The interplay of complexity and subjectivity in opinionated discourse

Katharina Ehret and Maite Taboada
Discourse Processing Lab, Department of Linguistics
Simon Fraser University

Abstract

This paper brings together cutting-edge, quantitative corpus methodologies and discourse analysis to explore the relationship between text complexity and subjectivity as descriptive features of opinionated language. We are specifically interested in how text complexity and markers of subjectivity and argumentation interact in opinionated discourse. Our contributions include the marriage of quantitative approaches to text complexity with corpus linguistic methods for the study of subjectivity, in addition to large-scale analyses of evaluative discourse. As our corpus, we use the *Simon Fraser Opinion and Comments Corpus* (SOCC), which comprises approximately 10,000 opinion articles and the corresponding reader comments from the Canadian online newspaper *The Globe and Mail*, as well as a parallel corpus of hard news articles also sampled from *The Globe and Mail*. Methodologically, we combine conditional inference trees with the analysis of random forests, an ensemble learning technique, to investigate the interplay between text complexity and subjectivity. Text complexity is defined in terms of Kolmogorov complexity, i.e., the complexity of a text is measured based on its description length. In this approach, texts which can be described more efficiently are considered to be linguistically less complex. Thus, Kolmogorov complexity is a measure of structural surface redundancy. Our take on subjectivity is inspired by research in evaluative language, stance and Appraisal and defined as the expression of evaluation and opinion in language. Drawing on a sentiment analysis lexicon and the literature on stance markers, a custom set of subjectivity and argumentation markers is created. The results show that complexity can be a powerful tool in the classification of text into different text types, and that stance adverbials serve as distinctive features of subjectivity in online news comments.

Keywords
discourse analysis, complexity, subjectivity, opinionated discourse, corpus linguistics, text linguistics

1 Introduction

The research we present in this paper is situated at the nexus of discourse analysis, text linguistics and theoretical complexity research. Our paper is inspired by the current boom in studies on language complexity which is all about defining, measuring and explaining complexity at various linguistic levels (e.g., Mufwene et al. 2017; Baechler and Seiler 2016; Baerman et al. 2015; Kortmann and Szmrecsanyi 2012). The interest in language complexity sprung up about two decades ago in the sociolinguistics-typological community and originally centered around the question of whether some languages are simpler or more complex than others (e.g. McWhorter 2001). Although most of this research focuses on language complexity from a typological perspective, some publications explore language complexity in different text types. Recently, it has been shown that different text types systematically vary in regard to their complexity, e.g., newspaper writing tends to be more complex than conversation or email (Ehret 2018, 2017; Szmrecsanyi 2009). Furthermore, complexity variation in different text types is closely linked to the formality of the functional-communicative context and discourse situation (Ehret 2018). As a matter of fact, in text linguistics, complexity has been known to be a distinguishing characteristic between different text types, most notably academic writing vs. conversation, for some time (Biber and Gray 2016; Biber et al. 2011; Biber and Gray 2010). In discourse analysis, however, complexity is a less explored construct—despite the fact that it would work well as a descriptive feature for discourse categorisation.

In contrast, notions of subjectivity have been extensively studied in discourse analysis, corpus linguistics, and related fields: Subjectivity features prominently in the study of opinion and sentiment (Wiebe et al. 2004; Wiebe and Riloff 2005), appraisal (Martin and White 2005) as well as analyses of stance (Biber and Finegan 1989; Englebretson 2007), and is the principal object of study in many corpus-based analyses of evaluation (Hunston and Thompson 2000b; Bednarek 2006).

Against this backdrop, we sketch a quantitative, corpus-based approach to discourse analysis by exploring the interplay between text complexity and subjectivity in opinionated discourse. Our paper thus fills an important gap in the literature on subjectivity and evaluation because next to nothing is currently known about how text complexity interacts with lexical markers of subjectivity and argumentation,

Corresponding author:
Maite Taboada, Department of Linguistics, Simon Fraser University, 8888 University Dr, Burnaby, BC, V5A 1S6, Canada
Email: mtaboada@sfu.ca

nor how text complexity characterises distinct types of opinionated discourse vis-à-vis arguably more objective news articles. We thus specifically address questions such as whether increased levels of subjectivity result in increased complexity in texts, and whether lexical markers of subjectivity and text complexity correlate with the perceived subjectivity of different text types (such as the hard news, opinion articles and reader comments we investigate here).

Methodologically, the interplay between text complexity and subjectivity, and their respective impact on the formation of distinct text types is assessed by combining conditional inference trees, a type of classification and regression trees, with the analysis of random forests. In this paper, we draw on an innovative information-theoretic metric of complexity to assess text complexity at an overall, morphological and syntactic level. The measure is based on the notion of *Kolmogorov complexity* and measures the linguistic complexity of a text in terms of its description length: The shorter the information-theoretic description of a text, the lower is its linguistic complexity. In linguistic terms, Kolmogorov complexity is a holistic measure of structural surface redundancy and regularity. Specifically, an implementation of the measure known as the *compression technique* which has been adapted for the analysis of naturalistic corpora (Ehret 2017) is utilised. Our definition of subjectivity is guided by the interest in discovering the relationship between text complexity and subjectivity in a quantitative and usage-based fashion. Therefore, we take a practical view on subjectivity and define it in terms of specific lexico-grammatical items that serve the goal of conveying opinion, evaluation, attitude, stance and subjectivity in a text. To be more precise, the set of relevant linguistic features includes an extensive list of evaluative words (Taboada et al. 2011), stance adverbials (Biber 1988; Biber et al. 1999), modal verbs, and connectives which mark argumentative discourse relations (Prasad et al. 2007).

Overall, we propose a strongly quantitative, corpus-based and large-scale approach to investigate the complex relationship between complexity and subjectivity. Although *Discourse Studies* readers may be familiar with different approaches to subjectivity, we believe that the approach to complexity presented here, and its application to the study of subjectivity in different text types, presents a novel way of analysing discourse.

The results demonstrate that text complexity is a powerful tool in distinguishing different text types, particularly opinionated discourse from news articles, and that stance adverbials interact with complexity in online news comments. More generally, we find evidence for the fact that subjectivity in language is pervasive even in supposedly objective texts (cf. Alba-Juez and Thompson 2014).

This paper is structured as follows. In Section 2 we provide some background on the study of language complexity and approaches to analysing subjectivity. Section 3 introduces the data source. Section 4 gives an overview of the features and measures utilized. In Section 5 the statistical analysis is described and in Section 6 the results are presented and discussed. Section 7 offers some concluding remarks.

2 Complexity meets discourse analysis

The starting point of the current debate on language complexity was the question of whether all languages were equally complex or not (e.g., McWhorter 2001; Kusters 2003; Shosted 2006). Among sociolinguists and typologists, it had been the common belief throughout much of the twentieth century that, on the whole, all languages are equally complex (e.g., Edwards 1994; Crystal 1987). In the meantime, plenty of empirical evidence has shown that the complexity of languages and language varieties can and does differ. Theoretical complexity research is now primarily concerned with the definition and measurement of language complexity. Most of this research analyses the complexity of either single languages (e.g., McWhorter 2012), cross-linguistic datasets (Lupyan and Dale 2010) or different geographical varieties of the same language (e.g., Szmrecsanyi and Kortmann 2009). Of particular interest to the current paper, however, are a few studies investigating the complexity in different text types of English.[1]

Szmrecsanyi (2009) analysed the written and spoken texts sampled in the *British National Corpus* (BNC) in terms of three complexity measures, which relate to the frequency of specific grammatical markers. The results show that different written and spoken text types differ in their complexity, and that these differences correlate with the well-known dimensions of textual variation established by Biber (1988) (Szmrecsanyi 2009, 335-336). Specifically, Biber (1988) describes the following six dimensions: 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. A more recent analysis of text types in the BNC relies on Kolmogorov complexity to assess text complexity at the overall, morphological and syntactic level (Ehret 2018). Finding substantial variation in complexity across the various BNC text types, the study stresses that text type, or more generally, the situational context of production, is a crucial factor for language-internal complexity variation and provides empirical evidence for the fact that language adapts to the situational context of production. Furthermore, Kolmogorov complexity captures variation along Biber's *Involved vs. Informational Dimension* (Ehret 2018, 18-20), which distinguishes between highly involved production such as face-to-face conversation on the one hand, and abstract-informational production such as academic writing or newspaper writing on the other hand (Biber 1988, 107-108). For instance, Ehret (2018) reports that formal registers (e.g., newspapers, biography, academic writing) are overall complex and exhibit a lot of morphological variation while comparatively informal/involved registers (e.g., conversation, public debate) are marked by fixed word order (Ehret 2018, 16, 18-19). This is in line with prominent literature on text linguistics

[1] In this paper, we use the term "text type" rather than "register" or "genre", because the latter are used with different meanings in different frameworks (Van Dijk 1997). Our understanding of what constitutes a text type closely follows the register perspective described in Conrad and Biber (2000, 16): Different text types are characterised by the (functionally motivated) use of linguistic features in specific communicative/discourse situations.

and register variation, which has long since established that language varies according to formality and situational context (e.g., Biber 1988).

