the stance of the author, who makes an effort to disalign with 'persuasive behaviour' (for example, in 1) or align with it (example 3).

Category	Example (emphasis in bold is added by the researchers)
Negatively connoted	(1) I haven't been 'trying to **persuade** these CMV redditers to stop this behaviour'. I've been trying to explain why this behavior feels unsettling and upsetting to many women.
Neutral usage	(2) It's not a data driven approach, so maybe it won't **persuade** you, but I think my personal experience with Mexican immigration to California's construction industry helps highlight the positives and negatives of immigration.
Positively connoted	(3) However, I am not posting this message in order to seek attention. I have other motives, such as to **persuade** you and get my point across.

Table 3. Three types of semantic prosody for lemma *persuade* in the Delta corpus.

Many concordances involve users denying their attempts to persuade anyone, indicating that this particular suasive verb might have negative semantic prosody in the community (such as example 1, where the user takes issue with another commenter describing their actions as persuasion). This could be related to Rule 2 of CMV, which reads "Change My View is meant to be a place where a person with an unpopular view could go to learn about the other side of issue, to try and understand different perspectives and do so without fear of being attacked" and prohibits hostile, aggressive, or even sarcastic arguments. Since violation of Rule 2 leads to the deletion of the comment, users are sensitive to phrasing which might appear overly self-assertive or explicit. This interpretation is supported by the frequent discussions of the word meaning of 'persuade' on CMV. Examples 4 and 5 below, for instance, frame persuasion as manipulation and contrast it with evidence-based discourse.

(4) So, to try to find _some_ merit in what he says, the emphasis should be on swapping reasons and evidence, as opposed to having as a goal to **persuade** someone else.

(5) It's more about them being convinced and not merely **persuaded**, - convince I'd say involves being compelled to abandon a 'view' by encountering a view supported by better reasoning than it, whereas a person could be **persuaded** by all sorts of things including pandering, appeals to authority, bribery, coercion, etc.

When using the lemma *persuade* positively, users in this sample emphasise that they are keeping to the aims of the CMV community and staying on topic (as example 3 above illustrates). These aims can be referenced explicitly to explain why persuasion is a good thing:

(6) Engagement and willingness to be **persuaded** are requirements here, check the sidebar and re-read the rules if you need a refresher.

On the material of this random sample, the negatively connoted group outweighs the positive group, although we cannot be certain if this is representative of the corpus on the whole.

4.2 Talking about *change PRONOUN view*

The concordance of the *change PRONOUN view* cluster in the Delta subcorpus is dramatically different. It is overwhelmingly used with the first person subject to describe one's own discursive activity and with a direct object and recipient of the persuasive act in the second person. A case in point is example 7 below:

(7) I'm responding to **change your view** because you say you don't see where the ethical problem rests, so I'm explaining where it rests to **change your view.**

Other usages include requesting a delta for a change in view and asking the OP whether their view has been changed. All of these are essentially meta-discourse around the stated aims of the subreddit and are given legitimacy by the forum itself, in contrast to *persuade*, which requires users to engage in explanation and self-justification. This community acceptance makes the cluster especially interesting to examine for representations of the persuasion process and what the users see as main factors in a successful change of view.

One aspect of the persuasion process that the users cite repeatedly is the willingness of the OP to be persuaded:

(8) But your approach has been to say No, no, no! to everyone who has tried to change your view. You've given no indication of what would **change your view**.

They highlight that the writer needs to be open-minded, on the one hand, and that the view itself needs to have certain characteristics to be subject to change, on the other hand. The presupposition that some posters may in fact not be willing to change their view matches the suspicions uttered by some respondents in Jhaver et al. (2017) that some posters use CMV to post opinions rather than to be persuaded. Example 9 recognises the distinction between a fact-based view and a personal preference, with the latter being a matter of taste and therefore impossible to change through argumentation:

(9) We can't really **change your view** of a personal opinion. If you don't enjoy ET then you don't enjoy ET.

