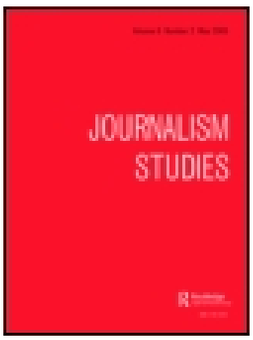

Journalism Studies

ISSN: 1461-670X (Print) 1469-9699 (Online) Journal homepage: http://www.tandfonline.com/loi/rjos20

Analysis of Linguistic Complexity in Professional and Citizen Media

Petro Tolochko & Hajo G. Boomgaarden

To cite this article: Petro Tolochko & Hajo G. Boomgaarden (2017): Analysis of Linguistic Complexity in Professional and Citizen Media, Journalism Studies, DOI: 10.1080/1461670X.2017.1305285

To link to this article: https://doi.org/10.1080/1461670X.2017.1305285

Published online: 29 Mar 2017.

ANALYSIS OF LINGUISTIC COMPLEXITY IN PROFESSIONAL AND CITIZEN MEDIA

Petro Tolochko and **Hajo G. Boomgaarden**

Structural linguistic characteristics are an important aspect of written communication. Previous research shows that linguistic complexity plays an important role in how people process information. With increasing popularity and readership of citizen journalism, questions of how structurally different this medium is from its professional counterparts and how this difference potentially affects readers become salient. Using automated content analysis methods, the present study investigates the differences in linguistic complexity across various citizen and professional journalism outlets. The analysis shows that the patterns of presenting political information across various media are different. These findings have direct implications for various branches of communication and journalism studies such as the knowledge gap hypothesis, language expectancy theory, and credibility research.

KEYWORDS citizen journalism; complexity; cross-media comparison; journalism; quality newspapers; tabloid newspapers

Introduction

Today "*citizen journalism*," "grassroots journalism or "*participatory journalism*" is an extremely widespread phenomenon. The proliferation of digital technologies, rapid growth of internet penetration, and the ease of access to a huge corpus of information, enables people of various professions to analyze, produce, and share political news content without the necessary condition of working at a traditional news media outlet. Citizen journalism is broadly defined as the participation of citizens in the political process by means of generation and dissemination of political information (e.g., Bowman and Willis 2003). However, the practice of citizens engaging in journalism may take various forms and mean different things depending on the context (Lasica 2003). A broader definition involves any action that disseminates political information, sometimes going as far as to term re-posts and social media "shares" as citizen journalism, while the more strict notions of citizen journalism refer to the creation of political content, such as commenting, writing a blog, etc. (e.g., Goode 2009). However, almost all of the definitions, regardless of how "strict" they are, share a few common threads. Citizen journalism is viewed as being an alternative to the mainstream media (Goode 2009), having a user-centric (as opposed to corporate) nature (Lewis, Kaufhold, and Lasorsa 2010), and being, at least to a larger extent, created by non-professionals.

The formal linguistic aspect of such communication, however, does not receive a lot of attention from journalism research in general, and definitely not in comparative accounts of traditional and newer forms of journalism. Language, nevertheless, plays a central role in politics, since it constituted simultaneously the ultimate medium through which political

Journalism Studies, 2017
http://dx.doi.org/10.1080/1461670X.2017.1305285

processes are communicated and the message of this communication. The structure of language in general, and language complexity in particular, affects the perceived credibility of a message (Jucks and Paus 2012). Political actors' credibility, for example, is very much dependent on how they structure their utterances (Wodak 1989, 115–118). Previous research has also shown that different linguistic characteristics of a message affect the way it is perceived by the audience (e.g., Kleinnijenhuis 1991). Structural characteristics of the message are thus a potentially important—but often overlooked—aspect of the product of journalists. The decrease of trust in the media for the past few decades (Moy and Scheufele 2000; Peters and Broersma 2013), together with the constantly multiplying modes of information dissemination and citizen engagement in the production process, make the question of analysis of various media, their structure, and effects on political processes as relevant as ever.

Political blogs are chosen as a prime example of citizen journalism to compare with traditional newspaper coverage. Political blogs are mostly user-centric and viewed in opposition to the professional, corporate media (e.g., Goode 2009). Although there are many manifestations of citizen journalism, political blogs resemble the conventional newspaper columns much more than any other form of citizen journalism, which makes them an appropriate candidate for a comparative structural analysis. It is important to understand structural differences in language complexity to then further understand the credibility competition between traditional and new news media.

Linguistic complexity is not extensively analyzed in the social sciences. The studies that have been interested in this topic, however, show evidence of its effects on political processes. One of the pioneering studies that investigated both textual and news complexity was Kleinnijenhuis' (1991) research that focused on the knowledge gap hypothesis (e.g., Tichenor, Donohue, and Olien 1970). This study concluded that newspaper complexity plays an important role in explaining the knowledge gap hypothesis. Other political processes, like political information recall and factual knowledge, are also affected by information complexity (e.g. Eveland and Cortese 2004). Thus, complexity plays an important role in how people process information. The current study attempts to comparatively analyze journalistic outlets, specifically, citizen journalism, quality newspapers, and tabloid newspapers, from a structural perspective, and tries to uncover the differences in linguistic complexity within these types of media. This project aims to answer the following research question:

> **RQ1:** Do professional newspapers (including quality and tabloid newspapers) differ in terms of structural linguistic complexity from citizen journalism media?

Theoretical Framework

Citizen journalism evokes a substantial amount of interest from journalism and political communication scholars since it effectively started a paradigm shift in media processes, and changes the framework of how information generation and dissemination occur. Citizen journalism blurs the lines between the creators and the consumers of political information. Traditional media rely on a hierarchical, top-down model, where information is created by an organization and "passed down" to the consumers. Citizen journalism removes the vertical distance between the creators and the consumers of information. Comparative research is a very popular approach when investigating the underlying mechanisms of citizen journalism. Since most of the theoretical tenets of journalism studies were

developed analyzing traditional media (e.g., gatekeeping research, credibility research, perceived roles, etc.), researchers nowadays apply these ideas to citizen journalism and cross-reference the differences and the similarities of these media models (see, e.g., Hanitzsch 2007; Reese et al. 2007).