In text linguistics, complexity, among other things, is also known as a linguistic feature characterising different text types. In particular, the complexity of conversation, as a prototypical spoken register, and academic writing, as a prototypical written register, has been extensively studied (Biber and Gray 2016, 2010). Both conversation and academic writing turned out to be complex, however, in structurally different ways. Complexity in conversation is related to subordination, while noun phrase structures and phrasal modifiers generate complexity in academic writing (Biber and Gray 2010, 2-4). Similarly, Biber et al. (2011) analysed the adequacy of measures commonly used in the assessment of grammatical complexity in second language writing. The authors found that common measures like T-unit length and increased subordination are characteristic of spoken language, i.e., conversation, rather than of academic writing, which is marked by phrasal complexity and a wide range of lexico-grammatical combinations (Biber et al. 2011, 28-31). In short, the complexity of different text types crucially depends on the use of specific grammatical structures which, in turn, are chosen according to the functional- situational and communicative setting of language production. In other words, complexity is a well- established metric for distinguishing different text types, and therefore well suited for the linguistic characterisation of opinionated discourse. Note, however, that the formulation of descriptive complexity in work by Biber and colleagues is not adopted here. In Biber and Gray (2016), the approach is structural and feature-specific, i.e., certain grammatical features are considered to be indicative of more complex writing. For instance, dependent clauses (conditional, non-finite complement clauses), relative clauses, and prepositional phrases as noun modifiers are markers of complex academic writing. In our approach, by contrast, complexity is not feature-specific, but holistic, in that it considers the entire structural complexity in texts, and can therefore be applied across text types and languages without the need to adjust which specific grammatical features should be considered.

Thus, our particular interest is in the role of complexity in opinionated discourse, specifically, in how it interacts with expressions of subjectivity and evaluation in opinion articles and reader comments. Taking a practical view on subjectivity, we aim at a narrow definition of subjectivity, mostly concerned with how opinion is conveyed through specific linguistic items. Numerous areas of linguistics and other social sciences study subjectivity, often under different terms, and in some cases, using very different methodology. Studies of stance, evidentiality, attitude, sentiment or appraisal are concerned with overlapping linguistic phenomena. Some of the first definitions of subjectivity describe the concept in terms of point of view, i.e. as the speaker's expression of themselves in an utterance, the definition that Lyons (1981) introduced (probably inspired by Benveniste (1966)), and sometimes from the perspective of a physical point of view (Langacker 1985). Hunston and Thompson (2000a) refer to subjectivity in terms of evaluation and propose that subjectivity comprises two aspects, namely, modality and a

point-of-view aspect, variously called evaluation, appraisal or stance. In some approaches, modality and evaluation are two distinct phenomena (Halliday 1985; Martin and White 2005), and in some cases the two expressions of subjectivity are combined under one label, such as stance (Biber and Finegan 1989; Conrad and Biber 2000). Bednarek (2006) proposes a tripartite classification of evaluation: perspective (the point-of-view aspect), affect (emotion and expression of the self), and modality (epistemic status of propositions). Biber and Finegan approach stance from a purely textual level and define it as "the lexical and grammatical expression of attitudes, feelings, judgments, or commitment concerning the propositional content of a message" (Biber and Finegan 1989, 93). This is the approach we take in this paper, coupled with previous work on the subjective nature of some discourse relations (Trnavac et al. 2016). The view of subjectivity adopted here is also informed by sentiment analysis, the computational extraction of subjective and evaluative meaning from text. We choose to use the term 'subjectivity' as a more general term, since we do not espouse the more specific approaches evoked by terms such as evaluation or stance.

In addition to Biber and colleagues' approach to lexical markers of stance, we also consider how certain discourse relations signal subjectivity. In the framework of Rhetorical Structure Theory (Mann and Thompson 1988), Trnavac and Taboada (2012) and Trnavac et al. (2016) have studied how different types of discourse relations are indicators of subjectivity and evaluation, following well-established distinctions in the literature between semantic and pragmatic relations (van Dijk 1979), external and internal discourse relations (Martin 1992) or subjective and objective relations (Sanders et al. 1993). Thus, in this paper, we focus on discourse connectives that are markers of subjective relations, as indicators of subjectivity in text.

Finally, sentiment analysis and opinion mining focus on how subjectivity is linguistically conveyed in texts (Wiebe and Riloff 2005; Wiebe et al. 2004). These fields do, however, not only encompass the study of evaluative expressions but are also concerned with their direction and strength: subjective expressions can be either positive (*love*) or negative (*hate*) and can vary in the degree of subjectivity expressed (*dislike* vs. *hate* vs. *loath*) (Taboada et al. 2011; Wiebe 2000). In sentiment analysis, subjectivity is essentially defined as the linguistic expression of opinion and evaluation.

As can be seen from this brief run-through of the literature on subjectivity—which by no means does justice to the vast amount of publications on the topic —subjectivity is a multi-faceted and hard-to-define concept. Maybe a core definition could be derived from the goal of most of the research on subjectivity and evaluation which is "[...] to explicate the range of lexical, grammatical, textual, and intertextual means by which speakers and writers laminate their language with their attitudes and points-of-view about its content." (Englebretson 2007, 16). Given that our research is quantitative and large-scale, we focus on the first of those means, the lexical level, adding some aspects of the grammatical level, in the form of grammatical structures that convey evaluation. Our definition of subjectivity, then, is one

where subjectivity is defined as the expression of evaluation and opinion through specific lexical items and grammatical structures, which can include a variety of devices, with modal expressions being one of them. Hunston and Thompson (2000a) use the umbrella term *evaluation* for this combination. This definition is congruent with the scope of phenomena under the Appraisal Framework (Martin and White 2005) and is also similar to stance in Conrad and Biber (2000) and Biber and Finegan (1989).

In broad strokes (a full description is provided in Section 3), subjectivity is operationalised as (i) specific evaluative words (adjectives, nouns, verbs and adverbs); (ii) stance adverbs; (iii) modal verbs; and (iv) connectives which mark argumentative discourse relations. Although our take on subjectivity and evaluation is quantitative and large-scale, these expressions occur in the context of naturalistic discourse (Alba-Juez and Thompson 2014), where they fully develop their meaning. The study of subjectivity and evaluation in context is therefore certainly within the scope of discourse analysis. As Du Bois (2007, pp.140-141) points out, when researching how stance works in language, "we find ourselves faced with a complex web of interconnections linking stance with dialogicality, intersubjectivity, the social actors who jointly enact stance, and the mediating frameworks of linguistic structure and sociocultural value they invoke in doing so."

3 Database

We tap the *Simon Fraser Opinion and Comments Corpus* (SOCC), the up-to-date largest collection of opinionated language. SOCC comprises roughly 10,000 opinion articles and 660,000 reader comments posted in response to these opinion pieces. Thus, a direct comparison between different types of opinionated discourse—the more formal opinion articles vs. the comparatively more personal and informal reader comments—is facilitated. The data spans the time period from 2012 to 2016 and was collected from the Canadian online newspaper *The Globe and Mail* (Kolhatkar et al. 2020). SOCC has been specifically designed to analyse the characteristics of online comments preserving the comments' thread-structure and providing extensive discourse annotation (e.g. Appraisal and negation annotation). In this paper the plain text version of the articles and comment threads is used.

As a benchmark for our measures and in order to compare the textual properties of opinionated discourse to (arguably) more objective discourse, we compiled a complementary corpus of news articles from *The Globe and Mail* following the sampling guidelines outlined in Kolhatkar et al. (2020).

We furthermore sample 'page-one stories' which cover a wide range of general and, at the time of their publication, topical issues to immunise the database against potential lexical bias. In other words, we argue that page-one-stories are more representative in terms of vocabulary and grammatical structures than, for instance, a database comprising a random sample of articles or articles of a particular category (e.g. political). The collected news corpus counts 4,600 articles.

