Are online news comments like face-to-face conversation? A multi-dimensional analysis of an emerging register

Katharina Ehret[1]* and Maite Taboada
Discourse Processing Lab, Department of Linguistics, Simon Fraser University
kehret@sfu.ca, mtaboada@sfu.ca

Abstract

This article focuses on the question of whether online news comments are like face-to-face conversation or not. It is a widespread view that online comments are like “dialogue”, with comments often being referred to as “conversations”. These assumptions, however, lack empirical back-up. In order to answer this question, we systematically explore register-relevant properties of online news comments using multi-dimensional analysis (MDA) techniques. Specifically, we apply MDA to establish what online comments are like by describing their linguistic features and comparing them to traditional registers (e.g. face-to-face conversation, academic writing). Thus, we tap the *SFU Opinion and Comments Corpus* and the Canadian component of the *International Corpus of English*.

We show that online comments are not like spontaneous conversation but rather closer to opinion articles or exams, and clearly constitute a written register. Furthermore, they should be described as instances of argumentative evaluative language.

1 Introduction

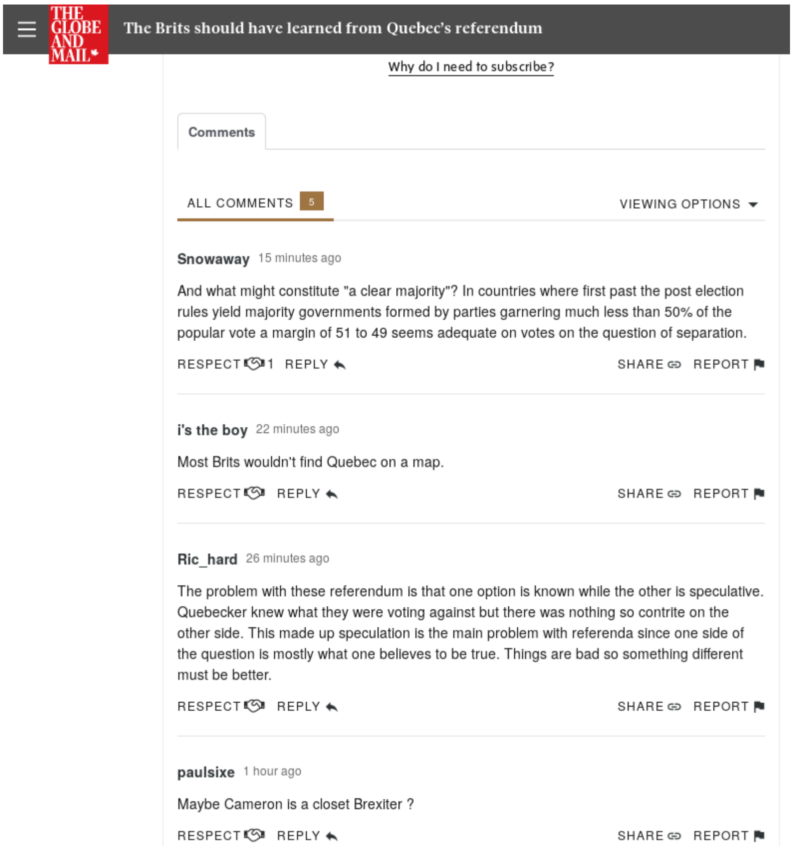
Online news comments and, more generally, social media language has become an increasingly popular topic among researchers from various disciplines (e.g. Herring 1996b; Mehler et al. 2010; Biber & Egbert 2018; Demata et al. 2018). Our interest in online news comments was sparked by the common assumption—while lacking empirical backup—that online comments are like face-to-face conversation. In Figure 1, for example, the thread structure supports replies and comments to the replies, seemingly engaging in a sequence of turns, akin to those in conversation. As a matter of fact, many journalists and editors label online comments as a “dialogue” (McGuire 2015) or “online conversations” (Woollaston 2013); some researchers refer to comments as conversations (Godes & Mayzlin 2004; North 2007). Are online comments really like face-to-face conversation? This, in a nutshell, is the question addressed in the present article.

In this spirit, we adopt a text-linguistic approach (Biber 1988) in order to investigate the register-relevant structural properties of online news comments in comparison to face-to-face conversation and other traditional registers (such as academic writing or broadcast talks). The corpus database consists of opinion articles and comment threads from the *Simon Fraser Opinion and Comments Corpus*, which were collected from the Canadian English-language daily *The Globe and Mail*, as well as the Canadian component of the *International Corpus of English*, to keep the register comparisons within the Canadian context. We thus analyse 25 registers covering a wide range of both spoken and written domains. On a methodological plane, we apply multi-dimensional analysis techniques (Biber 1988) based on a comprehensive set of

[1] * Corresponding author.

To appear in *Register Studies*. Current version: November 29, 2019.

lexico-grammatical features to determine the major dimensions of variation in our dataset, and to describe and locate online comments along the emerging dimensions.



The Brits should have learned from Quebec's referendum

Snowaway 15 minutes ago

And what might constitute "a clear majority"? In countries where first past the post election rules yield majority governments formed by parties garnering much less than 50% of the popular vote a margin of 51 to 49 seems adequate on votes on the question of separation.

i's the boy 22 minutes ago

Most Brits wouldn't find Quebec on a map.

Ric_hard 26 minutes ago

The problem with these referendum is that one option is known while the other is speculative. Quebecker knew what they were voting against but there was nothing so contrite on the other side. This made up speculation is the main problem with referenda since one side of the question is mostly what one believes to be true. Things are bad so something different must be better.

paulsixe 1 hour ago

Maybe Cameron is a closet Brexiter ?

Figure 1: News comment thread from *The Globe and Mail* Opinion section. https://www.theglobeandmail.com/opinion/, retrieved March 18, 2018.

The emerging six dimensions are somewhat similar to the six major dimensions of register variation in English established by Biber (1988). In particular, Dimension 1 strongly resembles Biber’s “Involved vs. informational production” and distinguishes typical written, informational and abstract language on one end of the continuum (e.g. academic writing) from typical spoken, involved language (e.g. conversation) on the other end of the continuum. Furthermore, Dimension 2, which we label “Overt expression of opinion”, comprises features representative of evaluative language (e.g. opinion articles), and is a crucial dimension for describing the structural properties of online news comments. In regard to whether or not online comments are like face-to-face conversation, the results show that online comments are not like spontaneous spoken conversation or dialogue after all. Rather, online comments are more like opinion articles, social letters or exams, and are clearly positioned among the written registers on Dimension 1. More generally, our findings contribute to theorising in usage-based linguistics (Diessel 2017), specifically to usage-based views on language as a complex adaptive system (Beckner et al. 2009), by showing how the situational context and medium of transmission are reflected in the structure of language. Moreover, the unique characteristics of social media language invite a re-thinking of the traditional two-mode classification of language into written and spoken, adding “online language” as a third category in present-day communication.

The terminology used in this article follows Biber and Conrad (2009) and deviates from the terminology used in Biber (1988) in that we use the term “register” rather than “genre”. A register is thus defined as a text variety whose linguistic structure is determined through the situation of usage, i.e. the context of production (Biber & Conrad 2009: 2, 6), rather than the stylistic choice of the language user. However, when referring to “style”, we deviate from this framework and mean style as in “manner” or “kind”: “A kind, sort, or type, as determined by manner of composition or construction, or by outward appearance” (Oxford English Dictionary, http://www.oed.com/).

This article is structured as follows. Section 2 gives a brief overview of research on social media, online comments and register analysis on the web. Section 3 describes the database and Section 4 introduces multi-dimensional analysis. In Section 5 the variational dimensions are described and interpreted. In Section 6 a discussion of the findings in light of the research question is provided. Section 7 offers some concluding remarks.

2 Social media language, online news comments and register analysis

Online registers are typically analysed from either the perspective of discourse analysis (e.g. Herring 1996a; Mehler et al. 2010; Benamara et al. 2018; Demata et al. 2018), or in the framework of text linguistics and register analysis (Biber & Egbert 2016, 2018). In essence, the aim of both lines of research is to describe the properties of social media language, and to categorise the vast amount of available online registers.

In this article, we adopt the analytical framework of register analysis pioneered by Douglas Biber (1988). The classic approach to register analysis is the multi-dimensional approach which is based on the quantitative analysis of a large set of linguistic features combined with the qualitative analysis of their functional-communicative properties and the situational context of the registers. Biber (1988) established six major dimensions along which English registers typically vary: involved vs. informational, narrative vs. non-narrative, explicit vs. situation-dependent reference, overt expression of persuasion, abstract vs. non-abstract information, and online informational elaboration (Biber 1988). Originally concerned with the analysis and description of “traditional” text types such as e.g. academic writing or face-to-face conversation (Biber & Finegan 1989, 1994), multi-dimensional analysis has been applied in recent studies to the description and categorisation of online registers (Biber & Egbert 2016, 2018), and the analysis of language in specific web domains such as Twitter or blogs (Clarke & Grieve 2019, 2017; Daems et al. 2013; Grieve et al. 2010). Interestingly, multi-dimensional analysis has so far not been applied to the study of online news comments.

Online news comments, in general, seem to be somewhat under-researched while other types of online commenting on various social media forums (e.g. Reddit, Twitter) seem much better documented (e.g. Marcoccia 2004; Reagle 2015; Kiesling et al. 2018). Marcoccia (2004), for example, analysed newsgroup messages in terms of their conversation structure and participation framework, characterising them as a form of polylogue, i.e. online communication with multiple levels of dialogue and different levels of participation. In a book-length review of online commenting on social media, ranging from book and restaurant reviews, YouTube messages and Tweets to news comments, Reagle (2015) characterised the “Bottom Half of the Web” (Reagle 2015: 1) as somewhat antagonistic and often meant to manipulate or aggravate, as is the case with trolling. Crucially, Reagle defined online commenting as reactive, yet asynchronous communication (Reagle 2015: 1-2). While this may seem obvious, it is, among other factors, this asynchronous nature that should naturally make online news comments different in linguistic characteristics from face-to-face conversation.

The earliest publication on online news comments (that we are aware of) is Bruce (2010), a discourse analysis of what is termed the "participatory news article", i.e. online news articles and the corresponding reader comments. Based on the analysis of ten articles, the paper describes online news comments as generally being subjective and referring either to the content presented in the article, or another previously posted comment. The general discourse pattern of these comments is described as containing "a statement of a view followed by some justification or argument that supports the comment" (Bruce 2010: 343). More recently, Cambria (2016) has explored the dialogic and interactive nature of news comments and their goal to express agreement or disagreement. As a matter of fact, a recurring feature in online comments in general, is their argumentative, evaluative or opinionated nature. For instance, Kiesling et al. (2018) examine comments on Reddit from the point of view of stance taking, showing that comments can be annotated along three dimensions of stance: affect, investment and alignment, and that each can be identified through specific lexical features. In some cases, the argumentative and opinionated nature of comments turns into face-threatening attacks where news comments criticise authors of articles rather than their writing (Weizman & Dori-Hacohen 2017).