Credibility Research, Traditional and Citizen Media

Research on the perceived credibility of (citizen) journalism is a relatively new branch and is gathering momentum. Credibility research is concerned with two aspects of communication: *source credibility* and *medium credibility* (Kiousis 2001). Source credibility research focuses on the characteristics of the communicator, and how they can affect the processing of the message (e.g., Hovland and Weiss 1951; Pornpitakpan 2004). The role of the communicator can be taken by an individual (e.g., journalist) or a group or institution (e.g., newspaper, publishing house). Conversely, medium credibility analyzes and focuses on the *channels* that are used to transmit information (e.g., print media, television or internet). The concept of a source in citizen journalism is somewhat of a debate; since citizen journalism is decentralized, it is often hard to pinpoint where exactly the information is originated, or who is actually responsible for the dissemination of this information. In traditional media models, journalists (who are themselves perceived as sources by the public) often rely on limited amounts of trusted sources to obtain information, therefore reducing the amount of resources to verify information (Williams and Delli Carpini 2004). Citizen journalism alters the established journalistic routines—ordinary citizens are more and more accepted as information sources by professional journalists and are considered among the most important sources of information (De Keyser, Raeymaeckers, and Paulussen 2011).

Credibility, however, is associated not only with who is the source, or how the message was transmitted, but also with the content and the structure of the message itself. There are two main arguments about how language complexity may influence credibility: *expertise*, on the one hand, and *comprehensiveness*, on the other. Studies on the effects of these two distinct dimensions produce mixed results (Blobaum 2016, 230–232). Technicality in language, for example, may come off as being complex and therefore increase the perceived credibility of the message (Jucks and Paus 2012). Moreover, students perceive academics that use complex language as more credible and their complicated explanations a function of the academic's expertise (Thiebach, Mayweg-Paus, and Jucks 2015). Contradictory to the aforementioned studies, however, Scharrer et al. (2012) have shown that written arguments are perceived as being more believable when they use less technical language; thus less complex—more credible.

These studies, while ending up with contradictory results, do agree on one thing—there are social expectations and assumptions as to what kind of language a certain social group should use, and when these assumptions are violated, the perceived credibility of the information changes. The social groups that are required to meet expectations include, but are not limited to, academics (Thomm and Bromme 2012), doctors (Blobaum 2016, 230), political actors (Pfau, Parrott, and Lindquist 1992; Dillard and Pfau 2002), and media actors (Pfau, Parrott, and Lindquist 1992; Burgoon, Denning, and Roberts 2002).

Language expectancy theory (LET), developed by Burgoon, Denning, and Roberts (2002), addresses the effects of the linguistic structure on the persuasiveness of the message. The main idea behind LET is that people develop social and cultural expectations

regarding the use of language, and these expectations further lead to classifying the message as being persuasive or not (Burgoon and Miller 1985). The linguistic patterns that are expected from a social group may be of a different nature: it could be aggressiveness (Pfau, Parrott, and Lindquist 1992), humor, irony, praise (e.g., Averbeck 2010) or the complexity of the language. A study of the credibility of online reviews, investigating the effects of lexical and semantic complexity, hypothesized that more complex messages would commit positive expectancy violations and, therefore, increase the credibility of the review (Jensen et al. 2013). The authors argued that, since the average semantic and lexical complexity in the review messages is relatively low, having a message with high complexity would seem more competent and, by extension, more credible in the eyes of the reader. Other studies have shown that the strong normative expectations of language use are not based on an individual level, but rather on a group or organizational level, for example, scientists and doctors are expected to conform to different linguistic patterns than manual labour workers (e.g., Buller et al. 2000; Jensen et al. 2013; Paus and Jucks 2011). Despite its advantages, LET is not applied extensively in journalism research. Interestingly enough, Burgoon initially thought of media analysis as being one of the main applications of LET (Burgoon, Denning, and Roberts 2002).

Research on journalistic roles is a branch that could greatly benefit from the incorporation of LET and linguistic complexity. For example, analyzing the difference in how professional and citizen journalisms see themselves, it was found that both groups take on a role of an interpreter, providing analysis and interpretation of complex problems and translating them for the public to read (Nah and Chung 2012). While both groups think of themselves virtually the same in this regard, knowing the linguistic complexity of their "interpretations" may provide further answers about their perceived roles, and how they fulfill these roles. Journalists also differ in what they think is credible—professional newspaper journalists often rate online news as being less credible than print newspapers because they are concerned that the proper journalistic norms and routines are not as rigorous (or absent altogether) from the new online environment (Singer 2004; O'Sullivan and Heinonen 2008).

Professional newspaper journalists, however, differ in their role perceptions among themselves as well. Tabloid media journalists, being more market-oriented, are placing higher values on the coverage of private sphere topics, entertainment, and human-interest stories, and are making less emphasis on investigative journalism, while quality journalists regard public sphere topics and investigative journalism as having high value (Beam 2003). Thus, market-oriented journalists do not necessarily see themselves as the ones who should portray the news as objectively as possible, but rather to provide the reader with the most interesting story. Tabloid journalists often see themselves employing a skill-set different from that of their colleagues in the quality media, as well as having a different view on what "quality" and "truth" in journalism actually mean (Deuze 2005).

It could be argued that linguistic structure (including textual complexity) has either a direct or an indirect connection with professional journalist routines—e.g., rigorous formatting of language, adherence to predefined style, barring the use of emotional words, colloquialism, etc. As much as journalists build up expectancies towards the structure of the news (Singer 2004), readers may also develop some preconceived notions as to how different media should be structured—e.g., they may think that professional quality newspapers should employ a sterile, objective language and articles should be rich in information, tabloid newspapers should employ captivating language and story structure, while

citizen journalism news should be easy to read and approachable. Also, as was already discussed, these expectations have direct implications for the trustworthiness and the credibility of news.

Complexity, Traditional and Citizen Media

Linguistic complexity is an important factor across different disciplines, and more importantly, across various branches of communication research. Understanding how complexity affects different media may open a way for a plethora of new research advancing already exciting theories. Combining LET with previous research on credibility and research that shows the effects of information complexity on various political processes, like political factual knowledge (e.g., Eveland and Cortese 2004), political socialization, acquisition of political information or political information retainment (e.g., Kleinnijenhuis 1991; Eveland and Dunwoody 2001), comparing the linguistic complexity of citizen journalism and traditional media may have implications for future research. The branch that seems most likely to benefit from knowing the effects of text complexity is credibility research. As discussed previously, complexity *directly* influences credibility. Yet, despite this fact, there is an evident lack of research concerned with complexity and *journalistic/news credibility*. First of all, does the language employed by different media actually differ? Does more complicated, technical, convoluted political language increase the perceived credibility of the medium, thereby contributing to more trust in the journalistic product? Or does explaining politics in layman's terms, therefore rendering it more comprehensible to citizens, imply a better understanding of a subject, and thus a better approach to formulate a credible message?