Text type	Number of texts	Number of words
Opinion articles	3,509	2,758,993
Comment threads	3,509	22,191,197
News articles	3,430	3,332,559
Total	10,448	28,282,749

Table 1. Number of texts per text type in the final database.

We then generate a random sample of SOCC articles and comment threads that approximates the news corpus in size and number of texts to obtain a database with roughly the same size per text type. In general, attention is furthermore restricted to texts with a minimum number of 700 words because the complexity measure utilized requires comparatively large texts to return representative results (see Section 4.1; for a discussion see Ehret 2017). A statistical overview of the final database is given in Table1.

4 Features and measures

4.1 Complexity

We use an unsupervised, information-theoretic metric of complexity to assess the complexity of our texts at an overall, morphological and syntactic level. The metric is based on the notion of Kolmogorov complexity, which can be approximated with text compression programs such as *gzip* or *WinRAR* commonly used to compress or minimise large files. The Kolmogorov complexity of a text is measured as the length of the shortest possible description that is necessary to recreate the original text (Li and Vitanyi 1997, 48). To illustrate, when compressing a text with a compression program, the text is minimised in such a way that the program can recreate the original, uncompressed text without loss of information or content from the compressed version (which is the shortest possible description of the text). Basically, the idea is that texts which can be compressed comparatively better, i.e., more efficiently, are linguistically comparatively less complex. Consider the two text strings below. Both strings count 20 characters, however, string 1a can be described more efficiently as the pattern *a rose tree* occurs twice, whereas there is no recurrent pattern in string1b that could be compressed. Measuring the complexity of these strings based on how well they can be compressed, string 1a is less complex than string 1b.

(1) a. a rose tree is a rose tree (20 characters)

b. a rose tree is very pretty (20 characters)

The Kolmogorov complexity metric was first introduced by Juola (1998, 2008), and substantially extended by Ehret (2017), who also adapted it for the application to naturalistic corpus data (Ehret 2018; Ehret and Szmrecsanyi 2016a). In this paper the implementation of the metric known as compression technique (Ehret 2017) is utilized.[2] Kolmogorov-based complexity as measured with the compression technique is a text-based and holistic metric of language complexity that can be linguistically interpreted as a measure of structural redundancy and regularity (Ehret 2017, 169-171; see also Ehret and Szmrecsanyi 2016b). As such, the metric is more or less agnostic about form-function pairings, yet it captures recurring linguistic structures and regularities. In other words, it is restricted to formal linguistic structures of entire texts as opposed to individual features or constructions. That said, Kolmogorov complexity measurements have been shown to dovetail with more traditional assessments of complexity, for example, in English text types (Ehret 2018) or learner essays (Ehret and Szmrecsanyi 2016a). Thanks to its radical (con)text-basedness Kolmogorov complexity measurements are inherently usage-based (because they are based on naturalistic language data) and contextualised (because they are based on the entire context of a text rather than being reduced to a selection of specific features). Therefore, it is ideal for assessing the complexity of our opinion database because we are interested in big-picture complexity and the global structural properties of the texts (rather than low-level categories of complex and simple features). Furthermore, the technique can be readily applied to large datasets whose manual annotation would otherwise be empirically expensive.

Overall complexity is here defined as the global structural complexity of an original text. It is based on two measurements for each text: the file size (in bytes) before compression and the file size (in bytes) after compression. Linear regression analysis of these file size pairings eliminates the trivial correlation between these measures and returns the *adjusted overall complexity scores*. Higher scores are interpreted as indicating higher overall linguistic complexity of a text, while lower scores are interpreted as indicating lower overall complexity of a text (Ehret 2017, 48-49).

Morphological and syntactic complexity can be indirectly assessed by manipulating the morphological and syntactic information, respectively, in the texts before compressing them (Juola 2008, 1998). The morphological and syntactic complexity scores presented here are therefore essentially indicators of how well the compression algorithm deals with the morphological/syntactic noise created through manipulation. Text manipulation is operationalized through random deletion. Morphological manipulation is performed by random deletion of 10% of all characters in each text. This creates new "word forms" and compromises morphological regularity. The rationale behind this manipulation is that morphologically complex texts contain overall a relatively large amount of word forms in any case, and

[2]The scripts for implementing the compression technique are available at `https://github.com/katehret/measuring-language-complexity`.

are thus less affected by random noise than morphologically less complex texts (which contain an overall smaller number of word forms). Hence, comparatively bad compression ratios after morphological distortion index low morphological complexity. Syntactic manipulation is performed by random deletion of 10% of all word tokens in each text. This procedure compromises word order regularities. Complex texts, i.e. texts with relatively fixed word order, are greatly affected because syntactic interdependencies and word order patterns are disrupted. Texts with comparatively free word order, in contrast, are less affected because they lack syntactic interdependencies which could be disrupted. Comparatively bad compression ratios after syntactic manipulation thus index comparatively high syntactic complexity (Ehret 2017, 49-50).

In linguistic terms, the morphological complexity score is a measure of structural word form variation and indicates the extent to which texts contain comparatively many (inflectional or derivational) forms as opposed to being invariant. Kolmogorov-based morphological complexity to some extent correlates with lexical diversity and the frequent use of varied phrasal and lexico-grammatical structures (cf. Ehret 2018, 17). The syntactic complexity score is a measure of word order flexibility and indicates the extent to which word order in a text is flexible or rigid. Or put differently, it is connected to how many different word order patterns occur in a text. Maximally flexible word order is defined as simple while maximally rigid word order is defined as complex (because there are more word order rules) (Ehret 2017, 48-50). This is in line with the common view in the theoretical complexity literature that more rules count as more complex (e.g. Arends 2001; McWhorter 2001).

On a technical note, the compression technique is applied with multiple random permutations and uses *gzip* (version 1.2.4., `http://www.gzip.org/`) to compress the corpus texts. The multiple random permutations are performed with $N = 1000$ iterations and ensure that the results are robust, comparable and representative (for a discussion see Ehret 2018, 2017). All complexity measurements reported in this paper are based on the arithmetic mean taken across these iterations of the compression technique, and were conducted in R (R Core Team 2019). We thus obtain one measurement each for overall, morphological and syntactic complexity per text in the database.

4.2 Subjectivity

Our take on subjectivity is empirically-oriented and defines subjectivity as lexico-grammatical expressions of opinion, evaluation, and subjectivity in a text. As described in Section 2, we employ four types of linguistic items as markers of subjectivity: (i) specific evaluative words (adjectives, nouns, verbs and adverbs); (ii) stance adverbs; (iii) modal verbs; and (iv) connectives which mark argumentative discourse relations. This subsection describes the sources which we tapped to obtain these markers and

provides some rationale for why these markers have been included in the analysis. In addition, corpus examples illustrate how these items and structures are used in context.

In order to obtain a comprehensive list of evaluative words, we draw on the *Semantic Orientation CALculator* (SO-CAL) which has previously been applied in the task of sentiment analysis (Taboada et al. 2011). SO-CAL comprises dictionaries of positive and negative adjectives, nouns, verbs and adverbs, amounting to a total of 5,042 evaluative words (2,252 adjectives, 1,142 nouns, 903 verbs, and 745 adverbs). The original dictionaries contain a fine-grained classification of sentiment which distinguishes between five valence categories indicating varying degrees of positive or negative sentiment.[3] Words can take valences ranging from −5 to 5, where higher positive or negative numbers are closer to the extremes of the evaluative continuum and thus indicate stronger evaluation. For example, *award-winning* has a valence of 5, *monstrosity* has a valence of −5, *informative* is 2, and *distasteful* is −3. In this paper, attention is restricted to evaluative words with a valence of −4/4 and −5/5 as markers of subjectivity because we are primarily interested in words that unambiguously (and strongly) convey evaluation and subjectivity (as in 2). Furthermore, the different word classes are collapsed into a simple binary categorisation of positive and negative evaluative words (regardless of word class) which we consider sufficient to capture subjectivity. Thus, a total of 674 evaluative words is analysed. These positive and negative words may, of course, be used with different polarity in context. For instance, a strongly negative phrase such as *not great at all* would be captured only in the positive adjective *great*. Similarly, sarcastic uses of words (*that's just great!*) are to be interpreted with the opposite polarity to that of their dictionary entry. Since we are concentrating on a few words, and their most typical usage, such concerns can be disregarded here, although they are of importance in practical applications such as sentiment analysis (Taboada 2016).