In point of fact, the often derogatory content of online news comments has been the source of work on comment classification and moderation (Diakopoulos 2015; Napoles et al. 2017) as online newspapers and news websites are increasingly faced with the challenge of maintaining "civil dialogue". Thus, other research on online news comments investigates their effect on social media and social behaviour (Nauroth et al. 2015; Rösner et al. 2016), and even much more of the linguistic research on news comments focuses on notions like constructiveness, toxicity, or civility (Coe et al. 2014; Kolhatkar & Taboada 2017a, 2017b). Coe et al. (2014), for instance, investigated the contextual factors for incivility in news comments, and report that uncivil and civil comments alike are supported by evidence and argumentation (Coe et al. 2014: 674). In a similar vein, Kolhatkar and Taboada (2017a) explore constructiveness and toxicity in online news comments in order to build classifiers for the identification of "good" and "bad" comments. Their results show that argumentation features such as adverbials and rhetorical relations are useful indicators for constructiveness and that toxicity is largely independent of constructiveness (Kolhatkar and Taboada 2017a: 15).

From a structural linguistic point of view, however, online news comments still remain largely uncharted territory. It is this gap in the literature, as well as the widespread and commonly held view that online news comments are conversation-like, which triggered the present investigation into the structural linguistic properties of online news comments.

3 Data

In order to analyse the linguistic characteristics of online news comments, we tap the *Simon Fraser Opinion and Comments Corpus* (SOCC, https://github.com/sfu-discourse-lab/SOCC), and, as reference database, the Canadian component of the *International Corpus of English*, version 1 (ICE, http://ice-corpora.net/ice/).

SOCC was specifically compiled for the study of online comments and is to date the largest available collection of online comments. It comprises about 660,000 reader comments and the corresponding 10,000 opinion pieces to which the comments were posted. Thematically, the opinion articles span a wide variety of current national and international topics and were written by 1,628 unique authors. The comments in the corpus originate from more than 34,554 different commenters. The data was harvested from the online version of the Canadian newspaper *The Globe and Mail* during the time period from 2012 to 2016 and is available in the form of three sub-corpora, the article corpus, the comments corpus, and the comments thread corpus which contains concatenated posts by multiple authors preserving the comment's original thread structure (Kolhatkar et al. 2019).

In this article, we utilise the articles and the comment threads corpus as the structure of comment threads is presumably more conversation-like than individual comments. The analysis is furthermore restricted to articles and comment threads with a minimum number of 700 words as a requirement for the subsequent analysis.[1] A random sample of 80 opinion articles and comment threads, respectively, is then generated to approximate the number of texts in the ICE register conversation.

ICE Canada comprises a wide range of spoken and written registers (15 spoken; 17 written) covering both private and public domains. The entire corpus counts 500 texts, which were collected in the 1990s, and amounts to approximately 1 million words in total (Newman & Columbus 2010). The 15 spoken registers were subsumed under 13 macro-registers, for instance, broadcast talks and broadcast interviews were combined in the macro-register broadcast; the 17 written registers were subsumed under 10 macro-registers. For example, the academic micro-registers social sciences, humanities, natural sciences, and technology were merged into the macro-register academic.[2] The final database thus counts 660 texts totalling 1,900,000 words and covering 25 registers (see Table 1).

Table 1. Overview of the database by register, number of texts, number of words, and corpus. Mode of the register (written vs. spoken) is indicated in parentheses.

Register	No. of texts	No. of words	Corpus
Academic (w)	40	96,518	ICE Canada
Broadcast (s)	30	64,019	ICE Canada
Business letters (w)	15	30,785	ICE Canada
Business transactions (s)	10	21,689	ICE Canada
Commentaries (s)	20	41,224	ICE Canada
Conversation (s)	90	187,993	ICE Canada
Cross-examinations (s)	10	22,156	ICE Canada
Demonstrations (s)	10	22,252	ICE Canada
Exams (w)	10	21,200	ICE Canada
Fiction (w)	20	42,582	ICE Canada
Instructional (w)	20	43,548	ICE Canada
Legal presentation (s)	10	21,439	ICE Canada
Lesson (s)	20	42,576	ICE Canada
Non-academic (w)	40	86,896	ICE Canada
Online comments (w)	80	773,062	SOCC
Opinion (w)	80	78,239	SOCC
Parliamentary debate (s)	10	20,703	ICE Canada
Persuasive (w)	10	20,766	ICE Canada
Reportage (w)	20	42,041	ICE Canada
Scripted broadcast (s)	40	84,475	ICE Canada
Scripted non-broadcast (s)	10	21,000	ICE Canada
Social letters (w)	15	31,469	ICE Canada
Student essays (w)	10	21,413	ICE Canada
Telephone (s)	10	22,190	ICE Canada
Unscripted speech (s)	30	66,918	ICE Canada
Total	660	1,927,153	

4 Multi-dimensional analysis

Multi-dimensional analysis (MDA) was first introduced by Douglas Biber in his landmark publication “Variation across speech and writing” (Biber 1988). It has since been a major approach to analysing variation across different text types. Essentially, MDA is a multivariate analysis technique based on the frequency of linguistic features and their co-occurrence patterns in texts, which are interpreted in functional and communicative terms to establish dimensions of textual variation across different texts and registers. In short, MDA is typically used to describe the linguistic properties of different texts/registers. In the present article, multi-dimensional analysis is applied as a diagnostic tool to establish whether or not online news comments are like face-to-face conversation. To this end, we largely follow the methodology outlined in Biber (1988: 71-93) and draw on his well-established catalogue of register-defining features (Biber 1988: 221-245). Unless otherwise indicated, all statistics discussed in this section and other supplementary materials can be downloaded from https://github.com/sfu-discourse-lab/MDA-OnlineComments.

4.1 Features and frequencies

The feature catalogue comprises 67 core features of English covering a wide-range of lexico-grammatical domains including modals, negation, pronouns, tense and aspect markers, and subordination (for an exhaustive list and detailed description see Biber 1988: 73, 221-245; for

POS-tags see Table 4, Appendix A). [3]

Thus, the SOCC-ICE database described in Section 3 is automatically annotated for the 67 features using the open-source *Multidimensional Analysis Tagger*, version 1.3 (MAT, https://sites.google.com/site/multidimensionaltagger/versions), which implements the algorithm described in Biber (1988). The tagger is based on the Stanford part-of-speech tagger (Toutanova et al. 2003), and extends the Stanford tagset with additional tags identifying the features listed in Biber's (1988) feature catalogue (Nini 2014).[4] After tagging, we retrieve the occurrence frequencies of the 67 features by means of a custom-made python script (see https://github.com/sfu-discourse-lab/MDA_project). The observed feature frequencies are normalised per 1000 words to ensure the comparability of the frequencies across differently sized texts. Exempt from normalisation are the two features, type token ratio (TTR) and average word length (AWL): Type token ratio is automatically calculated for the first 400 words in each text (cf. Biber 1988: 75); average word length is calculated as the number of characters per text divided by the number of words per text.

4.2 Factor analysis

The primary statistical tool of multi-dimensional analysis is exploratory factor analysis, a technique for variable reduction. We specifically utilise *maximum likelihood factor analysis*, as provided in the *stats* package in R (R Core Team 2018), and implement it with an oblique *promax rotation* (Biber 1988). All statistics and computing were conducted in R (version 3.5.0, R Core Team 2018).

The suitability of the data for performing factor analysis is tested by conducting *Bartlett's test of sphericity* and the *Kaiser-Meyer-Oelkin measure for sample adequacy* (MSA). Both tests return excellent results (Dziuban & Shirkey 1974: 358-359): Bartlett's test for sphericity returns a *p*-value = 0 and the overall MSA for our dataset is 0.9.

We thus proceed to perform the actual factor analysis. First, eigenvalues for all possible factors are calculated. All eigenvalues equal or greater than 1 are retained (Hair et al. 2014: 105) and visualised in a scree plot (Cattell 1966): The scree plot presented in Figure 2 follows a steep curve and exhibits a first break between Factor 6 and 7 before straightening into a horizontal line, suggesting that six factors should be sufficient (cf. Biber 1988: 82; Hair et al. 2014: 105). Next, we fit a model by stepwise adding a total of 7 factors to our solution in order to corroborate the number of factors suggested by the scree plot of eigenvalues. In other words, we start with a one-factor model, then add a second factor, a third factor and so on. For each added factor the variance explained by the individual factor and the total amount of variance explained by the model are examined, and weighed against the interpretability of the extracted factors. Using a conservative cut-off of |0.3| to determine statistically significant factor loadings, factors are considered interpretable in a meaningful linguistic way if they comprise at least five salient loadings (Biber 1988: 87). After considering all of the above described criteria, we settle on the 6-factor solution (see Table 5, Appendix B) as initially suggested by the scree plot which explains overall 41% of shared variance in the dataset.

Finally, factor scores for each text in the dataset are computed using the factor score option "regression" in R. The factor scores basically position each text on a given factor: The higher a given text loads on a factor, the more representative this text is of the underlying linguistic dimension of the factor (Biber 1988: 93). For example, the text of the file labelled comments2012.4527764.txt has a factor score of 0.26 on Factor 1 but a factor score of 1.37 on

Factor 2. This means that comments2012.4527764.txt contains more features that load high on Factor 2 than on Factor 1 and is therefore more representative of Factor 2. Subsequently, the factor scores are subjected to linear regression analysis testing the statistically significant differences between registers on each factor. All *p*-values (Table 2) are below the generally accepted threshold of significance ($p < 0.05$), so that on all six factors, register is a significant predictor for differences between factor scores. The amount of variance explained by the factors, however, fluctuates. For instance, differences in factor scores on Factor 1 account for 83% of the variance between registers indicating that Factor 1 is a very powerful register-distinguishing factor. In contrast, Factors 2 only accounts for 46% of register-specific variance.