Moreover, textual complexity may be of interest to researchers occupied with LET, especially in journalism research. If there is empirical evidence that certain expectations exist for social groups to conform to linguistic patterns, it is not unreasonable to assume that these expectations extend to media organizations, models, and journalists. An interesting extension to LET is the fact that various media are not only expected to conform to certain linguistic and stylistic standards, they also define these standards for themselves. While not directly related to citizen journalism studies, research on newspaper language has shown that broadsheet newspapers use a completely different style from that of tabloid newspapers, even when covering the same topic (Fowler 2013). What is even more important here is that the language that one medium considers to be appropriate is frowned upon by the other medium (Bagnall 1993). It is often the case that media have very distinct audiences—professional broadsheet newspapers, for example, feeling threatened by the proliferation of digital information, are beginning to cater to a shrinking circle of "elite" readership (Meyer 2008). Citizen journalism, on the other hand, strives to spread information as freely as possible (Bruns, Highfield, and Lind 2012). Some researchers argue that different audience characteristics of the media force these media to employ different structures of language (e.g., Fowler 1991). For example, quality newspapers could be expected to use more complex language than blogs, etc. And if they do, how do expectancy violations by these media affect their credibility? These are potentially important avenues for journalism research in times of declining trust in traditional news outlets and increasing success of alternative forms of news distribution. To begin systematically addressing these issues, one would need to have empirical data showing the difference (or lack thereof) in linguistic patterns in traditional and citizen journalism. This study is intended to be a first step in bridging this empirical and theoretical gap.

Such a comparative study of the structural difference of language between different news media formats seems even more important if we consider the fact that most people select the medium they consume, and the fact that social stratification and education are a very strong predictor of media choice (Chan and Goldthorpe 2007). In the same study, the authors argue that the choice of the medium may be highly associated with the reader's information-processing capacity. This argument ties back to the knowledge gap hypothesis and consequently the question of linguistic complexity is even more relevant.

The definition of complexity and its operationalization differs across studies, and most of the studies focus on a single aspect of linguistic complexity. Three different dimensions, however, can be outlined—*semantic complexity*, *syntactic complexity*, and *information entropy*. Semantic complexity is the hardest to both define and measure. It is most often operationalized by proxy of *lexical richness* and *lexical diversity*, that is, in the simplest terms, it measures how many unique words are used in a selected text (e.g., Malvern and Richards 2002). Syntactic complexity, on the other hand, is related to the formal structure of the text: length of its words, sentences, clauses, etc. One of the first scales developed to measure text complexity was the Flesch Reading Ease Test (Flesch 1948), which effectively measures syntactic complexity as well. Finally, complexity can be measured as information entropy (Shannon and Weaver 1964). This measure was used in Kleinnijenhuis' study to determine news complexity, however it can also be applied to estimate semantic characteristics of the text (Dale, Moisl, and Somers 2000, 551). The present study aims to use *all* of the aforementioned dimensions of text complexity and apply them in a comparative design to investigate whether professional newspapers differ between themselves and from citizen journalism media in terms of complexity.

Even though there is only a limited body of research on the structural differences in different media, drawing from previous studies on citizen journalism and traditional media a series of hypotheses can be formulated. First, citizen journalism, quality newspapers, and tabloid newspapers are expected to differ regarding the structure of the language they use. There are clear structural differences between the three media to expect such variation. Traditional newspaper articles go through a long process of editorial selection that most of the citizen journalism outlets either lack or these routines are not as rigorous (e.g., Goode 2009). The labor-intensive work of correcting bad grammar and adjusting the language to the newspaper's standard is viewed as a service to the reader and a pride of a newspaper (Thurman 2008). Editing changes not only the content, but the structure of language used in a final article, the process that sometimes is referred to as the "creation of language" (Bell 1991, 80–85). The first hypothesis, therefore, is:

> **H1:** Quality newspapers, tabloid newspapers, and citizen journalism articles will be significantly different on scores of text complexity.

For reasons discussed above, and because technical, dense texts are expected to be highly syntactically complex (Miller and Miller 2011, 65), it expected that quality newspaper articles will be more syntactically complex than citizen journalism articles and tabloid newspaper articles.

> **H2:** Quality newspaper articles will be more syntactically complex than both tabloid newspaper articles and citizen journalism articles.

Expectations regarding semantic differences between the three formats are more tentative. While quality newspapers undergo a strict editorial process during their

production, which quite possibly standardizes the language used in the articles, citizen journalism and tabloid media articles may use a greater variety of styles (Borjars and Burridge 2013), jargon, and emotional language, hence allowing for a broader, less sterile use of language (Timuçin 2010). This would result in higher semantic variety. We thus cautiously expect that:

> **H3:** Citizen journalism and tabloid newspaper articles will be more semantically complex than traditional newspaper articles.

Method

Addressing RQ1, a series of analyses were performed on a large sample of English-language newspapers and English-language political blogs. Additionally, a sample of German-language quality and tabloid newspapers were included to determine whether the findings regarding newspaper differences would hold across linguistic and journalistic environments.[1] The newspaper samples were obtained from digital newspaper repositories. To gather political blog articles' data, a series of Python scripts was written to scrape articles from their respective websites. For subsequent data preparation and data analyzes, scripts in Python and R programming languages were written. Natural language processing was performed with the spaCy Python module.

Sample

Two quality newspapers (*New York Times* and *Washington Post*), three tabloid newspapers (*New York Post*, *USA Today*, and *Los Angeles Daily News*), and five political blogs (*FiveThirtyEight*, *The Daily Beast*, *Breitbart*, *Politico*, and *The Wall Street Journal Blog*) were used in the English-language sample. The German sample comprised of two quality newspapers (*Frankfurter Allgemeine Zeitung* and *Süddeutsche Zeitung*) and a tabloid newspaper (*Bild*). The newspapers were chosen by largest average daily circulation and blogs were chosen by popularity. Both for the newspaper sample as well as for the blog sample only texts pertaining to politics were analyzed, while every other topic was filtered out. A total of 927,593 texts were analyzed. The sample represented five years of political coverage, from 2011 to 2015.

Variables

Syntactic complexity. A few simple measures, such as the average word length and the average sentence length, were used as rudimentary metrics of syntactic complexity. A more advanced metric was the *syntactic depth measure* (e.g., Yngve 1960). This measure determines the length of a parsed syntactic tree for the base word to the terminal word. The *syntactic dependency* metric takes a different approach to generating syntactic structures. Only words are used as nodes in a dependency tree (unlike verb and noun phrases in a syntactic tree) (e.g., Nivre 2005). Syntactic measures also included the *automated readability index* (ARI)—a readability measure that is independent of language-specific linguistic features (such as the number of syllables), and therefore is better suited for performing analyses on multiple languages.