On the whole, users explicitly appeal to the role of facts and evidence as the main instrument of persuasion, at the same time assuming that the facts will be interpreted in more or less the same way by all the readers. In fact, 'change your view' is often understood as 'provide evidence that your view is incorrect', such as example 10: an assumption that has more in common with communities of practice sharing a codified set of reasoning procedures (e.g. Western biochemists, cf. Gilbert & Mulkay 1984), rather than with mundane conversation.

(10) First, you can **change your view** multiple times, and you claimed secular buddhism is the most common, and I disproved that with numbers.

The values and descriptions brought up in the meta-discourse about persuasion lead us to conclude that CMV users perceive the appropriate genre features of the subreddit as being close to

those of academic writing. Such features are the organisational structure based on the Problem-Solution pattern (Flowerdew 2000) and the moves characteristic of empiricist writing such as reference to previous research, explanation, exemplification and deduction (Swales 1990). Thus, examples 8 and 9 fall short of appropriate evidence-based discourse by failing to identify a solvable problem; example 7 explicitly references the explanation move; example 10 references existing research and assumes the corresponding deduction on the part of the addressee.

The cues that community members attempt to *give* include many of the features we identified as 'formal': use nominalisations (*immigration*, *reasoning*, *indication*) and complex vocabulary (*data driven approach, secular Buddhism, being compelled to abandon*), hedge one's statements (*merely, I'd say*), construct sentences of several clauses strung together by logical connectors (see example 4) and avoid imperatives and emphatics that can be seen as overly colloquial. This confirms that the community members recognise the formality/informality continuum as a resource to draw on in constructing a persuasive message.

5. Delta and non-Delta corpora on the formality continuum: corpus findings

5.1 Results of pairwise comparison of formality markers in the two subcorpora

Pairwise comparisons of the formality/informality markers in Table 2 above established that no systematic difference of large or medium effect size between the Delta and non-Delta corpora can be detected. This means that insofar as our corpus-assisted approach accurately models formality and informality, there is no difference between DACs and non-DACs that would be interesting to us in terms of linguistic register markers (see Appendix 2 for the detailed results of this analysis). While comparing the frequencies has yielded significant p-values ($p<0.001$) in all cases but emphatics and average word length, significance is strongly affected by the number of observations, and even very small discrepancies are significant in very large samples such as ours. Indeed, the Vargha and Delaney A effect size statistic demonstrated only negligible effect sizes for all the detected differences. In other words, we can establish with high confidence for most measures that there is indeed a difference between DACs and non-DACs, but must concede at the same time that this difference is likely too small to be of any relevance.

Table 3 below documents the average values for the formality markers in the Delta and non-Delta corpus. Values expressed in frequencies (starting from row 5 in Table 4) were normalised per 1000 words, while the proportional values (Type-token ratio, Average word length, Average sentence length, lexical density) are given as is.

N	Linguistic feature	M, Delta corpus	M, non-Delta corpus
1	Type-token ratio	59.7	57.5
2	Average word length	5.1	5

3	Average sentence length	19.8	19.4
4	Lexical density	64.4	64
5	First person pronoun singular (I)	0.02	0.02
6	Second person pronoun (you)	0.02	0.02
7	Suasive verbs	0.003	0.003
8	WH questions	0.001	0.001
9	Emphatics	0.01	0.01
10	Imperatives	0.001	0.001
11	Hedges	0.001	0.001
12	Elaborate connectors	0.003	0.003
13	Nominalisations	0.02	0.02

Table 4. Average values for the formality markers in Delta and non-Delta corpora.

To put these descriptive statistics into perspective, the values are similar to other informational, prepared genres of English. For example, a study of a corpus of academic law articles (Breeze 2013) yielded an average word length of 4.99 and average sentence length of 20.22[3]. This suggests that when it comes to the dimension of formality/informality, the corpus is relatively homogenous and the texts place closer to the formality end of the continuum.