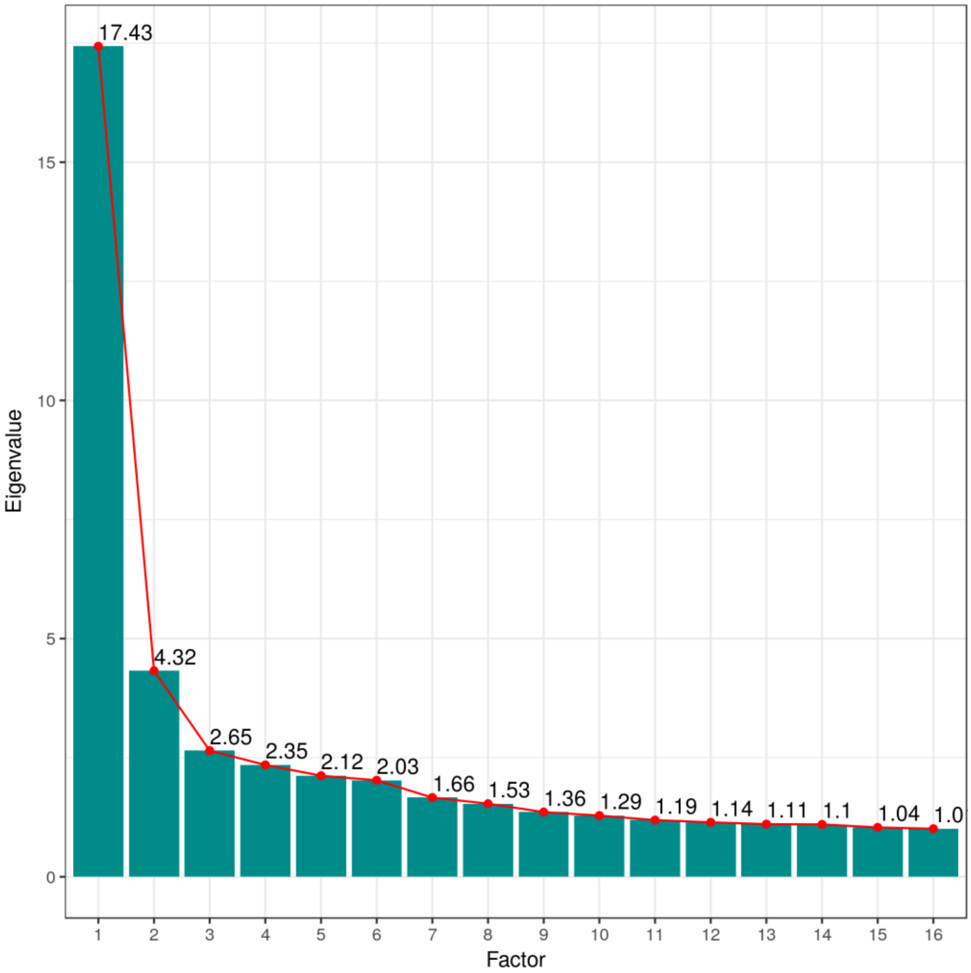


Figure 2: Scree plot of factors (on the *x*-axis) with eigenvalues equal or greater than 1 (on the *y*-axis).

Table 2: *F*-statistic, *p*-value and *R*-squared value (in percent) of linear regression analysis with factor scores as dependent variable and register as predictor variable.

Factor	***F*-statistic**	***p*-value**	***R*-squared in %**
Factor 1	131.2	< 2.2e-16	83%
Factor 2	24.34	< 2.2e-16	46%
Factor 3	11.35	< 2.2e-16	27%
Factor 4	21.52	< 2.2e-16	43%
Factor 5	20.41	< 2.2e-16	41%
Factor 6	11.86	< 2.2e-16	28%

In analogy to factor scores for individual texts, scores can be calculated which position individual registers on a given factor with respect to the underlying linguistic dimension (Biber 1988: 121; for illustration and discussion see Section 6). In this article, these scores are called *mean factor scores* as they are obtained by taking the arithmetic mean of all individual texts belonging to one register. Thus, mean factor scores as well as their standard deviation are computed for each register in the database.

5 Variational dimensions

In this section the factors extracted through exploratory factor analysis are interpreted as *variational dimensions* based on the shared functional-communicative properties of their most salient features, taking the complementary distribution of feature patterns into account (Biber 1988: 91-92). In order to determine the functional properties of the features, we mainly, but not exclusively, refer to Biber's detailed descriptions and functional definitions of the features outlined in Biber (1988: 221-245). In this interpretation, features with loadings greater or equal to |0.35| are given greater importance. Furthermore, the interpretation of the factors as dimensions accounts for cross loadings if these features exhibit the same polarity, i.e. features that load on more than one factor and exhibit the same polarity are given more importance on the factor where they load highest. Such a purely qualitative interpretation of the factors as variational dimensions is, of course, not unproblematic. The interpretation of the factors is therefore confirmed by examining the distribution of ICE and SOCC registers across the six factors (for a detailed discussion see Section 6). A summary of the six factors is given in Table 3.

Table 3: Summary of the 6-factor solution including only salient features with loadings ≥ |0.3|. Loadings ≥ |0.35| are italicised. Negative polarity indicates complementary distribution; positive polarity indicates co-occurrence of the features. Parentheses indicate that a given feature has a higher loading of the same polarity on another factor.

Factor 1		**Factor 2**	
Contractions	*0.954*	*Adjectives predicative*	*0.76*
Private verbs	*0.859*	*BE main verb*	*0.621*
THAT deletion	*0.814*	*(Adverbs*	*0.342)*
Present tense	*0.788*	Split auxiliaries	0.319
2nd ps. pronouns	*0.774*	(Emphatics	0.311)
Hedges	*0.772*	Adjectives attributive	0.3
Demonstrative pronouns	*0.723*	*Stranded prepositions*	*-0.377*
Discourse particles	*0.723*	Past participle clauses	-0.309
1st ps. pronouns	*0.714*		
Analytic negation	*0.714*		

Pronoun IT	*0.687*	**Factor 3**	
WH clauses	*0.669*	*Public verbs*	*0.738*
DO pro-verb	*0.643*	*Suasive verbs*	*0.51*
Emphatics	*0.615*	*Modals predictive*	*0.429*
(BE main verb	*0.585)*	*TO infinitives*	*0.378*
Adverbs	*0.507*	(Subordinator condition	0.332)
Stranded prepositions	*0.523*	*Non-phrasal coordination*	*-0.36*
WH questions	*0.456*		
Subordinator condition	*0.418*		
Existential THERE	*0.348*	**Factor 4**	
Subordinator cause	*0.347*	*Demonstratives*	*0.653*
Non-phrasal coordination	*0.332*	*THAT relatives obj.*	*0.467*
Prepositions	*-0.914*	*THAT adjective complements*	*0.351*
Average word length	*-0.894*	*(Existential THERE*	*0.346)*
Adjectives attributive	*-0.793*	THAT verb complements	0.332
Type-token ratio	*-0.787*	Amplifiers	0.324
Nouns	*-0.768*	(Stranded prepositions	0.331)
Nominalisations	*-0.662*	No negative features	
Phrasal coordination	*-0.657*		
Perfect aspect	*-0.559*		
Conjuncts	*-0.461*	**Factor 5**	
BY passive	*-0.451*	*Nominalisations*	*0.493*
Past participle WHIZ deletion	*-0.49*	*Agentless passives*	*0.392*
Agentless passives	*-0.448*	Past participle WHIZ deletion	0.34
Downtoners	*-0.369*	Conjuncts	0.334
THAT relatives subj.	*-0.35*	Pied-piping relatives	0.329
Pied-piping relatives	-0.316	*Place adverbials*	*-0.522*
Split auxiliaries	-0.313	*Time adverbials*	*-0.436*
WH relatives subj.	-0.312		
		Factor 6	
		Past tense	*0.984*
		3rd ps. pronouns	*0.352*
		Present tense	*-0.499*

As visualised in Figure 3 (Appendix C), most of the 67 features weigh on the first factor: In total, Factor 1 counts 32 salient features with loadings ≥|0.35|. As with most of the factors, Factor 1 consists of features with positive and negative loadings indicating a complementary distribution of the features, i.e. features with positive loadings frequently co-occur in the same texts while features with negative loadings do not frequently occur in or are absent from these texts. The top ten positive features are contractions, private verbs, THAT deletion, present tense verbs, second person pronouns, hedges such as *more or less*, demonstrative pronouns, discourse particles such as *anyway*, or *well*, first person pronouns and analytic negation. All of these features are indicators for a colloquial, informal style (contractions, THAT deletion) marked by disconnected speech and hesitations (hedges, discourse particles) typical of spontaneous spoken language where language is planned on-line and careful editing is not possible. First and second pronouns further indicate personal involvement of the language user(s). Other features loading high on this factor which are typical of spontaneous spoken language are the pronoun IT, WH-clauses, DO as proverb, as in Example (1), emphatics, adverbs and stranded prepositions. Most of the top ten negative features, are usually identified as markers of high information density (prepositions, attributive adjectives, nouns, nominalisations) and lexical specificity (type-token-ratio, average word length). Their co-occurrence with passive aspect and conjuncts such as *however, e.g., therefore* (see Biber et al. 1999: 562), indicate a formal style which is, for instance, typical of academic writing, e.g., Example (2). This is congruent with the literature on argumentation which considers linking adverbs as markers of argumentative writing (e.g. Tseronis 2011, van Eemeren et al. 2007). Thus, the complementary distribution of features that are typical of spontaneous spoken and involved language versus features which are typical of highly informative and formal language clearly constitutes a fundamental register-distinguishing dimension in English, namely, the gradual distinction between spoken and written language. As it substantially overlaps with Biber's dimension "Involved vs. informational production" (1988), we label it as Dimension 1 "Involved vs. informational".

(1) [...] I really have to speak to her you know can you get her out of the meeting and they [**did**]$_{\text{proverb do}}$. (conversationS1A-008[5], ICE)

(2) [**Thus**]$_{\text{conjunct}}$ much of the prerogative power of the Crown [**is determined**]$_{\text{BY-passive}}$ exclusively [**by Cabinet]**$_{\text{BY.passive}}$ rather than [**by the monarch]**$_{\text{BY-passive}}$. (academicW2A-019, ICE)

Factor 2 counts eight salient features, four of which have loadings equal or greater than |0.35|. The most characteristic positive features are predicative adjectives and BE as main verb. Biber identifies these two features as reduced surface structures which are common in conversational language (Biber 1988: 228-229). However, the construction *be* + adjective is also known to frequently occur with adjectives that mark attitude such as *bad*, *nice* or *true* (Biber et al. 1999: 437-440). Furthermore, *be* + (adverb) adjective is analysed as a marker of personal stance when occurring with first and second person pronouns as in Example (3a) (Biber & Finegan 1989). It goes without saying that it can also occur with other subjects to mark stance in a more general manner as in Example (3b). In discourse analysis, *be* + adjective occurs as part of various lexico-grammatical patterns, called "frames", which are used to convey evaluation (Hunston 2011; White 2003) such as in (4a-b). Typical canonical frames are e.g. *I feel* + adjective, *It be* + adjective *for/of him/her to*-infinitive (White 2003: 173), but slightly deviating forms taking other subjects such as *He/she be* + adjective (4c) etc. are also common in natural speech.