Semantic complexity. *Type/token ratio* is the simplest way to determine the lexical characteristics of the text. The metric represents the ratio of the unique words in the

text to all words (text length). While the simple type/token ratio is the most common approach to measuring the lexical diversity of a text, some argue that its applicability is limited only to short texts because in a longer text the words inevitably start to repeat (e.g, Fergadiotis, Wright, and West 2013). *MTLD* (Measure of Textual Lexical Diversity) is a measure that is maximally independent of the text size (Covington and McFall 2010). The algorithm calculates the type/toke ratio sequentially and counts the number of sequences that do not fall under the specified threshold. Another measure of textual semantic diversity was proposed by a mathematical statistician Yule (2014). Yule's *I* metric is an attempt to fit a hypergeometric distribution to word frequencies, and it shows a probability of two randomly sampled elements from a set being the same. Finally, *semantic entropy* is a logarithmic measure based on the information theory entropy (Dale, Moisl, and Somers 2000, 551). Entropy is a measure of uncertainty. For example in a string "AAAAAAA", entropy is 0, since there exists no uncertainty as to what the next symbol would be. However, if the string was composed of a random sequence of "A"s and "B"s, entropy would be 0.5.

Content complexity. This measure was used to determine the newspaper complexity as well as frame complexity in Kleinnijenhuis (1991) and Kleinnijenhuis, Schultz, and Oegema (2015). As semantic entropy, this metric also measures uncertainty. The uncertainty here is associated with the information in the text. Named entity recognition algorithms were used to extract entities from a text (people, geographic locations, and organizations). High content complexity score indicates a uniform distribution of named entities throughout the text, implying a higher complexity of the content presented in the text.

To facilitate easier understanding of the results and further analyses, all scores for the different measures were normalized (mean = 0; SD = 1).

Results

Descriptive Statistics

As evident from Figure 1, variables' mean scores differ depending on the medium (blogs versus tabloids versus quality newspapers). Variables measuring *syntactic text complexity*—mean length of sentence, mean sentence depth, syntactic dependency depth, and ARI—appear to be much higher for quality newspapers than for citizen journalism blogs and tabloid newspapers. However, the reverse is true for most scores corresponding to the *semantic complexity*—lexical diversity, semantic entropy, and Yule's *I* are all higher for the tabloid newspapers, followed by the citizen journalism blogs and quality newspapers. Finally, the variation in the *content complexity* between the different types of media is not as striking as in the previous dimensions. These preliminary observations suggest that, while having a more complex structure, political articles found in quality newspapers are less lexically and semantically diverse.

Cluster Analysis and Factor Analysis

Data reduction techniques are applied to come to a more general conclusion regarding medium differences. For theoretical reasons, the content complexity variable was not included in the dimension reduction model, but was treated as a separate dimension.

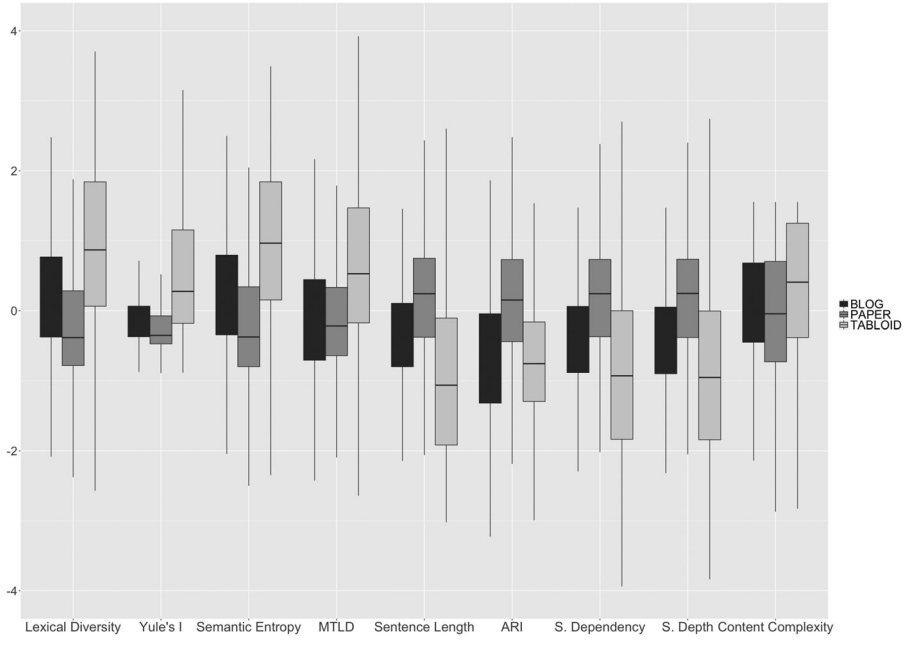

FIGURE 1
Descriptive boxplots for all variables entered in the data reduction model. All variable scores are normalized (mean = 0; SD = 1). Black boxplots: citizen journalism blogs; dark grey: quality newspapers; light grey: tabloid newspapers

Content complexity is not a measure of linguistic characteristics in the sense that the other presented variables are. It is bound to mathematically co-vary with the other variables, since it is based on word counts, however, it measures how complex the information encased in the text is, rather than how complex the language in the text is. First, a hierarchical cluster analysis was performed to be able to visually inspect how different variables load on separate complexity dimensions. Conforming with expectations, the analysis yielded two large clusters—corresponding to semantic and syntactic dimensions of textual complexity. The visualization of the hierarchical cluster analysis is presented in Figure 2.

Factor analysis was then performed to further investigate whether the measured variables correspond to the expected two-dimensional model. A two-factor solution was the optimal model fit—only two factors had eigenvalues higher than 1, 4.90 and 1.92, the eigenvalues dropped steeply after—with third and fourth factors scoring only 0.45 and 0.32, respectively. The data satisfied the statistical assumptions of sampling adequacy ($KMO = 0.75$) and homogeneity of variances—Bartlett test of sphericity ($p < 0.001$).[2]

To provide an additional face validation to the method, a model with German-language data was also estimated. After separating the sample by language (English and German)—the results stay very similar—which indicates that the dimensionality of textual complexity is not an isolated incident, but rather a more general model of text. Descriptives for the full factor analysis model, as well as German and English models are provided in Table 1.