The conclusion of the quantitative analysis section is that there is no variation between delta and non-delta posts along the formality continuum, or indeed within any post groupings within the CMV corpus (at least formality described by the 13 surface markers derived from literature). In the next section, we look at community discourses around persuasion in order to understand the linguistic norms. This will help us understand why formality/informality is not a relevant dimension in making CMV posts more persuasive.

5.2 Looking at keywords in the Delta corpus

The keywords presented in Table 4 indicate which terms were used significantly more frequently in the Delta subcorpus. We use two keyness metrics here: %DIFF and TD%DIFF.

In terms of register-related differences, the keyword lists confirm our main finding from the previous section, which is that formality/informality does not appear to be a relevant dimension to distinguish between DACs and non-DACs. The top ten keywords cited here, and a further hundred keywords that we examined, indicate exclusively topic-related differences. As Table 5 demonstrates, differences between %DIFF and TD%DIFF metrics are minimal, both of them yielding topic-related

[3] In contrast, an average word length is 4.2 for general fiction and 3.9 for personal letters (Biber 1988: 256-262). An independent t-test showed that these differences are significant (for personal letters vs. CMV at p = 1.331e-12, for general fiction vs. CMV at p < 2.2e-16, for academic writing vs. general fiction at p < 2.2e-16).

keywords. Based on the keyword lists, we can conclude that successfully persuasive posts made unusually frequent references to space, team, or the number five – a finding that is interesting in terms of 'aboutness' of the posts but not indicative of register. The only thematic category that emerges with any consistency is a topos of FILM (subsuming keywords *audience, film, characters* and possibly *team* and *style*). While we can only speculate what might be the cause of this difference (perhaps an OP with a CMV post on the topic of film, who was very generous awarding deltas, or possibly a general tendency for film-related comment threads to award many deltas), it is thematic and not related to linguistic variation.

Frequency-based Keyness				Text Dispersion Keyness		
Keyword	%DIFF	LL		Keyword	TD%DIFF	TD_LL
audience	58.5	80.1		film	51.4	40.8
film	56.7	86.2		audience	44.6	37.5
team	52.3	128.1		delta	38.1	33
delta	51	67.1		dr	37.5	34.2
characters	45.3	79.5		characters	36.1	30.4
style	44.4	48.1		tends	34.9	22.2
space	44.3	119.8		adding	34.8	22.2
tends	41.4	32.8		finally	34.2	38.7
candidates	38.5	55.2		five	34.1	22.6
adding	37.3	27.2		concerns	32.8	22.7

Table 5. Top 10 positive keywords in Delta corpus (reference corpus: non-Delta), ranked by %DIFF and TD%DIFF, word occurring in minimum 540 (~1%, n=54,680) texts.

These content-based findings aside, the conclusion of our keyness analysis is that no "DAC-specific register" can be identified using the CADS methodology of grouping keywords into topoi, which suggests that a "DAC register" is not a useful category. The finding is then that people use similar register throughout the CMV, and that this register can be described as quite formal and standard-like. We can further test this conclusion in the analysis below using n-gram comparison and a closer look at concordances in Delta corpus.

5.3 Looking at N-Grams in the Delta and non-Delta subcorpora

In this section, we move on to the last corpus-driven approach to investigating register differences between the two subcorpora: comparing n-gram ranking. N-grams, especially 3- and 4-grams, have been shown to be better markers of genre than individual words (see section 3.2).
Neither of the rankings of bi-, 3- or 4-grams reveal differences between the two subcorpora. Delta and non-Delta lists both include n-grams of function words such as "of the" or "I don't" at the top, in very similar ranking order (see Table 6). It is possible to make predictions about the genre of the corpus based on the n-gram lists, especially when looking at the longer clusters. For example, 3- and 4-grams reflect a text expressing the author's opinion, as clusters such as "don't think" and "don't know" in combination with the first person singular pronoun indicate. Taken together with the expressions of epistemic modality such as "I'm not sure", stance expressions "I think it's" and discourse organising devices such as "in the first place", the n-grams suggest a literate-like metadiscursive genre. Again, the language is standard English, with a medium degree of formality as indicated by conventional contractions.