(3) a. **You** [**are oblivious**]$_{\text{be + adjective}}$. (comments2013.10374471, ICE)

b. **The battle** [**is pointless**]$_{\text{be + adjective}}$. (examsW1A-011, ICE)

(4) a. **It'**[**s true**]$_{\text{be + adjective}}$ **that** federal-provincial practice has meant [...].
(*It be* + adjective *that*, opinionsocc_2015_22660102, SOCC)

b. **It** [**is perfectly reasonable**]$_{\text{be + adjective}}$ **to expect** all health workers [...] to be vaccinated against influenza. (*It be* + adjective *to*-infinitive, opinionsocc_2013_8028220, SOCC)

c. All you have to do is look at the election results to see that **voters [are not**

ready]$_{\text{be + adjective}}$ **to face reality**. (comments2014.19155297, SOCC)

Considering the less prominent positive features, some of which actually load higher on another factor, this interpretation can be confirmed: The systematic co-occurrence of split auxiliaries and adverbs suggests that the main verb is modified, and presumably evaluated. In the same vein, the frequent use of attributive adjectives suggests, firstly, that noun entities are modified and evaluated. In fact, adjectives, be they predicative or attributive, are used as a primary marker for determining subjectivity (including evidentiality) in sentiment analysis (Bruce & Wiebe 1999: 13-14; Taboada et al. 2011: 268). Secondly, it suggests that information is carefully integrated in the texts (Biber 1988: 237). Somewhat contradictory, then, is the presence of emphatics which suggest an involvement of the language user (Biber 1988: 241) and the absence of past participle clauses, a means of structural elaboration in written discourse (Biber 1988: 233). The notable absence of stranded prepositions is difficult to interpret but might indicate that Factor 2 characterises written rather than spoken texts, as stranded prepositions tend to be more frequent in spoken texts (with the exception of fiction and news) (Biber et al. 1999: 105-106). In summary, Factor 2 seems to represent a casual and involved, slightly informational style which is used to express attitude, evaluation and opinion. Factor 2 is thus labelled Dimension 2 “Overt expression of opinion”.

The third factor comprises six positive salient features and only one negative feature. Public verbs, suasive verbs and predictive modal verbs weigh highest on this factor. Public verbs denote public (and official) speech acts such as *announce, declare* and *testify* (for a full list see Quirk et al. 1985: 1180-1181). Suasive verbs express various degrees of persuasion (e.g. *suggest*, *instruct*, *command*, *rule*), concession (e.g. *allow, grant*), and future intentions (e.g. *desire*, *intend*) as well as certain speech acts (e.g. *beg*, *ask*) (for a full list see Quirk et al. 1985: 1182-1183). Generally, predictive modal verbs (*will*, *would*, *shall*) are more frequent in conversation, where they are used to indicate prediction and volition. Finally, TO infinitives are used to add complementary information (Biber 1988: 232). The absence of non-phrasal coordination, or independent clause coordination (Example 5), which is typically used in spontaneously produced language (cf. Biber 1988: 245) cannot be interpreted in a straightforward manner, especially as it is the only salient negative feature on Factor 3. Based on the pattern of positive features, Factor 3 denotes public spoken language with a strong persuasive slant. Dimension 3 is named “Public persuasive presentation”

(5) Ya [**and**]$_{\text{non-phrasal coordination}}$ what the road is like. (conversationS1A-046, ICE)

Factor 4 is composed of seven salient features of which three have loadings equal or greater than |0.35|. There are no negative features. The interpretation of Factor 4 is rather straightforward: Demonstratives used as determiners load highest on this factor and are used to create referential cohesion (Biber 1988: 241) in texts or speech as they reference specific noun entities; demonstratives could also be interpreted as “overt markers” of reference. THAT relatives in object position, THAT adjective complements and THAT verb complements. Examples (6a-c) provide further information, specification or elaborate a statement. We therefore dub Dimension 4 “Descriptive elaboration”.

(6) a. And uh on many of the flights [**that I have taken to various places in the north**]$_{\text{that relative obj.}}$ [...] (parliamentS1B-059, ICE)

b. Now in this case I’m satisfied **that you will all apply yourselves diligently and conscientiously to the task**]$_{\text{that adjective complement}}$ [...] (legal-presentationS2A-

066, ICE)

c. And I don’t think [**that we can afford the luxury of saying “this is too depressing”**]$_{\text{that verb complement}}$ [...]. (scripted-brdcasttS2B-029, ICE)

Overall, Factor 5 comprises seven salient features in total and four features with loadings equal or greater than |0.35|. Factor 5 is the only factor where the negative features are more distinctive than the positive features. The two negative features are place and time adverbials, clear markers of situational and temporal context; their character is also deictic in nature as time and place references are often context-dependent. These features of deictic reference occur in complementary distribution with nominalisations, agentless passives and, to a lesser extent, past participle WHIZ deletions, i.e. past participle clauses used as reduced relative clauses (Example 7) which all indicate formal, abstract and informational language. This interpretation is confirmed by the “secondary” features on this factor: conjuncts are typical of formal language and link sentences, and pied-piping relative clauses provide additional information, as illustrated in Example (8). We thus interpret Factor 5 as Dimension 5 “Abstract-informational vs. deictic reference”.

(7) [...] and suppresses the savage aspect of the individual to conform with [**the values prescribed by society**]$_{\text{past participle WHIZ deletion}}$. (student-essaysW1A-010, ICE)

(8) Scalar quantum field theory, [**in which massive spinless particles interact via another, massive or massless, scalar field]**$_{\text{pied-piping relative}}$ has long been a favorite model [...]. (academicW2A-039, ICE)

Factor 6, which was included in the model to avoid the conflation of underlying linguistic constructs, only counts three features in total. The complementary distribution of past tense and third person pronouns with present tense is reminiscent of the narrative dimension identified in Biber (1988), and, indeed, fiction is one of the registers that weighs high on this factor. Yet, a reliable linguistic interpretation of Factor 6 is not possible, and it will therefore be excluded from the following discussion.

6 Are online news comments like face-to-face conversation?

Before turning to a detailed discussion of the variational dimensions relevant for the categorisation and description of online news comments, let us take a brief look at the general distribution of the analysed registers across the various dimensions. Figure 4 shows the mean factor scores for each register on each of the six dimensions (see Appendix D for Figure 4 and Table 6). The mean factor scores provide information on the position of individual registers on a given dimension. For example, the mean factor score for commentaries, specifically spontaneous spoken sports commentaries, on Dimension 5 is -1.79. Dimension 5 represents abstract-informational language on the positive pole and deictic reference on the negative pole. Commentaries are positioned on the negative pole of Dimension 5, thus their mean factor score is visualised as a dark red bar. What does the mean factor score tell us about the nature of commentaries then? Commentaries are characterised by the frequent use of place and time adverbials because these features load on the negative pole of Dimension 5, and the absence of nominalisations and passives because they load on the positive pole of Dimension 5. On an interpretational plane, the frequent use of deictic markers in sports commentaries is not surprising as commentators have to frequently reference the place or position of objects or people and place actions within a certain time frame, e.g. when describing an ice hockey match as exemplified in (9).

> (9) Pullishy'll pick it up at the centre ice stripe. Plays it back for Goodkey. **Ahead** to Aaron Zarowny and he'll clear it in **across** the line. Bears will change. They send Esposito Degner and Cam Sherben out [...]. They're **now** two for four on the powerplay. This is their fifth **now** [...]. (commentaries2SA-005.txt, ICE)

Generally, most of the spoken registers weigh positively on Dimension 1, while all written registers exhibit—to varying degrees—negative weights on the first dimension. This is congruent with the interpretation of Dimension 1 as distinguishing between informational production on the one end of the continuum and involved production on the other end of the continuum. On the extreme ends of this continuum, prototypical written (academic writing) and prototypical spoken registers (conversation, telephone conversation) are located, highlighting the fundamentally distinctive characteristic of Dimension 1. That said, some of the spoken registers, most notably, scripted non-broadcast, scripted broadcast, reportage, and parliamentary debate have negative weights on Dimension 1, i.e. these spoken registers have an informational rather than involved focus and tend to be prepared/edited before orally delivered. In contrast, the written register social letters is positioned in the center of the two poles and combines clearly informational features with markers of personal involvement.[6]

Although the other dimensions are not as fundamentally distinctive as Dimension 1, they clearly add to the multifaceted description of these registers. Dimension 2 represents registers that overtly convey opinion and subjectivity, thus online comments, opinion articles, exams (i.e. mainly essays on literature) and social letters have high positive scores on this dimension while, for instance, commentaries, demonstrations, legal presentation and scripted broadcast have high negative scores indicating that evaluation and attitudes are not as overtly conveyed in these registers. Moreover, the patterning of these registers (i.e. written registers on the positive pole vs. spoken registers on the negative pole) is in tandem with the feature distribution on Dimension 2 which suggests that it characterises written rather than spoken language (see Section 5 for details). On Dimension 3 “Public persuasive presentation” only two registers have high positive loadings: reportage and legal presentation. These two registers utilise a comparatively large number of public and suasive verbs as well as expressions of prediction and volition. This dovetails with the situational context of the registers: both reportage and legal presentations give account of events and their circumstances and express (future) intentions or, in the case of legal presentation, varying degrees of persuasion, as shown in Example (10). In contrast, the registers with relatively high negative loadings—academic writing, commentaries, exams and lessons—are marked by a significant absence of these verbs. As can be seen from the emergence of the two distinct Dimensions 2 and 3, there is a crucial difference between the linguistic expression of opinion and subjectivity on the one hand, and persuasion on the other hand: opinion and persuasion are expressed through different linguistic devices (see also Section 5 for dimension-characteristic features) so that persuasion does not necessarily convey opinion overtly. Dimension 4 denotes discourse characterised by “Descriptive elaboration” such as is common in parliamentary debates, legal presentations and cross-examinations but also in student essays, scripted non-broadcast, demonstrations and exams where detail matters, and reference needs to be specific and marked overtly. As mentioned in the introductory paragraph of this section, Dimension 5 characterises abstract-informational language on the positive pole and deictic reference on the negative pole. Apart from commentaries, the registers fiction and social letters are positioned on the negative pole of this Dimension indicating that the use of deictic markers of place and time is very common in these registers. In academic writing, student essays, cross-examinations and instructional writing, on the contrary, this type of deictic reference is notably absent while, at the same time, passives and nominalisation characterise a rather formal and abstract discourse. In summary, the distribution of the analysed registers across the five Dimensions confirms the qualitative interpretation of the factors provided in Section 5.[7]