(2) a. Experts say that the post-generation is less polarized over Europe – their parents either *adored* or *detested* continental unification. (GLOB000020120622e86m00009.txt)

b. Funny, they never seem to find government candidates to be *unacceptable*. (comments2014.17829753.txt)

c. . . . I doubt anyone on earth would deny the *magnificence* of Rio or Vancouver or the *awesomeness* of Boston. (comments2013.11033911.txt)

Stance adverbs, as illustrated in 3 include adverbs that indicate epistemic, attitudinal or style stance, according to Biber and Finegan (1989) and Conrad and Biber (2000). Our list includes 78 adverbials and

[3]There are also two words, *better* and *worse*, with a valence of ***0***.

prepositional phrases (e.g., *undeniable, of course, to my surprise*) culled from Biber (1988) and Biber et al. (1999).

Modal verbs (see Example 4) are included as markers of subjectivity, as Hunston and Thompson (2000a) suggest. Our set of modal verbs includes the full list of modals (including abbreviated and negated forms) commonly found in English, i.e. *will, would, might, shan't, 'll*.[4]

(3) a. *Frankly*, as a US citizen if I were living abroad and earning my living outside the US the IRS could go pound sand for all I would care. (comments2013.6994760.txt)

b. I have zero issue with most immigration, *in fact*, most in canada are immigrants. (comments2016.30609024.txt)

(4) a. And who *will* be burning all that fossil fuel Mike. the wackos *won't* care but the law *will* by then. (comments2016.29805674.txt)

Finally, our list of markers of subjective discourse relations was compiled from the *Penn Discourse Treebank* annotation manual (Prasad et al. 2008) following the assumption that some discourse relations are subjective and more frequent in argumentative text. By discourse relations we mean those that connect discourse units and can be assigned labels such as Elaboration, Contrast or Summary (Mann and Thompson 1988; Prasad et al. 2008). Finding such relations automatically is still an open question, but it is well known that connectives and discourse markers are robust indicators of discourse relations. For that reason, we extracted all discourse markers listed as signals for six relations that are typically considered argumentative (Peldszus and Stede 2016; Moens et al. 2007): Cause, Comparison, Condition, Contrast, Evaluation and Explanation. The list of argumentative markers includes a total of 137 items comprising connectives such as *additionally*, *if* or *specifically* as well as multi-word markers such as *on the ground that*, *except insofar as*.

(5) a. Both Canada and the U.S. are high immigration countries *but because* Canada has the right model we have far more immigrants, *yet* far less social dislocation. (opinion2016.29294152.txt)

b. *But* the regime they conveniently champion is *ultimately* self-defeating and constraining as far as the broader goals of these sovereignty movements is concerned. (comments2014.20598876.txt)

[4]Semi-modals are not included because their semi-modal use cannot easily be distinguished from other uses (e.g., semi-modal *have to*, auxiliary verb *have* or *have* as a verb indicating possession). Furthermore, in this quantitative approach, different types of modality are not distinguished, as the primary interest is in modality in general as a means of conveying subjectivity.

Marker category	Number of markers
Evaluative words, positive	348
Evaluative words, negative	326
Stance adverbs	78
Modal verbs	18
Connectives	137
Total	907

Table 2. Number of markers per marker category

The complete set of subjectivity markers (see Table 2 for a tabular overview) analysed in this paper thus comprises 907 items. These markers of subjectivity were automatically extracted from our database using a custom-made python script. Subsequently, the frequencies of all markers were normalised per 1,000 word tokens as is customary in corpus linguistic frequency analyses in order to control for differences in text length.

5 Trees and forests

In order to analyse how text complexity and markers of subjectivity interact and contribute to the characterisation of distinct types of opinionated discourse, we utilise conditional inference trees (in the following simply referred to as trees), a type of classification and regression trees, and conditional random forests, an ensemble method for regression and classification based on an aggregate of single trees.[5] In general, trees and random forests are powerful tools for analysing and visualising complex interactions between many different predictor variables, as in our analysis with three complexity metrics and five different types of subjectivity markers, and assessing the importance of individual predictors (Strobl et al. 2009b). In particular, the combination of trees and random forests is well suited to explore our dataset as there are a couple of issues that cannot be easily handled by, e.g., standard general regression techniques (Tagliamonte and Baayen 2012, 161-165). First, text type is a categorical dependent variable with three levels (opinion articles vs. reader comments vs. news writing); second our predictor variables have different scales of measurement (complexity scores vs. relative token frequencies); and third, some of the predictors are somewhat correlated. These issues can be conveniently addressed by utilising conditional inference trees and random forests.

[5]As a detailed technical description of the methods is outside the scope of this paper, we refer the interested reader to Strobl et al. (2009b) and Hothorn and Zeileis (2006).

Our objective in this paper, however, is neither prediction nor classification. Rather, we aim to describe and characterise opinion and reader comments vis-à-vis arguably more objective newspaper writing and to explore the interplay between text complexity and markers of subjectivity. For this reason, the implementation of trees and forests and the statistical descriptions are kept straightforward. All statistics reported in this paper were conducted with the open-source software R (R Core Team 2019). For the implementation of trees and random forests we specifically draw on the R package *partykit* (Hothorn and Zeileis 2015) and largely follow the recommendations outlined in Strobl et al. (2009a). The R code, additional statistics, as well as the original dataset are available at `https://github.com/katehret/compinion`.

In this spirit, we construct a feature matrix with the individual texts as rows and each of the variables, i.e., text type, complexity measurements and frequency counts of the various subjectivity markers, as columns. The year in which a particular article or comment was posted is also included as predictor variable. In short, we analyse the descriptive power and interactions between nine predictor variables in 10,448 samples of naturalistic discourse.

In a first step, we construct a single tree to explore and visualise the complex interactions between the various predictor variables. In order to avoid overfitting, yet obtain a linguistically interpretable tree, we tune our tree model by splitting the data into training and test data. Subsequently, we compare how well the model performs in both datasets under different parameter settings. The accuracy of each parameter combination in the training and test dataset is reported in Table 4 (Appendix 7).[6] The "best" model is then selected based on the prediction accuracy in the test and training dataset as well as the linguistic interpretability of the tree. This tree has a prediction accuracy of 86% for the training data and 85% for the test data.

In a second step, we grow random forests to assess the importance of each individual predictor variable using the default settings for growing trees in the *partykit* package, which is recommended for unbiased forests. On a technical note, the data is split into a training and test dataset and random forests of varying sizes are grown to control for variable selection bias as well as to ensure that the results are robust. Specifically, we grow forests with $N = 500$, and $N = 1000$ and $N = 2000$ trees. To evaluate model performance, the prediction accuracy of the training and test set was calculated for each forest. The prediction accuracy across the differently sized forests (see Table 3 is virtually identical and the ranking of the predictors according to variable importance does not vary (for statistical details see `https://github.com/katehret/compinion`). In this paper, we therefore discuss the simplest

[6]We additionally test these settings across 20 random data splits to check whether the accuracy under the respective settings varies across different splits. We find that it does not.

model, i.e., the forest with $N = 500$ trees. The prediction accuracy for this forest is 93% for the training data and 92% for the test data.

Tree size	Training accuracy	Test accuracy
500	0.9265	0.9199
1,000	0.9271	0.9196
2,000	0.9260	0.9183

Table 3. Training accuracy and test accuracy for differently sized forests

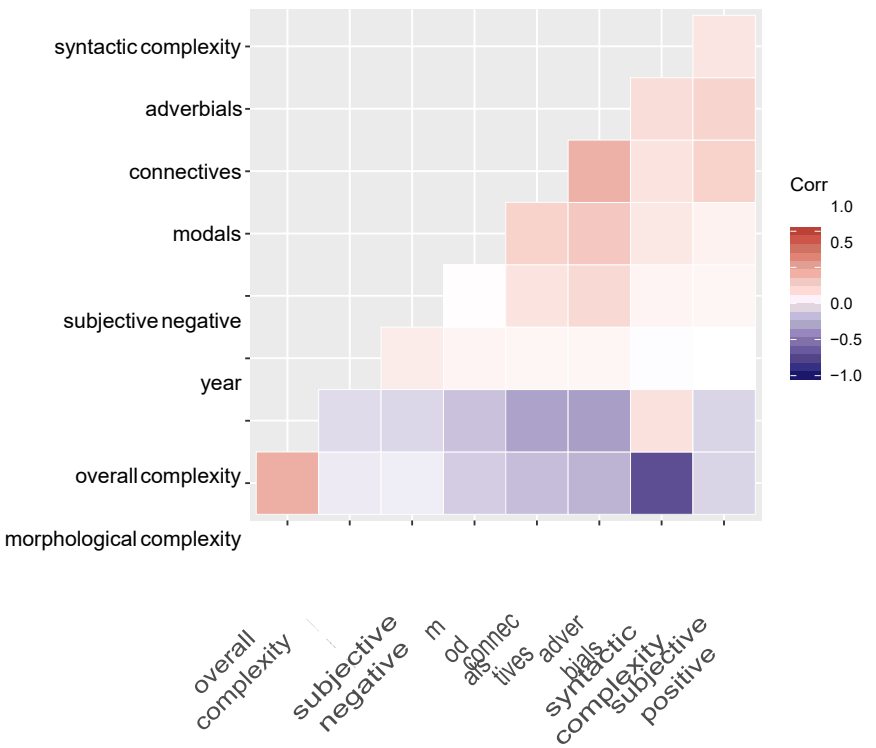

Figure 1. Pairwise correlations of predictor variables based on Pearson's correlation coefficient. Red squares indicate positive correlation; blue squares indicate negative correlation.