	Bigrams				3-grams				4-grams			
	Delta		non-Delta		Delta		non-Delta		Delta		non-Delta	
1	of the	50,120	of the	52,568	I don't	11,372	I don't	17,109	I don't think	3,771	I don't think	5,048
2	in the	38,078	don't	46,068	a lot of	7,529	a lot of	7,184	I don't know	1,427	I don't know	2,182
3	don't	34,780	in the	40,482	you don't	4,752	I'm not	6,335	don't want to	1,417	don't want to	1,806
4	to be	30,110	it's	34,373	don't think	4,590	don't think	6,323	I think it's	1,172	I don't see	1,548
5	it's	29,885	to be	34,024	I'm not	4,007	you don't	5,670	in the first place	1,058	in the first place	1,404
6	is a	23,124	is a	25,453	it's not	3,772	it's not	4,406	I don't see	982	I'm not sure	1,223
7	to the	21,758	to the	23,938	be able to	3,671	don't have	4,357	a lot of people	978	I think it's	1,132
8	it is	17,750	I'm	23,651	don't have	3,595	be able to	4,222	I'm not sure	944	but I don't	1,110
9	on the	16,170	it is	22,506	the fact that	3,009	don't know	3,994	to be able to	928	don't have to	1,041
10	I think	15,794	I do	22,394	don't know	2,865	the fact that	3,890	don't have to	914	to be able to	1,024

Table 6. Top 10 bi-, 3- and 4-grams for Delta and non-Delta corpora, ranked by frequency.

Returning to the results of section 4, we saw that CMV members consider formal language a community norm. This explains the lack of systematic variation on the formality/informality dimension between Delta and non-Delta subcorpora. Although, according to literature, simple language can be a powerful persuasive tool, it is not a tool that the CMV community recognises as appropriate in its discussions. Nor is it a strategy that we would have found to increase a poster's chances at receiving a delta, as this section has demonstrated. In other words, while the emic understanding of a successful comment with delta-potential may go against expectations raised by the extant literature, it is in tune with the linguistic patterns we actually observe in the data.

6. Conclusions

In order to investigate the genre of successfully persuasive CMV posts (DACs), we have employed a corpus-driven approach followed by a corpus-assisted qualitative analysis. The findings show that despite the expectations (which had been informed by existing literature on persuasive discourse in politics and advertising), DACs and non-DACs do not systematically vary along the formality/informality continuum.

Pairwise comparisons of lexico-grammatical indicators of formality and informality, which we had adopted from earlier studies of written English, yielded no differences between DACs and non-DACs of interesting effect size. The CMV posts that have successfully persuaded the OP and the posts that have not received a delta use personal pronouns, suasive verbs, emphatics, imperatives, elaborate connectors and WH-questions with similar frequency and employ vocabulary and syntax of similar complexity.

The next step in exploring the differences between DACs and non-DACs was to look at the rankings of keywords and n-grams in the subcorpora. Neither of these instruments revealed register differences. The keywords, measured using both word frequency and text dispersion metrics, described the aboutness of the posts without pointing at any functional linguistic differences. The n-grams described an identical register picture for both subcorpora: metadiscursive texts relying on first person statements with stance expressions and epistemic modality expressions and discourse organising devices.

Finally, we zoomed in on the way participants talk about their own persuasive activities. A qualitative analysis of 100 random concordance lines for the lemma *persuade* and the cluster *change PRONOUN view* in the delta-awarded posts painted a picture of a community that values factual, evidence-based discourse and openness to logical persuasion. Interestingly, the lemma *persuade* seems to have a pejorative connotation of convincing someone by means other than cold, hard facts. This suggests that *tickle*-based persuasion, effective as it may be in other contexts, is viewed critically in the CMV community.

Appendix 1.

Description is based on Nini (2019).