(10) He'[**d**]$_{\text{predictive modal}}$ uhm he'[**d**]$_{\text{predictive modal}}$ [**stipulate**]$_{\text{suasive}}$ what to put in there. [...] And was he incoherent when he [**instructed**]$_{\text{suasive}}$ you on that. Mhh I couldn't [**say**]$_{\text{public}}$ Sir. (legalS1B-065.txt, ICE)

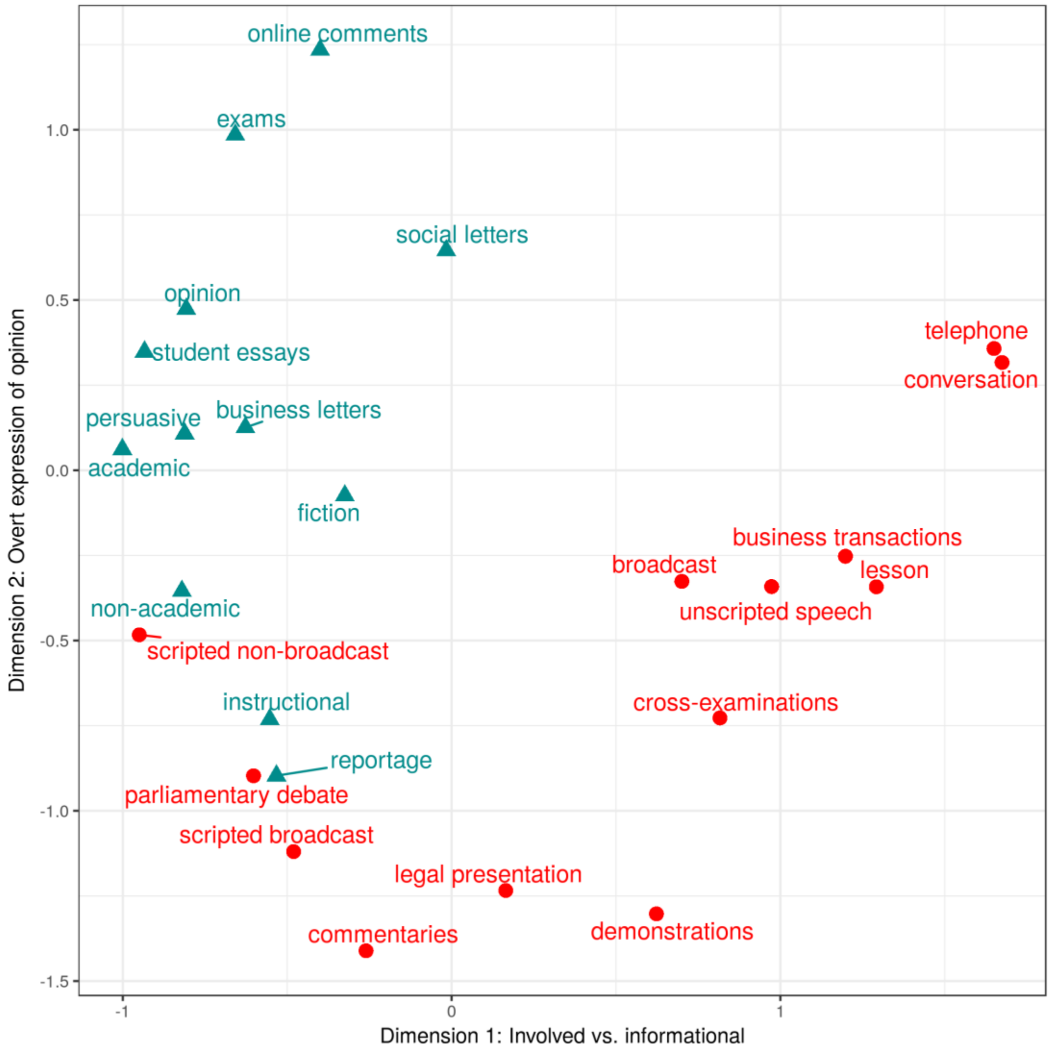


Figure 5: Scatterplot of the SOCC and ICE registers along Dimension 1 (abscissa; positive values index involved language; negative values index informational language) and Dimension 2 (ordinate). Green triangles represent written registers; red dots represent spoken registers.

Let us now turn to the key question of this article, namely, whether or not online news comments are like face-to-face conversation. In order to answer this question from a structural linguistic perspective, the position of online news comments and face-to-face conversation on Dimension

1 and 2, which are most distinctive for these two registers, will be compared. Figure 5 plots the analysed registers along the relevant dimensions “Involved vs. informational” and “Overt expression of opinion”. On Dimension 1 online comments and conversation are positioned at opposite ends (expressed by mean factor scores of -0.4 and 1.67, respectively): Online comments are much closer to academic writing (mean factor score -1.00) than to face-to-face conversation. Thus, online comments are rather characterised by features typical of informational, carefully edited, mostly (but not exclusively) written registers[8], while conversation is characterised by features marking involved discourse produced under the constraints of spontaneous production. Specifically, conversation is predominantly characterised by reduced surface structures such as contractions, hedges and discourse particles which are also used as time buying devices in on-line production, as well as the frequent use of private verbs and first/second pronouns which indicate involvement. Needless to say, some of these features do occur in online comments, however, they are not pervasive and hence not characteristic of online comments. All in all, online news comments need to be placed close to the written end of the spoken-written continuum, and are arguably not like face-to-face conversation which is at the spoken end of this continuum.

The question remains, then, what online comments are like vis-à-vis traditional English registers. In terms of their structural properties, online news comments closely resemble fiction (mean factor score -0.32) and scripted broadcast (mean factor score -0.48) on Dimension 1, and are very similar to exams (mean factor score 0.98), mainly discussions of literary works, and to a lesser extent social letters (mean factor score 0.65) and opinion articles (mean factor score 0.47) on Dimension 2.

Online comments contain comparatively many markers of information-density and carefully edited language such as prepositions, attributive adjectives, nouns and nominalisations. Their average word length and type-token-ratio are also high, indicating that commenters make careful and specific lexical choices. The presence of conjuncts can be interpreted—apart from being markers of a formal style of writing—as marking cohesive text (Halliday & Hasan 1976). As a matter of fact, conjuncts have been deployed as features of argumentation (Moens et al. 2007) and of constructiveness in the literature on comment moderation (Kolhatkar & Taboada 2017a, 2017b). Comments could be identified as constructive when, *inter alia,* commenters supported their arguments with examples, evidence or personal experiences which are often signalled by conjuncts (11) (Kolhatkar & Taboada 2017b). All of these features mark online comments as a register closer to written rather than spoken language.

11. a. I have been married and hated it [**although**]$_{\text{conjunct}}$ I will admit it would be nice to have some sort of a relationship. [**However**]$_{\text{conjunct}}$, I know so many people that hate their lives and take it out on their partner and I have no time for that. (comments2013.7536781, SOCC)

    b. [...] with our responsible party system the party, their voting and their positions are quite visible. In the U.S., [**for example**]$_{\text{conjunct}}$, with free-range voting you have your legislators bought and sold by special interests all over the place spouting populist rhetoric [...]. (comments2013.10749814, SOCC)

Furthermore, online comments exhibit a rather unique feature pattern which emerged as Dimension 2 “Overt expression of opinion” in our dataset. The principal features are predicative adjectives and BE as main verb which together occur in expressions of stance, evaluation and opinion. In online comments, these features co-occur with adverbs, split auxiliaries and attributive adjectives further stressing the evaluative character of online comments. Moreover, online comments combine an informational-involved style manifested by the presence of emphatics and the absence of past participle clauses on the one hand, and the presence of attributive adjectives—which can also be interpreted as markers of information integration—on the other hand. These findings dovetail with the situational context of online comments: the comments section of news websites is meant as a platform for readers to comment and exchange opinions on a wide variety of news content. In addition, the comments analysed in this article were posted in response to opinion articles, which, by virtue of their subjective nature, are due to elucidate an exchange of strongly evaluative replies. The informational-involved characteristic of online comments is certainly also due to the situational features of online commenting. Commenters are not under the pressure of having to reply spontaneously, or within a certain time frame as online commenting is not a real-time, synchronous means of communication. Thus, online comments can be carefully, or at least somewhat edited before being posted. At the same time, commenters are not constrained by formal linguistic requirements in the same way journalists are, and can use casual and emphatic language to express their personal opinions.

Last but not least, the similarity of comments to exams and social letters is not as surprising as it might seem at first glance. On the one hand, interpretations of literary works require the writer to evaluate the actions and characters of the protagonists. It goes without saying that such an interpretation further requires a coherent argumentation (12). Social letters, on the other hand, may contain subjective content because their writer comments on, gives their opinion on or evaluates, for instance, an event, a person (13a), or, as it turns out, the weather conditions (13b). In addition to these functional similarities, online comments, exams and social letters are produced in a similar situational context, i.e. the writer and their audience/the addressee share neither time nor place which is likely to impact on the linguistic structures used in these registers (cf. Collot & Belmore 1996).

In short, online comments can be best described as a type of informally written, argumentative evaluative language.

(12) a. [**However**]$_{\text{conjunct}}$, the ambivalent feelings remain in the tone despite these hints of future happiness. The butterfly has no future, has no story and [**therefore**]$_{\text{conjunct}}$ has no voice [...]. (examsW1A-015, ICE)

b. Example from exams with BEMA plus PRED

(13) a. **It** [**is larger**]$_{\text{be+adjective}}$ than the one she had last time she was here [...]. But **Rose** [**is eternally cheerful**]$_{\text{be+adjective}}$ and carries on as usual. (letterssW1B-002, ICE)

b. [**It's been hard**]$_{\text{be+adjective}}$ deciding what to take to wear since [**it's been so cold**]$_{\text{be+adjective}}$ here. (letterssW1B-005, ICE)

7 Concluding remarks

In this article, we presented the first-ever systematic description and multi-dimensional analysis of the language of online news comments as sampled in the *SFU Opinion and Comments Corpus*. Conducting a multi-dimensional analysis of online news comments vis-à-vis face-to-face conversation and other traditional registers in the Canadian component of the *International*

Corpus of English, we provide empirical evidence against the common view that online news comments are like face-to-face conversation. Despite their dialogic appearance, the analysis of their structural linguistic features shows that they must clearly be categorised as a written register. Online news comments are characterised by a distinctive combination of informational and involved features, and prominently contain feature patterns that are typical of argumentative and evaluative language. On the basis of these linguistic features, we describe the language of online news comments as an informally written, argumentative evaluative language, and argue that online news comments should be regarded as their own register.