The importance of individual predictor variables is assessed by calculating the *conditional permutation-importance measure* (Strobl et al. 2008, 2007) for each predictor because it can reliably assess the importance of predictors which are correlated. As mentioned at the beginning of this section, some of the predictor variables are correlated (Figure 1). To be more precise, connectives and adverbials (Pearson's correlation coefficient $r = 0.37$), as well as morphological and overall complexity (Pearson's correlation coefficient $r = 0.38$) exhibit a medium positive correlation. In contrast, overall complexity moderately negatively correlates with both connectives and adverbials (Pearson's correlation coefficient $r = -0.37$ and $r = -0.39$, respectively). We also observe the typical trade-off between morphological and syntactic complexity (discussed e.g., in Ehret (2018); Ehret and Szmrecsanyi (2016b)) which is expressed by a high negative correlation (Pearson's correlation coefficient $r = -0.74$).

6 The interplay of complexity and subjectivity

In this section, the interactions between text complexity and subjectivity are unravelled and the importance of the individual predictor variables is discussed.

Figure 2 plots the selected tree model with branches ending in terminal nodes. At each node binary splits of the data are performed depending on the value of the splitting variable. These splits result in different branches, which, basically, group the data into consecutively more homogeneous subsets (Strobl et al. 2009b, 326). Interactions occur where a predictor appears only in one of the resulting two branches after a previous split (Strobl et al. 2009b, 239). Interactions in trees are therefore always *local*, i.e., interactions between particular predictors apply to a specific subset of the data rather than the entire dataset. In this model, interactions between only four of the nine predictor variables can be observed: overall complexity and syntactic complexity, overall and morphological complexity, and morphological complexity and adverbials.

In the interpretation of the tree, we will move along the edges (the lines connecting the nodes) from the root node at the top to the terminal nodes at the bottom. For example, we can identify an interaction between morphological complexity and the presence of adverbials only in texts with an overall complexity score of ≤ 0.7 (node 1), a syntactic complexity score of > 0.94 (node 2), an overall complexity score of > -1.96 (node 8) and a morphological complexity score of less than -0.96 (node 10). In other words, the effect of adverbials depends on the values of morphological, overall and syntactic complexity.

It can thus be observed that the data is split into two major branches based on the overall complexity of the texts: a first branch grouping opinionated discourse, i.e., comments and articles, together (ending in the terminal nodes 5, 6, 7, 9, 12, 14, 15, and 16), and a second branch of news articles (ending in the terminal nodes 18 and 19). Overall complexity is therefore the most distinguishing predictor in this

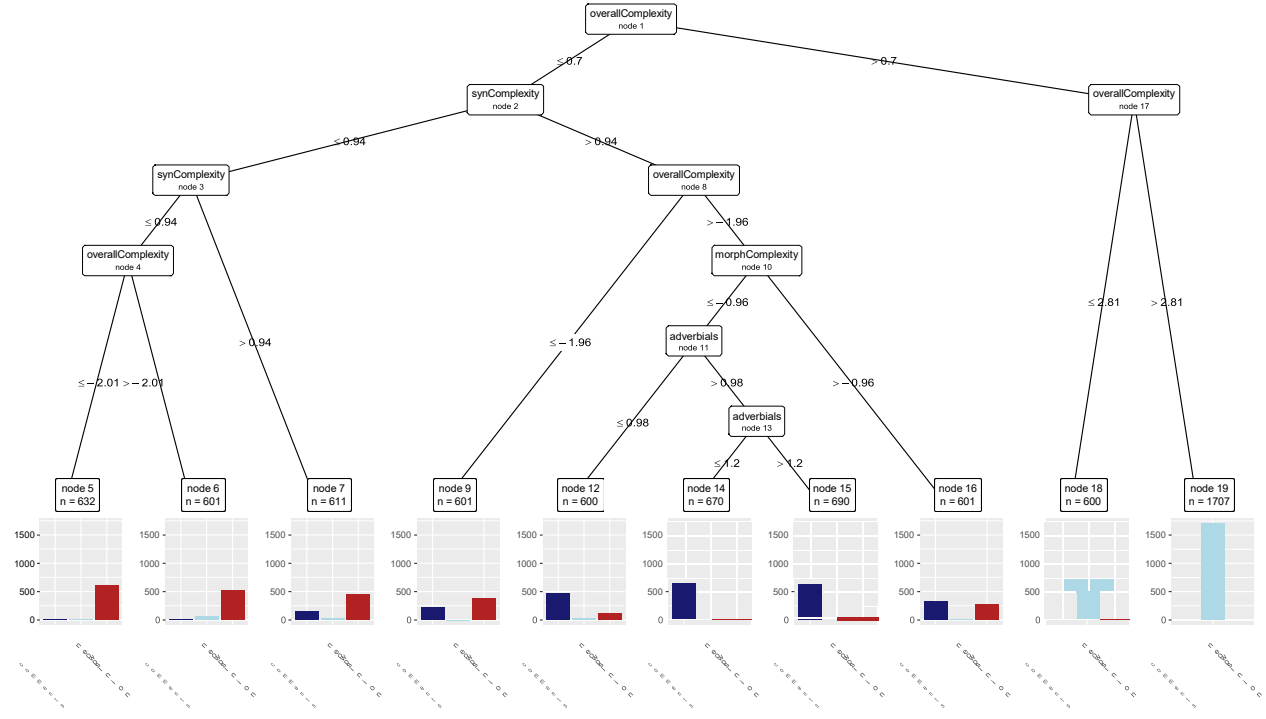

Figure 2. Tree model with 19 nodes displaying the complex interactions between overall complexity, syntactic complexity, morphological complexity, and adverbials. The values on the edges of the split nodes indicate at which complexity scores or frequencies the data was split. All values were rounded.

dataset (see also Figure 3). News articles are overall more complex (> 0.7, node 1) than comments and opinion articles (≤ 0.7, node 1). This finding is consistent with previous complexity research (Ehret 2018) in which newspaper writing, in particular broadsheet newspaper, was found to be overall the most complex register among a substantial number of written and spoken registers analysed. Furthermore, news writing is known to be characterised by a high type-token-ratio and information density, which are both indicators of lexical richness (Biber 1988). This dovetails with the high overall complexity scores observed in news texts, also bearing in mind that lexical richness and overall complexity are somewhat correlated (Ehret 2018, 14).

Turning to the left part of the plot, it can be seen that syntactic complexity roughly distinguishes between opinion articles and comments (node 2). Opinion articles tend to be less syntactically complex (≤ 0.94, node 2) than comments (> 0.94, node 2). Syntactic complexity is defined in terms of word order rigidity, i.e., more rigid word order is more complex while more varied word order counts as more syntactically simple. In this context, then, opinion articles are marked by less rigid and more varied word order than comments. This result matches with our intuitions, as opinion articles which are carefully edited should naturally contain more varied word order and syntactically elaborate structures (including patterns such as *Never was I more astonished*) than comments, which tend to be written 'on the fly'. This interpretation is supported by the split in node 3 which groups the texts further together based on their syntactic complexity. The terminal nodes 5 and 6 sample almost exclusively opinion articles, while in node 7 there are also some comments. Node 4 furthermore distinguishes between opinion articles of differing levels of overall complexity.

At this point, it is important to note that these observations do not mean that news writing is less syntactically complex than opinion articles. Rather, syntactic complexity is only a distinguishing criterion for this specific subsection of the data, i.e. for texts with an overall complexity of ≤ 0.7, as trees only depict local interactions.