Prevalence on the formal/informal end	Feature	Description
Informal	1st and 2nd person pronouns	*I, me, us, my, we, our, myself, ourselves*
Informal	1st and 2nd person pronominal possessives	*you, your, yourself, yourselves, thy, thee, thyself, thou*
Informal	WH questions	Any punctuation followed by a WH word (what, where, when, how, whether, why, whoever, whomever, whichever, wherever, whenever, whatever, however) and followed by any auxiliary verb (modal verbs in the form of MD tags or forms of DO or forms of HAVE or forms of BE). An intervening word was allowed between the punctuation mark and the WH word. Furthermore, we exclude WH words such as however or whatever that do not introduce WH- questions
Informal	general emphatics	This tag finds any of the items in this list: *just, really, most, more, real*+adjective, *so*+adjective, any form of *DO* followed by a verb, *for sure, a lot, such a*
Informal	Imperatives (sentence-initial)	Sentence-initial imperatives. We excluded other imperatives because within our methodological paradigm, there is no reliable way to distinguish imperatives from other verb base forms.
Formal	average word length	Mean length of the words in the text in orthographic letters. A word is any string separated by space in the text tokenised by the Stanford Tagger.
Formal	lexical complexity	TTR, lexical variety and lexical diversity metrics
Formal	sentence complexity	average finite verbs per sentence
Formal	general hedges	This tag finds any of the items in this list: *maybe, at about, something like, more or less, sort of, kind of* (these two items must be preceded by a determiner, a quantifier, a cardinal number, an adjective, a possessive pronoun or WH word)
Formal	conjunctions	This tag finds any of the items in this list: punctuation+*else*, punctuation+*altogether*, punctuation+*rather*, *alternatively, consequently, conversely, e.g., furthermore, hence, however, i.e., instead, likewise, moreover, namely, nevertheless, nonetheless, notwithstanding, otherwise, similarly, therefore, thus, viz., in comparison, in contrast, in particular, in addition, in conclusion, in consequence, in*

		sum, in summary, for example, for instance, instead of, by contrast, by comparison, in any event, in any case, in other words, as a result, as a consequence, on the contrary, on the other hand
Formal	nominalizations	Any noun ending in -tion, -ment, -ness, or -ity, plus the plural forms
Formal	suasive verbs	This tag finds any of the items listed by Quirk et al. (1985: 1182–3): *agree, agrees, agreeing, agreed, allow, allows, allowing, allowed, arrange, arranges, arranging, arranged, ask, asks, asking, asked, beg, begs, begging, begged, command, commands, commanding, commanded, concede, concedes, conceding, conceded, decide, decides, deciding, decided, decree, decrees, decreeing, decreed, demand, demands, demanding, demanded, desire, desires, desiring, desired, determine, determines, determining, determined, enjoin, enjoins, enjoining, enjoined, ensure, ensures, ensuring, ensured, entreat, entreats, entreating, entreated, grant, grants, granting, granted, insist, insists, insisting, insisted, instruct, instructs, instructing, instructed, intend, intends, intending, intended, move, moves, moving, moved, ordain, ordains, ordaining, ordained, order, orders, ordering, ordered, pledge, pledges, pledging, pledged, pray, prays, praying, prayed, prefer, prefers, preferring, preferred, pronounce, pronounces, pronouncing, pronounced, propose, proposes, proposing, proposed, recommend, recommends, recommending, recommended, request, requests, requesting, requested, require, requires, requiring, required, resolve, resolves, resolving, resolved, rule, rules, ruling, ruled, stipulate, stipulates, stipulating, stipulated, suggest, suggests, suggesting, suggested, urge, urges, urging, urged, vote, votes, voting, voted*

Appendix 2. The results of the pairwise comparison of formality markers for DAC and non-DAC.