Importantly, the structural description of online news comments presented in this article is in tandem with previous contributions about online news comments. In particular, online news comments have been repeatedly characterised as argumentative (Coe et al. 2014; Kolhatkar & Taboada 2017a), subjective and evaluative (Bruce 2010; Weizman & Dori-Hacohen 2017), features which seem to occur in other types of online comments as well (cf. Reagle 2015; Kiesling et al. 2018). Thus, our findings do not exclusively apply to *The Globe and Mail* comments discussed here but can be extended to online news comments in general.

Placing these results in a wider theoretical context, we can conclude, first of all, that our analysis is a case in point for usage-based linguistic theory which sees language as a system that is constantly adapting and evolving through the complex interactions between language users, their social (inter)actions and cognitive processes (Beckner et al. 2009: 2). In light of this theory, we find evidence for how language adapts to the situational context of language production, including the medium of transmission. Online news comments do not constitute real-time dialogue or synchronous communication (cf. Reagle 2015), rather the writer and their audience are separated in space and time, and language is transmitted in written form through a computer or mobile device. Thus, the production of online comments is not constrained by the pressures of spontaneous production but allows for editing of the posts. The context of online commenting platforms is informal rather than official, while still being public. Commenters address their comments to a wider public audience, but are at the same time not restricted by formal linguistic requirements. In terms of functional properties, online commenting invites the commenters to express their personal opinion. As a result of the interactions between these situational and functional properties, the language of online comments exhibits many features typical of (written) informational registers and combines these with some features of involved production as well as features of argumentation and evaluation.

Secondly, it has been pointed out that online news comments are clearly positioned on the written end of the written-spoken continuum, yet, are marked by an informational-involved style, thus "mixing" written and spoken features. Biber and Egbert (2018) also discuss the "hybrid nature" of online language in terms of situational and communicative purposes of the relevant registers. They found that many online registers combine situational and linguistic features characteristic of multiple registers such as narrative plus opinion (Biber & Egbert 2018: 199-208). These findings emphasise the uniqueness of social media communication and suggest, perhaps, that social media language as a whole should be regarded as a third mode of present-day communication which exists alongside written and spoken language, rather than somewhere "in between" or "as a combination of" (cf. Ko 1996; Yates 1996). In such a perspective, the appropriateness of the current descriptive norms would need to be questioned because they may not suffice to fully describe online registers. In want of more appropriate methods and frameworks, current research including the present article, then, is only a first step towards understanding the nature and dynamics of language on the web.

Third, the present description could find application in comment moderation and the classification of "good" and "bad" comments. In particular, the incorporation of *be* + adjective patterns, which were shown to be a crucially distinctive characteristic of online news comments in SOCC, could improve current classification algorithms, by pinpointing which parts of the comment are most highly evaluative (and potentially negative or abusive).

Finally, the present article leaves many further avenues to pursue. Since we have established that online news comments are their own type of register, and different from face-to-face communication, a natural next step is to compare them to other online registers. Furthermore, it would be interesting to turn from the analysis of comment threads to the (bottom-up) profiling of individual comments and the analysis of interactional dynamics in online news comments.

Acknowledgements

The first author gratefully acknowledges funding from the Alexander von Humboldt Foundation through a Feodor Lynen Postdoctoral Research Fellowship. This research was also supported by the Social Sciences and Humanities Research Council of Canada (Insight Grant 435-2014-0171 to M. Taboada) and by NVIDIA Corporation, with the donation of a Titan Xp GPU. Thanks to Jack Grieve for helpful feedback and advice on MDA and to Andrea Nini for advice on tagging text for MDA analyses. We are also grateful to the audience of the 4th International Conference of the American Pragmatics Association in Albany, to editor Jesse Egbert and to the two anonymous reviewers for their feedback and comments.

Notes

[1] Ideally, input texts for standard MDA should be of a minimum length of 500 to 1000 words in order to obtain reliable frequency estimates of the linguistic features retrieved and hence to obtain reliable results (Jack Grieve p. c.; for a discussion see also Biber 1993).

[2] Academic and non-academic each comprise: humanities, social sciences, natural sciences, technology. Instructional writing: skills and hobbies, administrative writing. Scripted broadcast: scripted broadcast news, scripted broadcast talks. Broadcast: broadcast discussion, broadcast interview.

[3] It is acknowledged that online news comments may contain additional linguistic features (e.g. abbreviations like IMHO, *in my humble opinion*) which are specific/common to different types of web-based communication (Herring 2004). These are not included because the principal aim of our analysis is not to give an account of linguistic features unique to web-based registers or online news comments—this is left for future investigations— but instead to establish what online news comments are like vis-à-vis traditional, non-web-based registers, and face-to-face conversation in particular.

[4] The tagger also includes an option to analyse the annotated corpus and to output several visualisations placing the input corpus to the closest similar text type. Our analysis does not use this option but merely employs the program as a tagger.

[5] We took the liberty to add the register to the original ICE file names.

[6] As pointed out by an anonymous reviewer, the gradual distinctions between different written and spoken registers along the written-spoken continuum are also known as *diamesic dimension* (cf. Bernini and Schwartz 2011). For reasons of consistency, we stick to the more common Biberian term "written-spoken continuum".

[7] Factor 6 is visualised for completeness sake but not discussed as a full and reliable interpretation is not possible.

[8] Note that spoken and written registers overlap to a considerable degree, for instance, carefully prepared and (scripted) spoken registers such as scripted broadcast, or parliamentary debate are not as spontaneously produced as other spoken registers and tend to be comparatively informational.

References

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning, 59*(s1), 1-26.

Benamara, F., Inkpen, D., & Taboada, M. (2018). Language in social media: Exploiting discourse and other contextual information. Special issue of *Computational Linguistics, 44*(4).

Bernini, G., & Schwartz M. (2006). *Pragmatic organization of discourse in the languages of Europe*. (Vol. 20-8). Berlin: Mouton de Gruyter.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243-257.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, D., & Egbert, J. (2016). Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics*, *44*(2), 95-137.

Biber, D., & Egbert, J. (2018). *Register variation online*. Cambridge: Cambridge University Press.

Biber, D., Egbert, J., & Davies, M. (2015). Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. *Corpora*, *10*(1), 11-45.

Biber, D., & Finegan, E. (1989). Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text*, *9*(1), 93-124.

Biber, D., & Finegan, E. (Eds.) (1994). *Sociolinguistic perspectives on register*. Oxford: Oxford University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Bruce, R. F., & Wiebe, J. M. (1999). Recognizing subjectivity: a case study in manual tagging. *Natural Language Engineering*, *5*(2), 187-205.

Cambria, M. (2016). Commenting, interacting, reposting: A systemic-functional analysis of online newspaper comments. In S. Gardner & S. Alsop (Eds.), *Systemic Functional Linguistics in the Digital Age* (pp. 81-95). Sheffield: Equinox.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245-276.

Clarke, I., & Grieve, J. (2019). Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018. *PLoS ONE 14*(9): e0222062.

Clarke, I., & Grieve, J. (2017). Dimensions of abusive language on Twitter. In *Proceedings of the First Workshop on Abusive Language Online* (pp. 1-10). Vancouver: Association for Computational Linguistics.

Coe, K., Kenski, K., & Rains, S. A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658-679.

Collot, M., & Belmore, N. (1996). Electronic Language: A new variety of English. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives* (pp. 13-28). Amsterdam: John Benjamins.

Daems, J., Speelman, D., & Ruette, T. (2013). Register analysis in blogs: Correlation between professional sector and functional dimensions. *Leuven Working Papers in Linguistics*, *2*(1), 1-27.

Demata, M., Heaney, D., & Herrring, S. C. (2018). Language and discourse of social media. New challenges, new approaches. Special issue of *Altre Modernità*, I-X.

Diakopoulos, N. (2015). Picking the NYT Picks: Editorial criteria and automation in the curation of online news comments. *ISOJ Journal*, *6*(1), 147-166.

Diessel, H. (2017). Usage-based linguistics. *Oxford Research Encyclopedia of Linguistics. Oxford Research Encyclopedia of Linguistics.* Oxford: Oxford University Press.

Dziuban, C. D., & Shirkey, E. C. (1974). When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychological Bulletin, 81*(6), 358–361.

Godes, D., & Mayzlin, D. (2004). Using online conversations to study word-of-mouth communication. *Marketing Science*, *23*(4), 545–560.

Grieve, J., Biber, D., Friginal, E., & Nekrasova, T. (2010). Variation among blog text types: A multi-dimensional analysis. In A. Mehler, S., Sharoff, & M. Santini (Eds.), *Genres on the Web: Computational Models and Empirical Studies* (pp. 303–322). New York: Springer.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis: Pearson new international edition, always learning* (7th ed.). London: Pearson Education Limited.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Harlow: Longman.

Herring, S. C. (1996a). *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives*. Amsterdam: John Benjamins.

Herring, S. C. (1996b). Introduction. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives* (pp. 1-10). Amsterdam: John Benjamins.

Herring, S. C. (2004). Slouching toward the ordinary: Current trends in computer-mediated communication. *New Media & Society, 6*(1), 26-36.

Hunston, S. (2011). *Corpus approaches to evaluation: Phraseology and evaluative language.* New York: Routledge.

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

Ko, K.-K. (1996). Structural characteristics of computer-mediated language: A comparative analysis of InterChange discourse. *Electronic Journal of Communication/La Revue Électronique de Communication*, *6*(3), 1-28.

Kolhatkar, V., & Taboada, M. (2017a). Constructive language in news comments. In *Proceedings of the First Workshop on Abusive Language Online* (pp. 11-17). Vancouver: Association for Computational Linguistics.

Kolhatkar, V., & Taboada, M. (2017b). Using New York Times Picks to identify constructive comments. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism* (pp. 100-105). Copenhagen: Association for Computational Linguistics.

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., & Taboada, M. (2019). The SFU Opinion and Comments Corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*. https://doi.org/10.1007/s41701-019-00065-w

Marcoccia, M. (2004). On-line polylogues: Conversation structure and participation framework in internet newsgroups. *Journal of Pragmatics, 36*(1), 115-145.

Maxwell, A. E. (1959). Statistical methods in factor analysis. *Psychological Bulletin*, *56*(3), 228-235.

McGuire, J. (2015, November 30). Uncivil dialogue: Commenting and stories about indigenous people. *CBC News*. Retrieved from https://www.cbc.ca/newsblogs/community/editorsblog/2015/11/uncivil-dialogue-commenting-and-stories-about-indigenous-people.html

Mehler, A., Sharoff, S., & Santini, M. (2010). *Genres on the web: Computational models and empirical studies*. New York: Springer.

Moens, M.-F., Boiy, E., Mochales Palau, R., & Reed, C. (2007). Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law* (pp. 225-230). Palo Alto: Association for Computing Machinery.