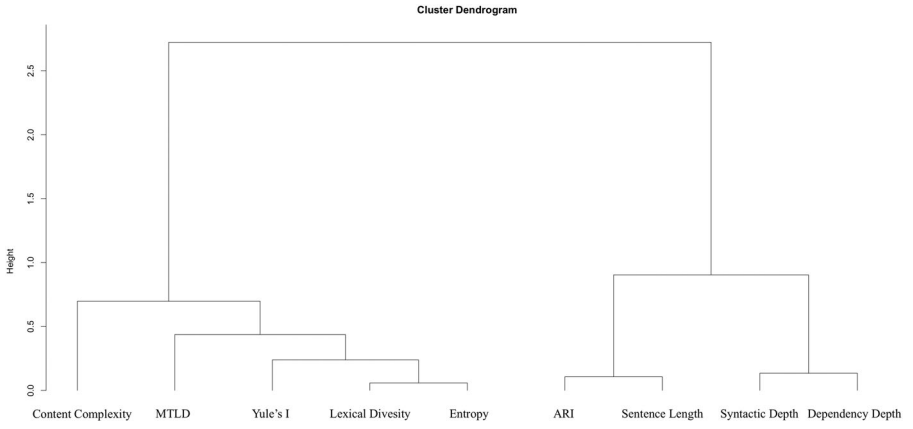

FIGURE 2
Cluster dendrogram for descriptive cluster analysis

Cross-media Comparison

After conducting the factor analysis, the extracted factors, as well as the content complexity variable, were then used to analyze the differences in complexity dimensions across the three different media (citizen journalism blogs, professional newspapers, tabloid newspapers). A boxplot for normalized factor scores (mean = 0; SD = 1) is provided in Figure 3.

As can be seen from the figure, professional newspapers score higher on both syntactic and content complexity. Citizen journalism blogs, however, are semantically more

TABLE 1
Model summary for factor analysis models

	Model I		Model II		Model III	
	F1	**F2**	**F1**	**F2**	**F1**	**F2**
Lexical diversity	0.93		0.89		0.94	
Yule's *I*	0.84		0.85		0.82	
Semantic entropy	94		0.91		0.95	
MTLD	0.86		0.88		0.78	
Sentence length		0.89		0.92		0.94
ARI		0.94		0.85		0.94
Syntactic depth		0.90		0.93		0.96
Syntactic dependency		0.90		0.95		0.96
Intra-factor correlations	−0.38***		−0.31***		−0.15***	
Eigenvalues	4.90	1.92	4.37	2.24	3.91	2.75
Variance	0.43	0.42	0.43	0.40	0.45	0.38
Cumulative variance	0.43	0.85	0.43	0.83	0.45	0.83
χ^2	16,4051.4***		34,934.76***		166,534.2***	

Model I: German and English texts combined; Model II: only German texts; Model III: only English texts.
***$p < 0.001$.

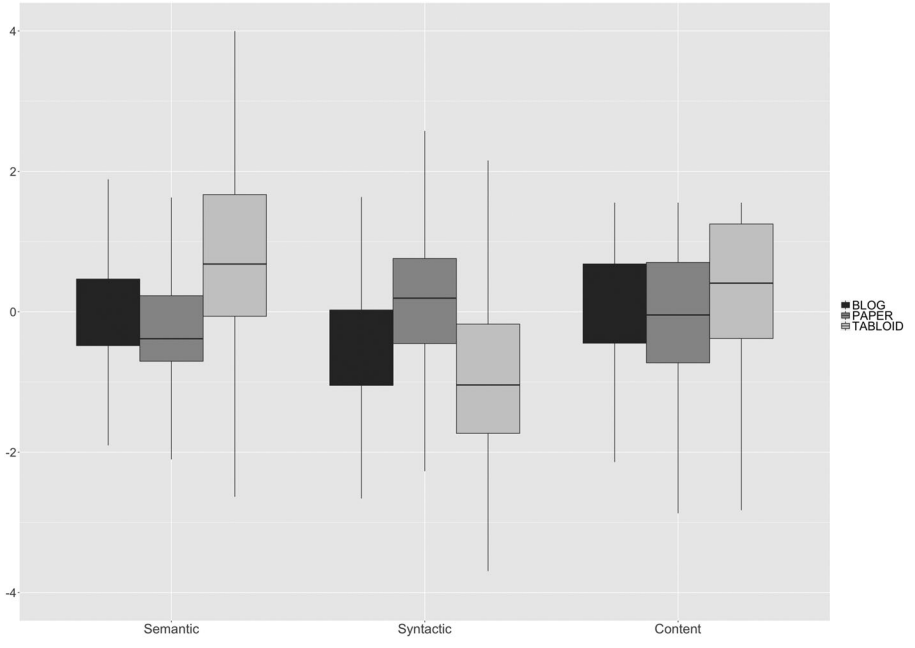

FIGURE 3
Descriptive boxplots for complexity dimensions. All variable scores are normalized (mean = 0; SD = 1). Black boxplots: citizen journalism blogs; dark grey: quality newspapers; light grey: tabloid newspapers

complex than their professional counterpart. Professional newspapers are more complex regarding the content of the articles. For clarification purposes a table with sample sentences from the analyzed outlets is provided in Table 2.

One-way ANOVA models were estimated to compare the complexity dimension means across media. All ANOVA models for complexity dimensions were significant at $p < 0.001$: $F(2, 927{,}590) = 49{,}413$ for the semantic model, $F(2, 927{,}590) = 61{,}504$ for the syntactic model, and $F(2, 927{,}590) = 5538$ for the content complexity model.

English-language quality newspapers score the highest on the syntactic complexity dimension (mean = 0.13; SD = 0.95) followed by tabloid papers (mean = −0.49; SD = 0.93) and citizen journalism (mean = −0.85; SD = 0.94), $F(2, 770{,}735) = 40{,}395$, $p < 0.001$. This is the only exception from the combined model, since here citizen journalism articles scored lower than tabloid articles. On the English-language semantic dimension, tabloids scored the highest (mean = 0.54; SD = 0.93), closely followed by citizen journalism (mean = 0.51; SD = 1.18); quality newspapers were the least semantically complex (mean = −0.10; SD = 0.95), $F(2, 156{,}853) = 20{,}392$, $p < 0.001$. Finally, regarding the content complexity of the text, English-language tabloid newspapers (mean = 0.21; SD = 0.99) are the most complex of all text types, followed by citizen journalism articles (mean = 0.18; SD = 0.89), followed by quality newspapers (mean = −0.03; SD = 0.93), $F(2, 770{,}735) = 2618.9$, $p < 0.001$.