Let us now focus on the centre of the tree. Node 8 splits opinionated discourse—the majority of comments, but also a couple of opinion articles—into texts with an overall complexity score equal to/less than and greater than -1.96, respectively. All of these texts are marked by very low overall complexity as indicated by the negative score. The articles and comments sampled in node 9 are the least overall complex texts in the dataset. At node 10, the data is divided according to morphological complexity: Most comments were categorised in the terminal nodes below. Therefore, morphological complexity is a distinctive characteristic for comments as opposed to opinion articles, such that most comments are less morphologically complex (≤ -0.96, node 10) than opinion articles (> -0.96, node 10). At the next two nodes below (nodes 11 and 13), an interaction between morphological complexity and subjectivity, i.e., stance adverbials, can be observed. Stance adverbials are an important characteristic for comments with a level of morphological complexity below -0.96. In other words, only in some comments, namely those

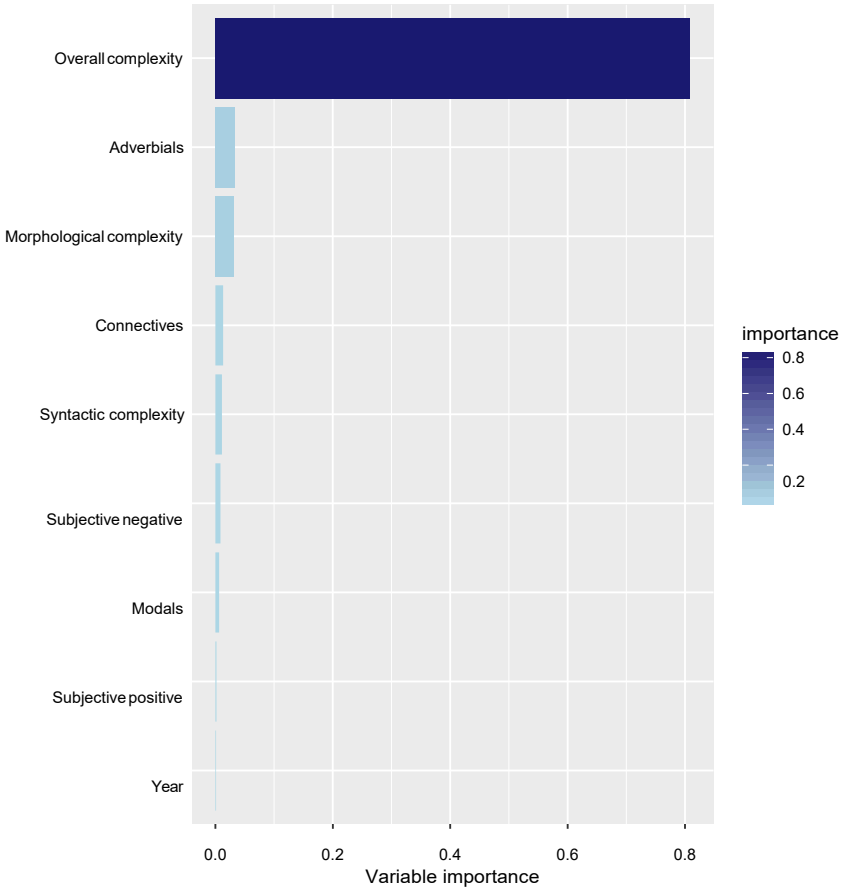

Figure 3. Conditional variable importance of the individual predictors.

of relatively low overall complexity, high syntactic complexity and low morphological complexity, are adverbials an important predictor. On an interpretational plane, these interactions indicate that comments tend to follow a comparatively rigid word order and make frequent use of stance adverbials such as *frankly, admittedly, seriously* to mark personal opinion and structure discourse.

In summary, the tree visualises how different levels of text complexity locally interact and clearly shows that, in general, overall text complexity is the major characteristic distinguishing between opinionated discourse and news articles. Furthermore, syntactic complexity distinguishes between the two text types opinion articles and comments. The ranking of variable importance presented in Figure 3 largely corroborates the findings described above. Overall complexity is by far the most important predictor, followed, at a considerable distance by stance adverbials and morphological complexity. This

reflects the local interaction between morphological complexity and adverbials which are distinctive predictors in the categorisation of comments. It is surprising that syntactic complexity is only on rank five, although it served as splitting variable in the tree at various levels. This might be explained by the strong negative correlation between syntactic and morphological complexity, which can only be accounted for in the calculation of the conditional permutation-importance. The ranking of variable importance, however, clearly shows that subjectivity markers other than stance adverbials barely matter in the classification of opinion articles, comments, and news articles. Neither the tree nor the variable importance ranking can illuminate this issue. In search for an explanation, we therefore survey, *post hoc*, the distribution of subjectivity markers in the three different text types. Figure 4 plots the percentage of each type of subjectivity marker in news articles, opinion articles and comments (see Appendix, Table 5, for raw frequencies). Strikingly, the different types of subjectivity markers, including positive and negative evaluative words, occur with roughly the same percentage in all three text types, irrespective of whether the texts are opinionated or supposedly objective.

In short, the various markers of subjectivity are not a distinctive characteristic in opinion articles, comments and news articles, and therefore do not serve as splitting variables in the tree nor are they ranked as important predictors in the variable importance ranking. Only in interaction with the various levels of text complexity do adverbials of stance make a contribution to the categorisation of the three text types at a local level, i.e., only for a small subset of the data.

7 Concluding remarks

We have offered a quantitative corpus-based analysis of subjectivity and complexity in opinionated discourse and news articles. By comparing three different text types (general news articles, opinion pieces and online comments), we explore the relationship between text complexity and linguistic markers of subjective and evaluative language. To capture text complexity, we use a holistic, information-theoretic measure of complexity which has previously been shown to be a reliable tool in text type analyses (Ehret 2018). We approach subjectivity from a lexical standpoint, compiling a list of subjectivity markers from the literature (Biber et al. 1999; Taboada et al. 2011), which comprise highly positive and negative words, stance adverbials, modals and argumentative connectives. The analyses include conditional inference trees for visualising complex interactions together with random forests, an established method for investigating the predictive power of individual variables.

Our main finding is that text complexity, and specifically overall complexity is by far the most powerful predictor distinguishing between opinionated discourse on the one hand, and news articles on the other hand: opinionated discourse is less overall complex than news articles. This dovetails with previous research on the complexity of different British English registers which reports that newspaper writing was

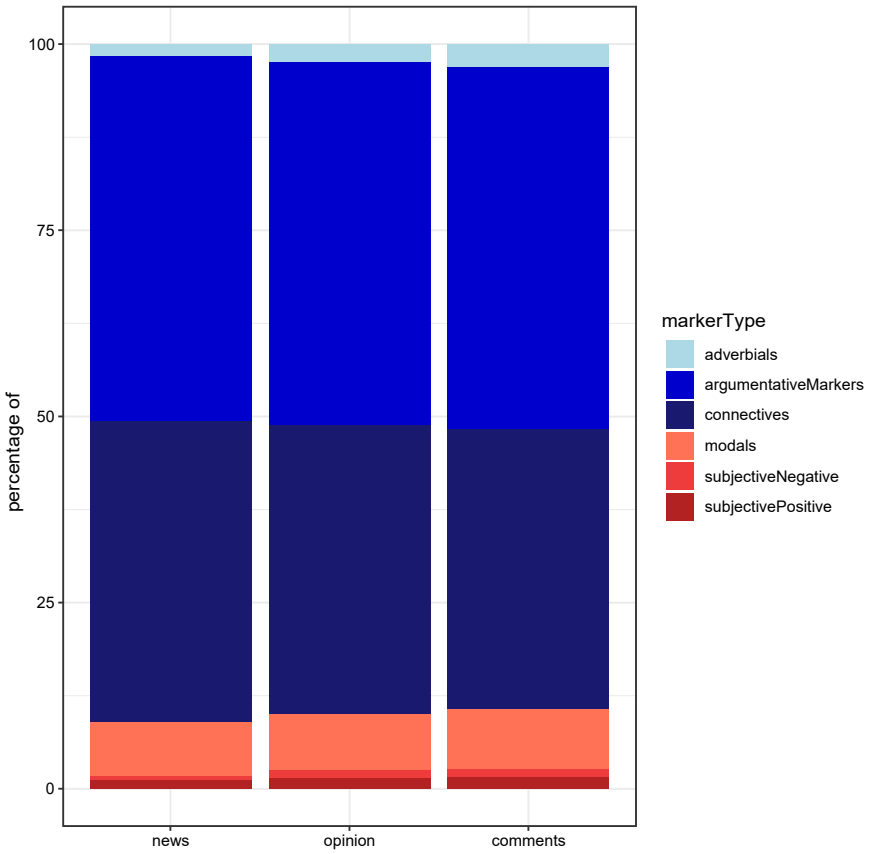

Figure 4. Distribution (in percentage) of the different subjectivity markers per text type in the entire dataset