Napoles, C., Tetreault, J., Rosato, E., Provenzale, B., & Pappu, A. (2017). Finding good conversations online: The Yahoo News Annotated Comments Corpus. In *Proceedings of the 11th Linguistic Annotation Workshop* (pp. 13-23). Valencia.

Nauroth, P., Gollwitzer, M., Bender, J., & Rothmund, T. (2015). Social identity threat motivates science-discrediting online comments. *PLoS One, 10*(2), e0117476.

Newman, J., & Columbus, G. (2010). *The ICE-Canada Corpus.* (Version 1). Retrieved from http://ice-corpora.net/ice/download.htm

Nini, A. (2014). Multidimensional Analysis Tagger - Manual (Version 1.3). Retrieved from http://sites.google.com/site/multidimensionaltagger

North, S. (2007). 'The voices, the voices': Creativity in online conversation. *Applied Linguistics*, *28*(4), 538-555.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Harlow: Longman.

R Core Team. (2018). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from URL http://www.R-project.org/

Reagle, J. M. (2015). *Reading the comments: Likers, haters, and manipulators at the bottom of the web*. Cambridge: MIT Press.

Rösner, L., Winter, S., & Krämer, N. C. (2016). Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior, 58*, 461–470.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267-307.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 252-259.

Tseronis, A. (2011). From connectives to argumentative markers: A quest for markers of argumentative moves and of related aspects of argumentative discourse. *Argumentation*, *25*(4), 427-447.

van Eemeren, F. H., Houtlosser, P., & Snoeck Henkemans, A. F. (2007). *Argumentative indicators in discourse: A pragma-dialectical study*. New York: Springer.

Weizman, E., & Dori-Hacohen, G. (2017). On-line commenting on opinion editorials: A cross-cultural examination of face work in the Washington Post (USA) and NRG (Israel). *Discourse, Context & Media*, *19*, 39-48.

White, L. (2003). *Second language acquisition and universal grammar*. Cambridge: Cambridge University Press.

Woollaston, V. (2013, September 30). Online conversations are damaging how we speak to each other in real life: Author claims people could soon “forget” how to handle social situations. *Daily Mail*. Retrieved from http://www.dailymail.co.uk/sciencetech/article-2439336/Online-conversations-damaging-speak-real-life-claims-author.html

Yates, S. J. (1996). Oral and written linguistic aspects of computer conferencing: A corpus based study. In S. Herring (Ed.), *Computer-Mediated Communication: Linguistic, social, and cross-cultural perspectives* (pp. 29-46). Amsterdam: John Benjamins.

Appendices

Appendix A

Table 4: Description of part-of-speech tags.

Part-of-speech tag	**Description**
AMP	Amplifiers
ANDC	Non-phrasal coordination
AWL	Average word length
BEMA	BE as main verb
BYPA	BY passive
CAUS	Subordinator cause
CONC	Subordinator concession
COND	Subordinator condition
CONJ	Conjuncts
CONT	Contractions
DEMO	Demonstratives
DEMP	Demonstrative pronoun
DPAR	Discourse particles
DWNT	Downtoners
EMPH	Emphatics
EX	Existential THERE
FPP1	1st person pronouns
GER	Gerunds
HDG	Hedges

INPR	Indefinite pronouns
JJ	Adjectives attributive
NEMD	Modals necessity
NN	Nouns
NOMZ	Nominalisations
OSUB	Subordinator other
PASS	Agentless passives
PASTP	Past participle clauses
PEAS	Perfect aspect
PHC	Phrasal coordination
PIN	Prepositions
PIRE	Pied-piping relatives
PIT	Pronoun IT
PLACE	Place adverbials
POMD	Modals possibility
PRED	Adjectives predicative
PRESP	Present participial clauses
PRIV	Private verbs
PRMD	Modals predictive
PROD	DO pro-verb
PUBV	Public verbs
RB	Adverbs
SERE	Sentence relatives
SMP	SEEM/APPEAR
SPAU	Split auxiliaries
SPIN	Split infinitives
SPP2	2nd ps. pronouns
STRP	Stranded prepositions
SUAV	Suasive verbs
SYNE	Synthetic negation
THAC	THAT adjective complements

THATD	THAT deletion
THVC	THAT verb complements
TIME	Time adverbials
TO	Infinitives
TOBJ	THAT relatives (object position)
TPP3	3rd person pronouns
TSUB	THAT relatives (subject position)
TTR	Type-token ratio
VBD	Past tense verbs
VPRT	Present tense verbs
WHCL	WH clauses
WHOBJ	WH relatives (object position)
WHQU	WH questions
WHSUB	WH relatives (subject position)
WZPAST	Past participle WHIZ deletion
WZPRES	Present participial WHIZ deletion
XX0	Analytic negation

Appendix B

Table 5: Promax-rotated factor solution with six factors. Negative polarity indicates complementary distribution; positive polarity indicates co-occurrence of the features.

Part-of-speech tag	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**	**Factor 6**
AMP	0.18	0.153	-0.195	0.324	0	-0.016
ANDC	0.332	-0.093	-0.36	0.179	0.07	0.173

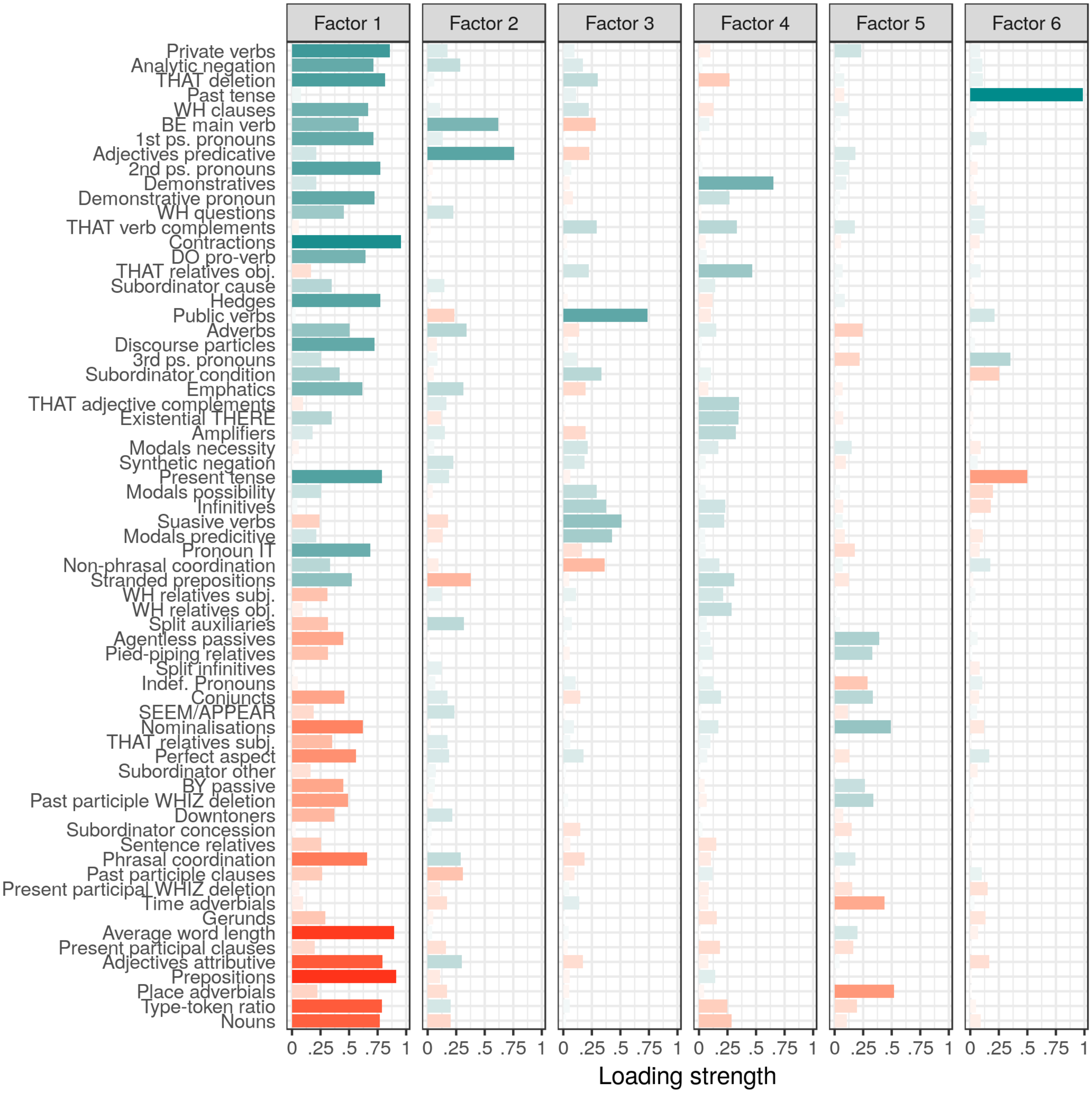
AWL	-0.894	0.042	0.047	-0.016	0.202	-0.068
BEMA	0.585	0.621	-0.284	0.09	0.05	-0.033
BYPA	-0.451	0.062	-0.004	-0.051	0.263	0.028
CAUS	0.347	0.145	-0.011	0.142	0.04	-0.018
CONC	-0.034	0.034	-0.15	0.029	-0.148	0.005
COND	0.418	-0.055	0.332	0.103	0.018	-0.253
CONJ	-0.461	0.172	-0.15	0.194	0.334	-0.079
CONT	0.954	-0.011	-0.032	-0.058	-0.058	-0.081
DEMO	0.215	-0.022	-0.057	0.653	0.105	-0.03
DEMP	0.723	-0.033	-0.083	0.266	0.046	-0.06
DPAR	0.723	-0.082	-0.043	0.028	-0.041	0.042
DWNT	-0.369	0.215	-0.005	0.012	-0.077	-0.035
EMPH	0.615	0.311	-0.193	-0.084	-0.072	-0.028
EX	0.348	-0.12	-0.021	0.346	-0.077	-0.028
FPP1	0.714	0.131	0.021	-0.022	0.023	0.143
GER	-0.291	-0.051	0.019	-0.154	-0.025	-0.135
HDG	0.772	-0.027	-0.036	-0.119	0.091	-0.037
INPR	-0.053	0.066	0.106	0.126	-0.288	0.106
JJ	-0.793	0.3	-0.174	-0.08	0.041	-0.168
NEMD	-0.062	0.057	0.216	0.171	0.149	-0.091
NN	-0.768	-0.204	0.026	-0.288	-0.108	-0.09
NOMZ	-0.622	-0.027	0.091	0.168	0.493	-0.121
OSUB	-0.161	0.073	0.025	0.012	-0.013	-0.066
PASS	-0.448	-0.015	0.035	0.1	0.392	0.062
PASTP	-0.263	-0.309	-0.1	0.121	-0.047	0.103
PEAS	-0.559	0.189	0.175	0.071	-0.131	0.168
PHC	-0.657	0.291	-0.184	-0.107	0.181	-0.003
PIN	-0.914	-0.109	-0.056	0.144	-0.006	-0.015
PIRE	-0.316	0.02	-0.057	0.12	0.329	0.02
PIT	0.687	0.016	-0.163	0.058	-0.177	-0.082
PLACE	-0.222	-0.171	-0.053	-0.046	-0.522	-0.011