It is important to note that these patterns hold for separate language models as well. German-language quality newspaper articles are significantly more syntactically complex (mean = 0.30; SD = 0.99) than German tabloid articles (mean = −0.77; SD = 1.01), $F(2,$

TABLE 2
Sample paragraphs from the analyzed media types

		Complexity		
Text	**Type**	**Semantic**	**Syntactic**	**Content**
Continuing public support for resettlement programs in the United States and other industrialized countries depends on research and public awareness of the long-term benefits of resettlement policy, the integration of refugees into their new communities, and the social, cultural and economic effect of resettlement on the United States. As we seek innovative solutions to refugee and migration crises and reform immigration policies, let's not focus only on deterrence strategies. Resettlement programs, if properly researched and established, offer the greatest potential for countries to absorb refugees, not just by sharing the costs of refugee crises, but by harnessing the great resources that refugees have to offer.	Quality newspaper	−0.57	1.23	0
President Barack Obama has seen the enemy, and it is the refusal to accept more Syrian refugees. From the tone of his post-Paris remarks, you'd think that a sophisticated terrorist assault on a major Western city is a setback; sentiment in the U.S. against taking more Syrian refugees is an atrocity. Obama warned this week against "that dark impulse inside of us," as if we were debating whether Syrian refugees should be drawn and quartered. He said that "slamming the door in their faces would be a betrayal of our values."	Online blog	0.15	−0.29	0.69
Thousands of refugees left their camp near the front lines as ISIS forces advanced. Khazer Camp had been filled with Iraqis who had already fled their besieged homes. US cargo planes dropped shipments of food and water for the second night in a row in areas near mountain camps to help the famished refugees, who are facing dehydration and food shortages. President Obama phoned Jordan's King Abdullah in a show of support for a key ally threatened by the stunning advances of ISIS.	Tabloid newspaper	0.45	−0.17	1

The semantic and syntactic complexity scores are scaled (mean = 0; SD = 1). For content complexity a score of 0 indicates no entropy (e.g., only one named entity), while a score of 1 indicates maximum entropy (e.g., multiple named entities, without repletion). The sample paragraphs may not be completely representative, since the automated content analysis method was designed to work with full articles. For example, content complexity of 1 or 0 is very improbable in the full articles.

156,853) = 47,399, $p < 0.001$. German tabloid articles are significantly more semantically complex (mean = 0.48; SD = 1.06) than quality media (mean = −0.19; SD = 0.91), $F(2, 156,853) = 15,406$, $p < 0.001$. Content-wise, German tabloid articles (mean = 0.09; SD = 1.01) are more complex than quality newspapers (mean = −0.04; SD = 0.99), $F(2, 156,853) = 588.01$, $p < 0.001$.

Thus, all three hypotheses in this study were confirmed. The three newspaper text types are significantly different from each other (H1); professional newspapers are more syntactically complex than both the tabloid newspapers and citizen journalism articles (H2); and citizen journalism texts and tabloid newspapers score significantly higher on the semantic complexity metric than professional newspapers (H3). Additionally, it was established that similar complexity patterns are occurring across English and German languages.

Conclusion and Discussion

In the present study, the structural characteristics of political journalistic texts from three different media types were compared, specifically professional newspaper articles, tabloid newspaper articles and political weblog articles. The aim of this research was to investigate whether textual complexity differs between these media, and in particular whether citizen journalism texts would diverge from the other two professional types of texts. First, a range of textual complexity measures was chosen from fields ranging from linguistics to social sciences and applied to ∼930,000 articles. It was determined that these measures actually gauge three different dimensions of textual complexity: syntactic, semantic, and content complexity. The content complexity measure was not entered with the other variables into a data reduction model because it is, unlike the rest, not a measure that determines the linguistic characteristics of a text.

A series of comparisons determined whether professional quality and tabloid newspapers differ in complexity dimensions from political blogs. Quality newspapers had higher syntactic complexity than the other two text types, but citizen journalism and tabloid newspaper articles scored higher on the semantic and content dimension. One possible explanation for these findings is the fact that professional newspaper articles go through a rigorous process of filtering (journalistic routines, gatekeeping, standardization of language, etc.) before actually being printed. For the blogs, this might not be the case. An individual writer of a blog article may not need to think about a "standard language template", or an emotional tone of the article, therefore utilizing various synonyms, colloquialisms, and jargon that is not available to the professional journalist, making the text lexically richer. Additionally, it has been noted that tabloid newspapers often employ a different style of language than that of quality media, even when covering the same topic (Fowler 2013), and also simplify their text to accommodate the audience (Zelizer et al. 2000). These characteristics of tabloid journalism may explain the high semantic complexity (e.g., very emotional language, non-standard colloquialisms, etc.) and low syntactic complexity (a relatively simple style). These patterns hold separately in German and English languages, indicating that the media landscape is quite similar in terms of language complexity across different languages.

An important conclusion from these results is that quality newspapers have the most complex syntactic structure, while citizen journalism blogs and tabloid newspapers include semantically rich text. This indicates that a syntactically complex text is not a necessary

prerequisite for a complex political story (high semantic and content complexity). The relevance of these findings is significant, since many branches of journalism research could benefit from the addition of complexity as a factor. For example, researchers may be interested in how complexity affects credibility, whether these effects are different across different media, or whether different dimensions of complexity have a different effect on credibility, among others.

The present study also included a number of limitations. For example, the measures chosen to determine the linguistic complexity, although covering a wide range of linguistic nuances, are in no way exhaustive. Experimenting with other complexity metrics and investigating how they are related to political texts would add precision to the research on complexity. The second major limitation is the fact that the present study did not discriminate between various topics within the broad category of political texts. It could be the case that different political topics vary widely in terms of text complexity even within one particular publication. Future researchers should also strive to include more languages into the comparative design. This would help to understand how journalism is affected by text complexity across both the linguistic and cultural environments. Addressing these issues, further research would be able to come even closer to understanding the intricate interplay between linguistic complexity and journalism.

Admittedly, the present study was interested solely in the structural linguistic differences between various types of journalistic outlets. The nature of the automated content analysis method is such that, while it allows the analysis of amounts of data that would be unimaginable for humans to deal with, thereby uncovering meaningful patterns that would otherwise be hard to see, it pushes the human dimension of communication to the background. This study suffers from this limitation as well. However, even though automated methods do not fully capture all aspects of analyzed text, they might provide an insight for future research, giving us the ability to investigate the interplay between the content and the structure.

In summary, notwithstanding the limitations associated with this study, it provides valuable insights into the role of text complexity in journalism and how quality newspapers, tabloid newspapers, and citizen journalism articles differ in terms of text complexity and across its dimensions. It is now evident that in order to have a holistic understanding of textual complexity, future research must incorporate a range of measures to correctly determine the structural characteristics of text.