Variable	p-value	Vargha and Delaney A
ffp1_n	< 2.2e-16	0.4260696 (small)
ssp2_n	< 2.2e-16	0.4807493 (negligible)

imp_n	< 2.2e-16	0.4716099 (negligible)
suav_n	0.0003669	0.4942478 (negligible)
whq_n	< 2.2e-16	0.4647596 (negligible)
emph_n	0.2743	0.5019956 (negligible)
hdg_n	0.005605	0.496604 (negligible)
conj_n	9.94E-10	0.4928586 (negligible)
nomz_n	< 2.2e-16	0.4693828 (negligible)
finv_n	< 2.2e-16	0.4841676 (negligible)
func_n	< 2.2e-16	0.4770741 (negligible)
AWL	0.1599	0.5018279 (negligible)
ASL	< 2.2e-16	0.5289643 (negligible)
TTR	< 2.2e-16	0.5542589 (negligible)
lexdens	< 2.2e-16	0.5284346 (negligible)

References

Baker, Paul, Gabrielatos, Costas, Khosravinik, Majid, Krzyzanowski, Michal, McEnery, Tony and Ruth Wodak. 2008. “A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK Press.” *Discourse and Society 19(3):* 273-306.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. "quanteda: An R package for the quantitative analysis of textual data". *Journal of Open Source Software* 3(30): 774. https://doi.org/10.21105/joss.00774.

Benwell, Bethan, and Elizabeth Stokoe. 2006. *Discourse and Identity.* Edinburgh: Edinburgh University Press.

Biber, Doug. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad, Viviana Cortes. 2004. "If you look at ... Lexical bundles in university teaching and textbooks." Applied Linguistics 25(3): 371-405.

Blaette, Andreas, and Christoph Leonhard. 2020. "polmineR: Verbs and Nouns for Corpus Analysis." *The Comprehensive R Archive Network.* https://cran.r-project.org/web/packages/polmineR/

Blom, Jonas, Nygaard, Hansen, and Kenneth Reinecke. 2014. "Narrative and rhetoric strategies in commercial click journalism." Presentation, *JSS-ECREA conference* "Journalism in Transition: Crisis or Opportunity". Thessaloniki, 28–29 March 2014.

Boyland, Emma J., Joanne A. Harrold, and Tim C. Kirkham. 2011. "The extent of food advertising to children on UK television in 2008." *International Journal of Pediatric Obesity* 6(5-6): 455–461.

Breeze, Ruth. 2013. "Lexical bundles across four legal genres." *International Journal of Corpus Linguistics* 18(2): 229–253.

Carter, Ronald, and Michael McCarthy. 2006. *Cambridge Grammar of English: A Comprehensive Guide; Spoken and Written English Grammar and Usage.* Cambridge: Cambridge University Press.

Chafe, Wallace L. 1982. "Integration and involvement in speaking, writing, and oral literature." In *Spoken and Written Language: Exploring Orality and Literacy*, ed. by Deborah Tannen, 35–53. Norwood, NJ: Ablex.

Cockroft, Robert and Cockroft, Susan. 2005. *Persuading people: An introduction to rhetoric*. Houndmills: Palgrave Macmillan

Dedaić, Mirjana N. 2006. "Political speeches and persuasive argumentation". In *The Encyclopedia of Language and Linguistics*, ed. by Keith Brown, 700–707. Oxford: Elsevier.

Durant, Alan, and Marina Lambrou. 2009. *Language and Media: A Resource Book for Students*. London: Routledge.

Durant, Alan, and Janny Leung. 2016. *Language and Law: A Resource Book for Students*. London: Routledge.

Egbert, Jesse, and Doug Biber. 2019. "Incorporating text dispersion into keyword analyses." *Corpora* 14(1): 77–104.

Evert, Stefan, and Andrew Hardie. 2011. "Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium." *Proceedings of the Corpus Linguistics 2011 Conference*, University of Birmingham, UK: 1–21.

Fairclough, Norman. 2001. *Language and Power*. Harlow: Longman.

Findley, Jessica D., and Bruce D. Sales. 2012. *The Science of Attorney Advocacy: How Courtroom Behavior Affects Jury Decision Making.* Washington, DC: American Psychological Association.

Flowerdew, Lynn. 2000. "Using a genre-based framework to teach organizational structure in academic writing." *ELT Journal* 54(4): 369–378.