POMD	0.253	-0.046	0.292	0.058	0.048	-0.197
PRED	0.214	0.76	-0.228	0.005	0.184	0.02
PRESP	-0.199	-0.16	-0.037	-0.185	-0.165	-0.008
PRIV	0.859	0.176	0.097	-0.1	0.234	0.089
PRMD	0.215	-0.132	0.429	0.056	-0.09	-0.109
PROD	0.643	-0.017	0.036	0.067	0.013	-0.031
PUBV	0.033	-0.233	0.738	-0.103	-0.018	0.211
RB	0.507	0.342	-0.14	0.153	-0.245	0.01
SERE	-0.253	0.024	-0.061	-0.152	0.036	-0.017
SMP	-0.19	0.234	-0.019	0.021	-0.119	0.06
SPAU	-0.313	0.319	0.074	0.068	0.049	-0.015
SPIN	-0.021	0.123	0.005	0.032	0.014	-0.081
SPP2	0.774	-0.043	0.069	0.03	0.126	-0.065
STRP	0.523	-0.377	-0.053	0.311	-0.125	-0.029
SUAV	-0.241	-0.179	0.51	0.222	0.07	-0.012
SYNE	-0.014	0.226	0.185	0.061	-0.098	0.063
THAC	-0.098	0.164	0.021	0.351	0.033	0
THATD	0.814	0.056	0.301	-0.268	0.085	0.111
THVC	-0.059	-0.026	0.293	0.332	0.177	0.125
TIME	-0.096	-0.171	0.141	-0.083	-0.436	0.035
TO	0.046	-0.005	0.378	0.232	-0.077	-0.18
TOBJ	-0.167	-0.011	0.224	0.467	0.071	0.09
TPP3	0.257	0.084	0.119	0.005	-0.217	0.352
TSUB	-0.35	0.174	0.059	0.1	-0.005	-0.007
TTR	-0.787	0.202	0.054	-0.25	-0.197	-0.048
VBD	0.077	0.009	0.114	-0.006	-0.083	0.984
VPRT	0.788	0.186	-0.06	-0.008	0.003	-0.499
WHCL	0.669	0.107	0.222	-0.13	0.123	0.057
WHOBJ	-0.095	-0.008	-0.017	0.288	0.021	0.009
WHQU	0.456	0.224	0.031	-0.023	0.027	0.129
WHSUB	-0.312	0.124	0.111	0.21	0.023	0.047

WZPAST	-0.49	-0.043	0.04	-0.068	0.34	0.026
WZPRES	-0.065	-0.11	0.053	-0.088	-0.154	-0.154
XX0	0.714	0.285	0.171	0.029	0.047	0.106

Appendix C

Figure 3: Loading strength of individual features across the six factors. Green bars indicate positive loadings; red bars indicate negative loadings. Colour saturation indicates loading strength.



Appendix D

Table 6 and Figure 4.

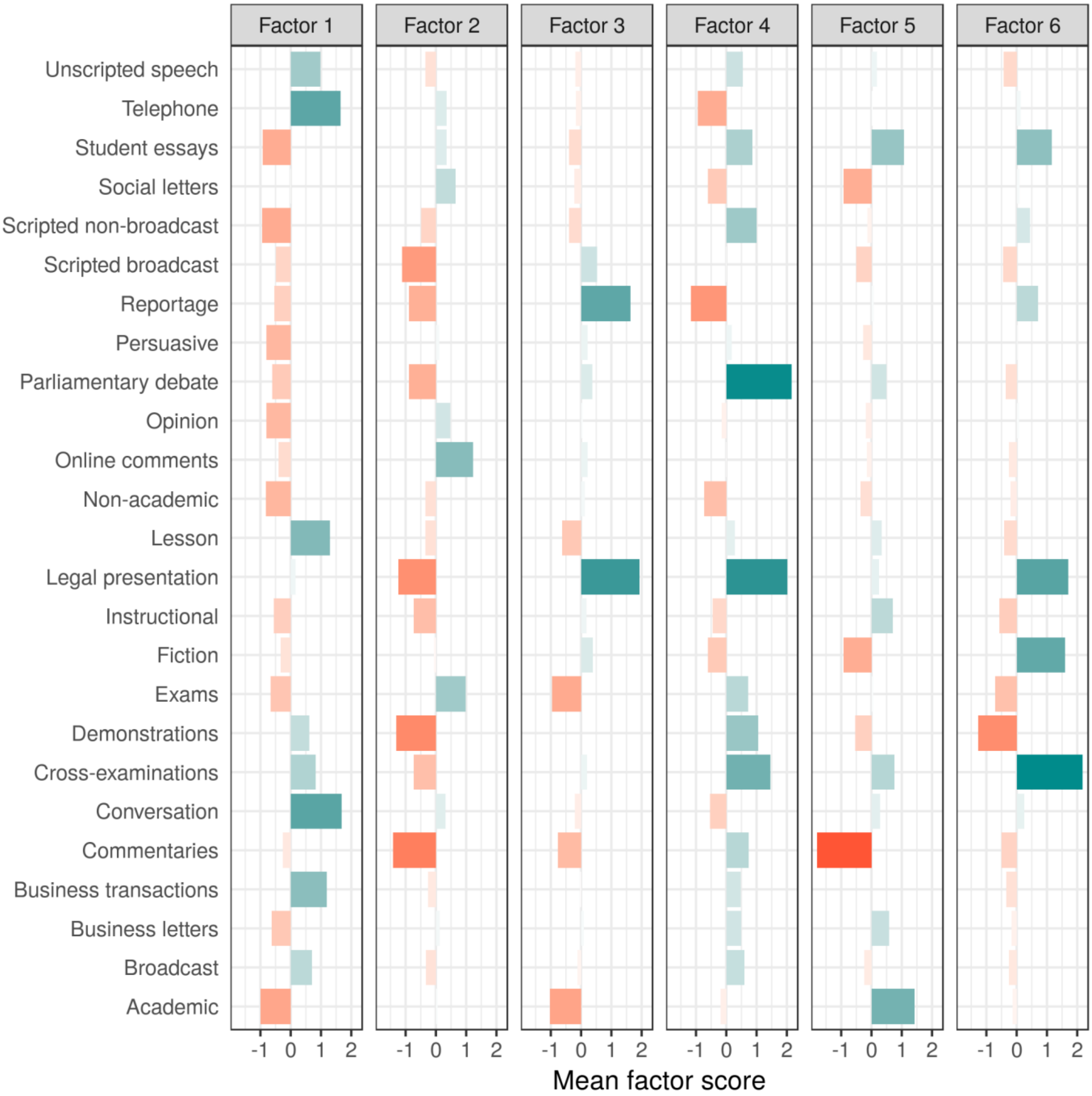
Table 6: Mean factor scores per register across the six factors/dimensions. Negative polarity indicates complementary distribution; positive polarity indicates co-occurrence of the features.

Values were rounded (for original unrounded values see https://github.com/sfu-discourse-lab/MDA-OnlineComments).

Register	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**	**Factor 6**
Academic	-1.0013	0.0618	-1.0281	-0.1845	1.4267	-0.1279
Broadcast	0.7004	-0.3263	-0.1231	0.5980	-0.2318	-0.2637
Business letters	-0.6272	0.1257	0.0954	0.4919	0.5822	-0.1602
Business transactions	1.1980	-0.2525	-0.2247	0.4817	0.0018	-0.3404
Commentaries	-0.2602	-1.4111	-0.775	0.7462	-1.79	-0.5010
Conversations	1.6743	0.3166	-0.2013	-0.5382	0.2927	0.2434
Cross-examinations	0.8163	-0.7272	0.1917	1.4517	0.7636	2.1728
Demonstrations	0.6230	-1.3027	-0.0025	1.0515	-0.534	-1.265
Exams	-0.658	0.9849	-0.962	0.7242	0.065	-0.7065
Fiction	-0.3247	-0.0745	0.3874	-0.6075	-0.9279	1.6025
Instructional	-0.553	-0.7317	0.1789	-0.4482	0.6995	-0.5668
Legal presentation	0.1648	-1.2341	1.9243	2.015	0.2571	1.6968
Lesson	1.2921	-0.3422	-0.6280	0.281	0.3412	-0.4097
Non-academic	-0.8203	-0.3547	0.1236	-0.7273	0.3633	-0.2123
Online comments	-0.3999	1.2342	0.2071	-0.0089	-0,1571	-0.253
Opinion	-0.8070	0.4733	0.0607	-0.1585	-0.1947	0.0698
Parliamentary debate	-0.6024	-0.8972	0.3684	2.1562	0.4954	-0.3684
Persuasive	-0.8119	0.1076	0.2097	0.1838	-0.2674	0.0189
Reportage	-0.5327	-0.8969	1.6304	-1.176	0.073	0.7073
Scripted broadcast	-0.4804	-1.1198	0.5283	-0.0042	-0.4978	-0.4441
Scripted non-broadcast	-0.9499	-0.4832	-0.4053	1	-0.1379	0.4387
Social letters	-0.0163	0.6454	-0.2196	-0.6	-0.9181	0.0881
Student essays	-0.9339	0.3481	-0.4016	0.8609	1.0779	1.1691
Telephone	1.6499	0.3576	-0.1768	-0.933	0.0442	0.1326
Unscripted speech	0.9734	-0.3415	-0.1933	0.544	0.1867	-0.4338

Figure 4: Mean factor scores per register for each of the six dimensions. Green bars indicate positive factor scores; red bars indicate negative factor scores. Colour saturation indicates

absolute weight of mean factor scores.


Factor 1
Factor 2
Factor 3
Factor 4
Factor 5
Factor 6
Unscripted speech
Telephone
Student essays
Social letters
Scripted non-broadcast
Scripted broadcast
Reportage
Persuasive
Parliamentary debate
Opinion
Online comments
Non-academic
Lesson
Legal presentation
Instructional
Fiction
Exams
Demonstrations
Cross-examinations
Conversation
Commentaries
Business transactions
Business letters
Broadcast
Academic
Mean factor score