one of the most overall complex registers (Ehret 2018). On a methodological plane, this finding highlights that overall text complexity may be a useful tool for text type classification and characterisation, with a potential of further application, e.g., in fake news identification and other applications that involve register classification. We emphasise that our approach is lexically-based and aggregate in nature. This might be considered a caveat, in that it does, on the one hand, not capture more subtle structures and patterns (Hunston 2011) that may be deployed in the expression of subjectivity such as modal-like patterns of the type "noun *of* + verb-*ing*" (*the advisability of, the likelihood of, the difficulties of*), and, on the other hand, not distinguish e.g., between different types of stance adverbials. That said, the strength of this approach lies in its applicability to large-scale naturalistic datasets whose (qualitative) analysis would otherwise be

extremely time consuming, empirically expensive and therefore outside the scope of traditional discourse analysis.

We further find that stance adverbials are the only markers of subjectivity analysed in this paper which have a distinguishing nature vis-à-vis the complex interactions of text complexity, yet only in a small subset of comments. There are several possible explanations for this. We pointed out earlier that less morphologically complex texts show this interaction, perhaps due to relatively rigid word order and comparatively simpler structures that rely on stance adverbials to convey opinion. Another explanation may be that (stance) adverbials are present overall in opinionated texts, but perhaps more saliently so in comments, suggesting that stance adverbials are a text type specific stylistic device (cf. Ehret and Taboada 2019).

Our work has some interesting theoretical implications. We analysed a large number of well-established subjectivity markers in naturalistic discourse. Contrary to our expectations, we found practically the same proportion of these markers across the three text types. More to the point, opinionated discourse, represented here by opinion articles and reader comments, does not contain more subjectivity markers than the arguably more objective news texts. This is particularly striking in regard to positive and negative evaluative words which, in our dataset, are the clearest and strongest markers of evaluation and expression of subjectivity. This raises the question of whether 'objective' discourse exists at all, or whether any text produced by language users is invariably subjective in the sense of Lyons (1981). Expressing one's opinion, or point of view, is one of the most fundamental functions of human language, one which language users might not be able to avoid. We thus conclude that naturalistic discourse, be it labelled opinionated or objective, always conveys to some extent the opinion, stance, or evaluation of the language user. As has been pointed out by a reviewer, the opinion, evaluation or stance conveyed in a text might not always be the writer's own opinion and we acknowledge that subjectivity attribution is outside the scope of our approach which is lexically-based and quantitative. Naturally, it also goes without saying that a different, more fine-grained distinction between positive and negative evaluative words might have led to different results and that the findings presented here should be further corroborated with different data covering more varied text types. We would furthermore like to point out that sarcastic and ironic uses of evaluative words (see Section 4.2) are not captured here. Nevertheless, our findings are supported by recent qualitative research investigating the connection between emotion and evaluation in discourse (Mackenzie and Alba-Juez 2019, see also Alba-Juez and Thompson 2014). In particular, Alonso Belmonte (2019) describes how journalists use emotional-evaluative language in a major Spanish newspaper. Furthermore, our results dovetail with a substantial body of literature on changes in journalism, or media discourse more generally, which also discuss the use of involved, informal or emotional language (e.g. Fairclough 1995; Zelizer 2009). As a result of such linguistic

changes, the distinctions between different news types, such as opinion and factual news, are no longer clear-cut (e.g. Peters 2011, 298).

However, the finding that news texts employ roughly as many lexical markers of subjectivity as opinion articles and reader comments, could also be seen as part of a more general (linguistic) development, namely, informalisation (Fairclough 1992) or colloquialisation (Hundt and Mair 1999) which refers to an increasing informal, colloquial language use in written discourse, most evident in newspaper writing (see also Peters 2011).

Finally, this paper advances the state-of-the-art in discourse analysis by showcasing how quantitative methods can be applied to explore discourse analytic concepts, such as subjectivity, in otherwise inaccessible datasets.

Funding

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.

Acknowledgements

Thanks to the members of the Discourse Processing Lab at Simon Fraser University, and especially to Laurens Bosman for help with data manipulation and feature extraction. We are also grateful for feedback from the audience of the 2019 Conference of the Canadian Association of Applied Linguistics in Vancouver. Many thanks to Laura Alba Juez, Monika Bednarek, and Bethany Gray, for feedback on a draft of this manuscript. Angela Chen, Haoyao Ruan and Ian Berkovitz from the Statistical Consulting Group at SFU provided feedback on the conditional inference tree analyses.

References

Alba-Juez L and Thompson G (2014) The many faces and phases of evaluation. In: Thompson G and Alba-Juez L (eds.) *Evaluation in Context*. Amsterdam: John Benjamins, pp. 3–23.

Alonso Belmonte I (2019) Victims, heroes and villains in newsbites. In: Mackenzie JL and Alba-Juez L (eds.) *Emotion in Discourse*. Amsterdam: John Benjamins, pp. 335–356.

Arends J (2001) Simple grammars, complex languages. *Linguistic Typology* 5(2/3): 180–182.

Baechler R and Seiler G (eds.) (2016) *Complexity, Isolation, and Variation*. Berlin, Boston: De Gruyter. ISBN 978-3-11-034896-5. URL http://www.degruyter.com/view/books/9783110348965/9783110348965-004/9783110348965-004.xml.

Baerman M, Brown D and Corbett GG (eds.) (2015) *Understanding and measuring morphological complexity*. New York: Oxford University Press.

Bednarek M (2006) *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*. London: Continuum.

Benveniste E (1966) *Problèmes de linguistique générale*. Paris: Gallimard.

Biber D (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

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

Biber D and Gray B (2010) Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9: 2–20.

Biber D and Gray B (2016) *Grammatical Complexity in Academic English. Linguistic Change in Writing*. Studies in English Language. Cambridge: Cambridge University Press.

Biber D, Gray B and Poonpon K (2011) Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly* 45(1): 5–35.

Biber D, Johansson S, Leech G, Conrad S and Finegan E (1999) *Longman Grammar of Spoken and Written English*. Harlow, Essex: Pearson Education.

Conrad S and Biber D (2000) Adverbial marking of stance in speech and writing. In: Hunston S and Thompson G (eds.) *Evaluation in Text: Authorial Distance and the Construction of Discourse*. Oxford: Oxford University Press, pp. 56–73.

Crystal D (1987) *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

Du Bois JW (2007) The stance triangle. In: Englebretson R (ed.) *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*. Amsterdam: Benjamins, pp. 139–182.

Edwards J (1994) *Multilingualism*. London: Penguin.

Ehret K (2017) *An information-theoretic approach to language complexity: variation in naturalistic corpora*. PhD dissertation, Freiburg.

Ehret K (2018) An information-theoretic view on language complexity and register variation: Compressing naturalistic corpus data. *Corpus Linguistics and Linguistic Theory* (Ahead of print).

Ehret K and Szmrecsanyi B (2016a) Compressing learner language: an information-theoretic measure of complexity in SLA production data. *Second Language Research (Sage Online First)* URL http://journals.sagepub.com/doi/abs/10.1177/0267658316669559.

Ehret K and Szmrecsanyi B (2016b) An information-theoretic approach to assess linguistic complexity. In: Baechler R and Seiler G (eds.) *Complexity and Isolation*. Berlin: de Gruyter, pp. 71–94.

Ehret K and Taboada M (2019) Are online news comments like face-to-face conversation? A multi-dimensional analysis of an emerging register. *Register Studies* 2(1): 1–36.

Englebretson R (ed.) (2007) *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*. Amsterdam: John Benjamins.

Fairclough N (1992) *Discourse and social change*. Cambridge: Polity press.

Fairclough N (1995) *Media Discourse: Voices*. London: Edward Arnold.

Halliday MAK (1985) *An Introduction to Functional Grammar*. 1st edition. London: Arnold.

Hothorn K Torsten tnd Hornik and Zeileis A (2006) Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics* 15(3): 651–674.