Moreover, the structural linguistic complexity aspect of the journalistic media also plays a very important role in bridging various branches of communication research. Credibility research, knowledge gap research, and research on LET are all related via the linguistic characteristics of media. Previous research indicates that the structure of the texts matters in these cases, now that we have empirical evidence that different media vary in the complexity of their texts, one could start asking questions about whether people perceive complex texts to be an indication of *expertise* or *comprehensiveness* (credibility research). For example, are higher levels of complexity perceived to be a positive characteristic since they imply expertise, while a different medium actually benefits from lower complexity levels, since people expect this medium to be comprehensible (LET); and how does that pan out for different types of complexity? Or maybe the readership expects different media to be more or less complex depending on the educational background (LET and the knowledge gap hypothesis). Application of automated text analysis in journalism research allows the incorporation of computationally intensive

problems, such as the analysis of linguistic complexity, and opens new directions for further investigations (see also Boumans and Trilling 2016). A longitudinal or a cross-national study investigating such phenomena as tabloidization (e.g., Esser 1999) on a very large scale would now be more feasible. These questions are now open for future research and, if pursued, would undoubtably enrich the theoretical body of communication and journalism research. The present study is but a first step towards finding one of the missing puzzle pieces—political news textual complexity—in the wide picture of political communication research.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

NOTES

1. In the German context, political blogs play so far only a marginal role and therefore the analysis is now restricted to newspapers only.
2. Eight variables were used in the factor analysis model: mean length of sentence, mean sentence depth, syntactic dependency depth, ARI, lexical diversity, semantic entropy, Yule's *I*, and MTLD. A null factor analysis model with eight factors and no rotation extracted two factors with Eigenvalues above 1 (4.90 and 1.92, respectively). A two-factor model with oblimin rotation was estimated. Oblimin rotation is appropriate in this case because cross-factor correlations may be discovered. The model explained approximately 83 percent of the variance in the data. The extracted components correspond to the expected two-dimensional complexity model—clearly showing semantic and syntactic constructs.

REFERENCES

Averbeck, Joshua M. 2010. "Irony and Language Expectancy Theory: Evaluations of Expectancy Violation Outcomes." *Communication Studies* 61 (3): 356–372.

Bagnall, Nicholas. 1993. *Newspaper Language*. London: Routledge.

Beam, Randal A. 2003. "Content differences between Daily Newspapers with Strong and Weak Market Orientations." *Journalism & Mass Communication Quarterly* 80 (2): 368–390.

Bell, Allan. 1991. *The Language of News Media*. Oxford: Blackwell.

Blobaum, Bernd. 2016. *Trust and Communication in a Digitized World: Models and Concepts of Trust Research*. New York: Springer.

Borjars, Kersti, and Kate Burridge. 2013. *Introducing English Grammar*. New York: Routledge.

Boumans, Jelle W., and Damian Trilling. 2016. "Taking Stock of the Toolkit: An Overview of Relevant Automated Content Analysis Approaches and Techniques for Digital Journalism Scholars." *Digital Journalism* 4 (1): 8–23.

Bowman, Shayne, and Chris Willis. 2003. "*We media. How audiences are shaping the future of news and information.*"

Bruns, Axel, Tim Highfield, and Rebecca Ann Lind. 2012. "Blogs, Twitter, and Breaking News: The Produsage of Citizen Journalism." *Produsing Theory in a Digital World: The Intersection of Audiences and Production in Contemporary Theory* 80 (2012): 15–32.

Buller, David B., Michael Burgoon, John R. Hall, Norman Levine, Ann M. Taylor, Barbara Beach, Mary Klein Buller, and Charlene Melcher. 2000. "Long-term effects of Language Intensity in Preventive Messages on Planned Family Solar Protection." *Health Communication* 12 (3): 261–275.

Burgoon, Michael, and Gerald R. Miller. 1985. "An Expectancy Interpretation of Language and Persuasion." In *Recent Advances in Language, Communication, and Social Psychology*, edited by H. Giles and R. N. Clair, 199–229. London: Lawrence Erlbaum.

Burgoon, Michael, Vickie Pauls Denning, and Laura Roberts. 2002. "Language Expectancy Theory." In *The Persuasion Handbook: Developments in Theory and Practice*, edited by James Price Dillard and Michael Pfau, 117–136. Thousand Oaks, CA: Sage Publications.

Chan, Tak Wing, and John H. Goldthorpe. 2007. "Social Status and Newspaper Readership 1." *American Journal of Sociology* 112 (4): 1095–1134.

Covington, Michael A., and Joe D. McFall. 2010. "Cutting the Gordian Knot: The Moving-average type–token ratio (MATTR)." *Journal of Quantitative Linguistics* 17 (2): 94–100.

Dale, Robert, Hermann Moisl, and Harold Somers, eds. 2000. *Handbook of Natural Language Processing*. New York: Marcel Dekker.

De Keyser, Jeroen, Karin Raeymaeckers, and Steve Paulussen. 2011. "Are Citizens Becoming Sources?" In *Journalists, Sources and Credibility: New Perspectives*, edited by B. Franklin and M. Carlson, 139–151. New York: Routledge.

Deuze, Mark. 2005. "Popular Journalism and Professional Ideology: Tabloid Reporters and Editors Speak out." *Media, Culture & Society* 27 (6): 861–882.

Dillard, James Price, and Michael Pfau. 2002. *The Persuasion Handbook: Developments in Theory and Practice*. Thousand Oaks: Sage Publications.

Esser, Frank. 1999. "Tabloidization' of News. A Comparative Analysis of Anglo-American and German Press Journalism." *European Journal of Communication* 14 (3): 291–324.

Eveland, William P., and Juliann Cortese. 2004. "How Web site Organization Influences Free Recall, Factual Knowledge, and Knowledge Structure Density." *Human Communication Research* 30 (2): 208–233.

Eveland, William P., and Sharon Dunwoody. 2001. "User Control and Structural Isomorphism or Disorientation and Cognitive Load? Learning from the Web Versus Print." *Communication Research* 28 (1): 48–78.

Fergadiotis, Gerasimos, Heather H. Wright, and Thomas M. West. 2013. "Measuring Lexical Diversity in Narrative Discourse of People with Aphasia." *American Journal of Speech-Language Pathology* 22 (2): S397–S408.

Flesch, Rudolph. 1948. "A New Readability Yardstick." *Journal of Applied Psychology* 32 (3): 221.

Fowler, Roger. 1991. *Language in the News: Discourse and Ideology in the Press*. London: Routledge.

Fowler, Roger. 2013. *Language in the News: Discourse and Ideology in the Press*. London: Routledge.

Goode, Luke. 2009. "Social News, Citizen Journalism and Democracy." *New Media & Society* 11 (8): 1287–1305.

Hanitzsch, Thomas. 2007. "Deconstructing Journalism Culture: Toward a Universal Theory." *Communication Theory* 17 (4): 367–385.

Hovland, Carl I., and Walter Weiss. 1951. "The Influence of Source Credibility on Communication Effectiveness." *Public Opinion Quarterly* 15 (4): 635–650.