Gabrielatos, Costas, and Anna Marchi. 2011. "Keyness: matching metrics to definitions." *Corpus Linguistics in the South: Theoretical-methodological Challenges in Corpus Approaches to Discourse Studies and Some Ways of Addressing them*. University of Portsmouth. http://eprints.lancs.ac.uk/51449.

Gerlitz, Carolin, and Anne Helmond. 2013. "The like economy: Social buttons and the data-intensive web." *New Media & Society* 15(8): 1348–1365. https://doi.org/10.1177/1461444812472322

Gilbert, Nigel and Michael Mulkay. 1984. *Opening Pandora's Box: A Sociological Analysis of Scientists' Discourse.* Cambridge: Cambridge University Press.

Giménez-Moreno, Rosa. 2011. "Register variation in electronic business correspondence." *International Journal of English Studies* 11(1): 15-34.

Glinert, Lewis. 2005. "TV commercials for prescription drugs: A discourse analytic perspective." *Research in Social and Administrative Pharmacy 1(2):* 158–184.

Goffman, Erving. 1956. *The Presentation of Self in Everyday Life.* New York: Doubleday.

Greaves, Chris, and Martin Warren. 2010. "What can a corpus tell us about multi-word units?" In *The Routledge Handbook of Corpus Linguistics,* ed. by Anne O'Keeffe, and Michael McCarthy, 240–254. London: Routledge.

Halliday, Michael A. K. 1978. *Language as Social Semiotic: The Social Interpretation of Language and Meaning.* London: Edward Arnold.

Hidey, Christopher, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathleen McKeown. 2017. "Analyzing the semantic types of claims and premises in an online persuasive forum." *Proceedings of the 4th Workshop on Argument Mining:* 11–21. https://doi.org/10.18653/v1/W17-5102

Honnibal, Matthew, and Ines Montani. 2020. "Industrial-strength natural language processing in python." *spaCy.* https://spacy.io.

Janoschka, Anja. 2004. *Web advertising: new forms of communication on the Internet.* Amsterdam: John Benjamins.

Jhaver, Shagun, Pranil Vora, and Amy Bruckman. 2017. "Designing for civil conversations: lessons learned from ChangeMyView." *GVU Technical Report.* Georgia Institute of Technology.

Koteyko, Nelya. 2009. "'I am a very happy, lucky lady, and I am full of Vitality!' Analysis of promotional strategies on the websites of probiotic yoghurt producers." *Critical Discourse Studies* 6(2): 111–125.

Kreis, Ramona. 2017. "The 'tweet politics' of President Trump." *Journal of Language and Politics* 16(4): 607–618.

Kuiken, Jeffrey, Anne Schuth, Martijn Spitters, and Maarten Marx. 2017. "Effective headlines of newspaper articles in a digital environment." *Digital Journalism 5(10)*: 1300-1314.

Labrador, Belén, Noelia Ramón, Héctor Alaiz-Moretón, and Hugo Sanjurio-González. 2014. "Rhetorical structure and persuasive language in the subgenre of online advertisements." *English for Specific Purposes* 34(1): 38–47.

Lakoff, Robin. 1982. "Persuasive Discourse and Ordinary Conversation, with Examples from Advertising." In *Analyzing Discourse: Text and Talk*, ed. by Deborah Tannen, 25–42. Washington DC: Georgetown University Press.

Langlotz, Andreas and Locher, Miriam A. 2013. "Ways of communicating emotional stance in online disagreements." *Journal of Pragmatics* 44(12): 1591-1606.

Liu, Dilin, and Lei Lei. 2018. "The appeal to political sentiment: An analysis of Donald Trump's and Hillary Clinton's speech themes and discourse strategies in the 2016 US presidential election." *Discourse, Contect & Media* 25: 143–152.

Locher, Miriam A. 2013. "Relational work and interpersonal pragmatics." *Journal of Pragmatics* 58: 145–149.

Locher, Miriam A., and Richard J. Watts. 2008. "Relational work and impoliteness: negotiating norms of linguistic behaviour." In *Impoliteness in Language: Studies on its Interplay with Power in Theory and Practice*, ed. by Derek Bousfield, and Miriam A. Locher, 77–99. Berlin: Mouton de Gruyter.