Hothorn T and Zeileis A (2015) partykit: A Modular Toolkit for Recursive Partytioning in R. *Journal of Machine Learning Research* 16: 3905–3909.

Hundt M and Mair C (1999) 'Agile' and 'uptight' genres: the corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4: 221–242.

Hunston S (2011) *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*. New York: Routledge.

Hunston S and Thompson G (2000a) Evaluation: An introduction. In: Hunston S and Thompson G (eds.) *Evaluation in Text: Authorial Distance and the Construction of Discourse*. Oxford: Oxford University Press, pp. 1–27.

Hunston S and Thompson G (eds.) (2000b) *Evaluation in Text: Authorial Distance and the Construction of Discourse*. Oxford: Oxford University Press.

Juola P (1998) Measuring linguistic complexity: the morphological tier. *Journal of Quantitative Linguistics* 5(3): 206–213.

Juola P (2008) Assessing linguistic complexity. In: Miestamo M, Sinnemäki K and Karlsson F (eds.) *Language Complexity: Typology, Contact, Change*. Amsterdam, Philadelphia: Benjamins, pp. 89–107.

Kolhatkar V, Wu H, Cavasso L, Francis E, Shukla K and Taboada M (2020) The SFU Opinion and Comments Corpus: A Corpus for the Analysis of Online News Comments. *Corpus Pragmatics* 4: 155–190.

Kortmann B and Szmrecsanyi B (eds.) (2012) *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Lingua & Litterae. Berlin/Boston: Walter de Gruyter.

Kusters W (2003) *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*. Utrecht: LOT.

Langacker RW (1985) Observations and speculations on subjectivity. In: Haiman J (ed.) *Iconicity in Syntax*. Amsterdam and Philadelphia: John Benjamins, pp. 109–150.

Li M and Vitanyi PMB (1997) *An introduction to Kolmogorov complexity and its applications*. New York: Springer.

Lupyan G and Dale R (2010) Language Structure Is Partly Determined by Social Structure. *PLoS ONE* 5(1): 1–10. DOI:10.1371/journal.pone.0008559.

Lyons J (1981) *Language, Meaning and Context*. London: Fontana.

Mackenzie JL and Alba-Juez L (eds.) (2019) *Emotion in Discourse*. Amsterdam: John Benjamins.

Mann WC and Thompson SA (1988) Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8(3): 243–281.

Martin JR (1992) *English Text: System and Structure*. Amsterdam and Philadelphia: John Benjamins.

Martin JR and White PRR (2005) *The Language of Evaluation*. New York: Palgrave.

McWhorter J (2001) The world's simplest grammars are creole grammars. *Linguistic Typology* 6: 125–166.

McWhorter J (2012) Complexity hotspot. In: Kortmann B and Szmrecsanyi B (eds.) *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*, Linguae & Litterae. Berlin: Walter de Gruyter, pp. 243–246.

Moens MF, Boiy E, Palau RM and Reed C (2007) Automatic detection of arguments in legal texts. In: *Proceedings of the 11th International Conference on Artificial Intelligence and Law*. Stanford, California: ACM, pp. 225–230.

Mufwene S, Coupe´ C and Pellegrino F (2017) *Complexity in language: Developmental and evolutionary perspectives*. Cambridge/New York: Cambridge University Press.

Peldszus A and Stede M (2016) Rhetorical structure and argumentation structure in monologue text. In: *Proceedings of the 3rd Workshop on Argument Mining, ACL*. Berlin, pp. 103–112.

Peters C (2011) Emotion aside or emotional side? Crafting an 'experience of involvement' in the news. *Journalism* 12(3): 297–316.

Prasad R, Lee A, Dinesh N, Miltsakaki E, Campion G, Joshi AK and Webber B (2008) Penn Discourse Treebank version 2.0. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. Marrakesh, Morocco, pp. 2961–2968.

Prasad R, Miltsakaki E, Dinesh N, Lee A, Joshi AK, Robaldo L and Webber B (2007) The Penn Discourse Treebank 2.0 Annotation Manual. Technical report, University of Pennsylvania.

R Core Team (2019) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Sanders T, Spooren W and Noordman L (1993) Coherence relations in a cognitive theory of discourse representation. *Cognitive Linguistics* 4(2): 93–133.

Shosted RK (2006) Correlating complexity: A typological approach. *Journal of Linguistic Typology* 10: 1–40. Strobl

C, Boulesteix AL, Kneib T, Augustin T and Zeileis A (2008) Conditional Variable Importance for Random Forests. Technical Report 23, Department of Statistics, University of Munich. URL `http://www.slcmsr.net/boulesteix/papers/condimp.pdf`.

Strobl C, Boulesteix AL, Zeileis A and Hothorn T (2007) Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics* 8(25): 1–21. DOI:doi:10.1186/1471-2105-8-25.

Strobl C, Hothorn T and Zeileis A (2009a) Party on! *The R Journal* 1(2).

Strobl C, Malley J and Tutz G (2009b) An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods* 14(4): 323–348.

Szmrecsanyi B (2009) Typological parameters of intralingual variability: Grammatical analyticity versus syntheticity in varieties of English. *Language Variation and Change* 21(3): 319–353.

Szmrecsanyi B and Kortmann B (2009) The morphosyntax of varieties of English worldwide: a quantitative perspective. *Lingua* 119(11): 1643–1663.

Taboada M (2016) Sentiment-analysis: An overview from linguistics. *Annual Review of Linguistics* 2: 325–347.

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

Tagliamonte S and Baayen HR (2012) Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language variation and change* 24(2): 135–178.

Trnavac R, Das D and Taboada M (2016) Discourse relations and evaluation. *Corpora* 11(2): 169–190.

Trnavac R and Taboada M (2012) The contribution of nonveridical rhetorical relations to evaluation in discourse. *Language Sciences* 34(3): 301–318.

van Dijk TA (1979) Pragmatic connectives. *Journal of Pragmatics* 3: 447–456. Van

Dijk TA (1997) *Discourse as Structure and Process*, volume 1. London: Sage.

Wiebe J (2000) Learning subjective adjectives from corpora. In: *Proceedings of 17th National Conference on Artificial Intelligence (AAAI)*. Austin, Tx, pp. 735–740.

Wiebe J and Riloff E (2005) Creating subjective and objective sentence classifiers from unannotated texts. In: *Proceedings of Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005)*. Mexico City, Mexico, pp. 1–12.

Wiebe J, Wilson T, Bruce R, Bell M and Martin M (2004) Learning subjective language. *Computational Linguistics* 30(3): 277–308.

Zelizer B (ed.) (2009) *The Changing Faces of Journalism : Tabloidization, Technology and Truthiness*. London and New York: Routledge.

Appendix

Mincriterion	Minbucket	Maxsurrogate	Training accuracy	Test accuracy
0.90	600	0	0.86	0.85
0.95	600	0	0.86	0.85
0.99	600	0	0.86	0.85
0.90	800	0	0.85	0.85
0.95	800	0	0.85	0.85
0.99	800	0	0.85	0.85
0.90	1,000	0	0.84	0.83
0.95	1,000	0	0.84	0.83
0.99	1,000	0	0.84	0.83
0.90	600	3	0.86	0.85
0.95	600	3	0.86	0.85
0.99	600	3	0.86	0.85
0.90	800	3	0.85	0.85
0.95	800	3	0.85	0.85
0.99	800	3	0.85	0.85
0.90	1,000	3	0.84	0.83
0.95	1,000	3	0.84	0.83
0.99	1,000	3	0.84	0.83
0.90	600	6	0.86	0.85
0.95	600	6	0.86	0.85
0.99	600	6	0.86	0.85
0.90	800	6	0.85	0.85
0.95	800	6	0.85	0.85
0.99	800	6	0.85	0.85
0.90	1,000	6	0.84	0.83
0.95	1,000	6	0.84	0.83
0.99	1,000	6	0.84	0.83

Table 4. Training accuracy and test accuracy for different parameter settings used in tree model selection. Mincriterion: value of the test statistic; minbucket: the minimum sum of weights in a terminal node; maxsurrogate: the number of alternative predictors to consider at each split.

Marker Type	News	Opinion	Comments
Adverbials	8,387	13,323	144,851
Connectives	228,177	218,333	1,817,503
Modals	40,257	42,322	386,113
Argumentative	276,821	273,978	2,348,467
Subjective positive	6,758	8,615	76,043
Subjective negative	3,631	5,639	53,536

Table 5. Raw frequencies of subjectivity and argumentation markers across the three text types.