Jensen, Matthew L., Joshua M. Averbeck, Zhu Zhang, and Kevin B. Wright. 2013. "Credibility of Anonymous Online Product Reviews: A Language Expectancy Perspective." *Journal of Management Information Systems* 30 (1): 293–324.

Jucks, Regina, and Elisabeth Paus. 2012. "What makes a Word Difficult? Insights into the Mental Representation of Technical Terms." *Metacognition and Learning* 7 (2): 91–111.

Kiousis, Spiro. 2001. "Public Trust or Mistrust? Perceptions of Media Credibility in the Information Age." *Mass Communication & Society* 4 (4): 381–403.

Kleinnijenhuis, Jan. 1991. "Newspaper Complexity and the Knowledge Gap." *European Journal of Communication* 6 (4): 499–522.

Kleinnijenhuis, Jan, Friederike Schultz, and Dirk Oegema. 2015. "Frame Complexity and the Financial Crisis: A Comparison of the United States, the United Kingdom, and Germany in the Period 2007–2012." *Journal of Communication* 65 (1): 1–23.

Lasica, Joseph D. 2003. "What is Participatory Journalism." *Online Journalism Review* 7 (8). http://www.ojr.org/ojr/workplace/1060217106.php.

Lewis, Seth C., Kelly Kaufhold, and Dominic L. Lasorsa. 2010. "Thinking about Citizen Journalism: The Philosophical and Practical Challenges of User-generated Content for Community Newspapers." *Journalism Practice* 4 (2): 163–179.

Malvern, David, and Brian Richards. 2002. "Investigating Accommodation in Language Proficiency Interviews Using a New Measure of Lexical Diversity." *Language Testing* 19 (1): 85–104.

Meyer, Philip. 2008. "The Elite Newspaper of the Future." *American Journalism Review* 30 (5): 32–35.

Miller, James Edward, and Jim Miller. 2011. *A Critical Introduction to Syntax*. New York: Continuum.

Moy, Patricia, and Dietram A. Scheufele. 2000. "Media Effects on Political and Social Trust." *Journalism & Mass Communication Quarterly* 77 (4): 744–759.

Nah, Seungahn, and Deborah S. Chung. 2012. "When Citizens meet both Professional and Citizen Journalists: Social Trust, Media Credibility, and Perceived Journalistic roles among Online Community News Readers." *Journalism* 13 (6): 714–730.

Nivre, Joakim. 2005. "Dependency Grammar and Dependency Parsing." *MSI Report* 5133 (1959): 1–32.

O'Sullivan, John, and Ari Heinonen. 2008. "Old Values, New Media: Journalism Role Perceptions in a Changing World." *Journalism Practice* 2 (3): 357–371.

Paus, Elisabeth, and Regina Jucks. 2011. "Depressive or just in a bad Mood? Laypersons' Assumptions about their Knowledge of Medical Vocabulary." *Studies in Communication Sciences* 11 (1): 51–71.

Peters, Chris, and Marcel J. Broersma. 2013. *Rethinking Journalism: Trust and Participation in a Transformed News Landscape*. New York: Routledge.

Pfau, Michael, Roxanne Parrott, and Bridget Lindquist. 1992. "An Expectancy Theory Explanation of the Effectiveness of Political Attack Television Spots: A Case Study." *Journal of Applied Communication Research* 20 (3): 235–253.

Pornpitakpan, Chanthika. 2004. "The Persuasiveness of Source Credibility: A Critical Review of Five Decades' Evidence." *Journal of Applied Social Psychology* 34 (2): 243–281.

Reese, Stephen D., Lou Rutigliano, Kideuk Hyun, and Jaekwan Jeong. 2007. "Mapping the Blogosphere Professional and Citizen-based Media in the Global News Arena." *Journalism* 8 (3): 235–261.

Scharrer, Lisa, Rainer Bromme, M. Anne Britt, and Marc Stadtler. 2012. "The Seduction of Easiness: How Science Depictions Influence Laypeople's Reliance on Their Ownevaluation of Scientific Information." *Learning and Instruction* 22 (3): 231–243.

Shannon, C., and W. Weaver. 1964. *The Mathematical Theory of Information*. Urbana: Illinois Press.

Singer, Jane B. 2004. "More than Ink-stained Wretches: The Resocialization of Print Journalists in Converged Newsrooms." *Journalism & Mass Communication Quarterly* 81 (4): 838–856.

Thiebach, Monja, Elisabeth Mayweg-Paus, and Regina Jucks. 2015. ""Probably True" says the Expert: How two Types of Lexical Hedges Influence Students' Evaluation of Scientificness." *European Journal of Psychology of Education* 30 (3): 369–384.

Thomm, Eva, and Rainer Bromme. 2012. "'It should at Least Seem Scientific!' Textual Features of 'Scientificness' and Their Impact on Lay Assessments of Online Information." *Science Education* 96 (2): 187–211.

Thurman, Neil. 2008. "Forums for Citizen Journalists? Adoption of User Generated Content Initiatives by Online News Media." *New Media & Society* 10 (1): 139–157.

Tichenor, Phillip J., George A. Donohue, and Clarice N. Olien. 1970. "Mass Media Flow and Differential Growth in Knowledge." *Public Opinion Quarterly* 34 (2): 159–170.

Timuçin, Metin. 2010. "Different Language Styles in Newspapers: An Investigative Framework." *Journal of Language and Linguistic Studies* 6 (2): 104–126.

Williams, Bruce A., and Michael X. Delli Carpini. 2004. "Monica and Bill all the Time and Everywhere the Collapse of Gatekeeping and Agenda Setting in the New Media Environment." *American Behavioral Scientist* 47 (9): 1208–1230.

Wodak, Ruth, ed. 1989. *Language, Power and Ideology: Studies in Political Discourse*. Vol. 7. Amsterdam: John Benjamins Publishing.

Yngve, Victor H. 1960. "A Model and an Hypothesis for Language Structure." *Proceedings of the American Philosophical Society* 104 (5): 444–466.

Yule, C. Udny. 2014. *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.

Zelizer, Barbie, S. Elizabeth Bird, Rod Brookes, Andrew Calabrese, Peter Golding, Jostein Gripsrud, Ágnes Gulyás, et al. 2000. *Tabloid Tales: Global Debates Over Media Standards*. Lanham: Rowman & Littlefield Publishers.

Petro Tolochko (author to whom correspondence should be addressed), Department of Communication, University of Vienna, Austria. E-mail: petro.tolochko@univie.ac.at

Hajo G. Boomgaarden, Department of Communication, University of Vienna, Austria. E-mail: hajo.boomgaarden@univie.ac.at