Nini, Andrea. 2019. "The Multi-Dimensional Analysis Tagger." In *Multi-Dimensional Analysis: Research Methods and Current Issues*, ed. by Tony Berber Sardinha, and Marcia Veirano Pinto, 67–94. New York: Bloomsbury Academic.

Ng, Sikh and James Bradac. 1993. *Power in Language. Verbal Communication and Social Influence*. Newbury Park: SAGE Publications.

Ochs, Elinor. 1979. "Planned and unplanned discourse." In *Discourse and Syntax*, ed. by Givon Talmy, 51-80. New York: Academic Press.

Popova, Ksenia. 2018. "Persuasion strategy in online social advertising." *Training Language and Culture* 2(2): 55–65. https://doi.org/10.29366/2018tlc.2.2.4.

Priniski, John, and Zachary Horne. 2018. "Attitude change on Reddit's Change My View." *Proceedings of the 40th Annual Meeting of the Cognitive Science Society* 40: 2276–2281.

Quam, Justin, and Mariana Ryshina-Pankova. 2016. "'Let me tell you…': Audience engagement strategies in the campaign speeches of Trump, Clinton, and Sanders." *Russian Journal of Linguistics* 20(4): 140–160.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

R Core Team. 2020. "R: A language and environment for statistical computing." *The R Project for Statistical Computing.* Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Rudolf Von Rohr, Maire-Therese. 2018. *Persuasion in Smoking Cessation Online: An Interpersonal Pragmatics Perspective*. PhD Dissertation, University of Basel. Freiburg: NIHIN.

Segal, Judy Z. 2005. *Health and the rhetoric of medicine*. Carbondale, IL: Southern Illinois University.

Senft, Theresa M. 2013. "Microcelebrity and the branded self." In *A Companion to New Media Dynamics*, ed. by John Hartley, Jean Burgess, and Axel Bruns, 346–354. Hoboken: Blackwell.

Simpson, Paul. 2001. "'Reason' and 'tickle' as pragmatic constructs in the discourse of advertising." *Journal of Pragmatics* 33(4): 589–607.

Simpson, Paul, and Andrea Mayr. 2010. *Language and Power. A Resource Book for Students*. London, New York: Routledge

Stab, Christian, and Iryna Gurevych. 2014. "Identifying argumentative discourse structures in persuasive essays." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*: 46–56.

Swales, John. 1990. *Genre Analysis.* Cambridge: Cambridge University Press.

Tan, Chenhao, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 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

Tannen, Deborah. 1982. "Oral and Literate Strategies in Spoken and Written Narratives." *Language* 58: 11-21.

Tannen, Deborah. 1985. "Relative focus on involvement in oral and written discourse." In *Literacy, Language, and Learning: The Nature and Consequences of Reading and Writing*, ed. by David R. Olson, Nancy Torrance, and Angela Hildyard, 124–147. Cambridge: Cambridge University Press.

Thurnherr, Franziska, Rudolf von Rohr, Marie-Thérèse and Locher, Miriam A. 2016. "The functions of narrative passages in three written online health contexts." *Open Linguistics* 2(1): 450-470.

van Dijk, Teun. 2008. *Discourse and power.* Basingstoke: Palgrave Macmillan.

Vestergaard, Torben, and Kim Schroder. 1985. *The Language of Advertising*. Oxford: Blackwell.

Wei, Zhongyu, Yang Liu, and Yi Li. 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.

Wright, Kevin. 2015. "Computer-mediated support for health outcomes: psychological influences on support processes." In *The handbook of the psychology of communication technology*, ed by. S. S. Sundar, 488-506. Chichester: Wiley-Blackwell.

Yaqub, Ussama, Soon Ae Chun, Vijayalakshmi Atluri, and Jaideep Vaidya. 2017. "Analysis of political discourse on twitter in the context of the 2016 US presidential elections." *Government Information Quarterly* 34(4): 613–626. https://doi.org/10.1016/j.giq.2017.11.001