



Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Analysis of Textual Complexity in Danish News Articles on Climate Change

Meier, Florian Maximilian; Eskjær, Mikkel

Published in:

Proceedings of DHNB 2024 Digital Humanities in the Nordic and Baltic Countries 8th Conference

Creative Commons License
Unspecified

Publication date:
2024

Document Version
Accepted author manuscript, peer reviewed version

[Link to publication from Aalborg University](#)

Citation for published version (APA):

Meier, F. M., & Eskjær, M. (2024). Analysis of Textual Complexity in Danish News Articles on Climate Change. In *Proceedings of DHNB 2024 Digital Humanities in the Nordic and Baltic Countries 8th Conference* Advance online publication.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy

If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Analysis of Textual Complexity in Danish News Articles on Climate Change

Meier Florian¹, Fugl Eskjær Mikkel¹

¹*Department of Communication and Psychology, Aalborg University, Copenhagen*

Abstract

Structural linguistic features are often overlooked yet potentially important aspects of journalistic practice. Especially in news reporting on climate change, these features can play a crucial role as the proper use of language is tied to message credibility, processing fluency and knowledge retention, which can positively influence the reader to take more climate action. This article analyzes language use in Danish news articles on climate change using a sample of around 32,000 articles from four different outlet types (quality news, niche papers, tabloids, and public service broadcasters) published from 1990 to 2021. We create a machine-learning model of text complexity covering this concept's semantic and syntactic dimensions. Our findings confirm expected differences in complexity between news outlets, highlighting tabloid articles as engaging with higher semantic complexity, while quality papers and niche papers exhibit higher syntactic complexity. We observe a significant decrease in semantic complexity and a slight increase in syntactic complexity over time, a trend towards more generic language, and an increased use of pronouns, verbs, and adverbs. Most of these changes can be attributed to the emergence of articles by public service broadcasters. Articles by public service broadcasters are characterised by high syntactic complexity, which we consider problematic due to their popularity among the general public.

Keywords


Climate change, Newspaper, Text complexity, Machine learning,


1. Introduction and background

The way individuals perceive, understand, and engage with information about climate change (CC), including the potential to enhance knowledge on the subject, is influenced by the complexity of the language with which it is presented. Simpler (vs. more complex) language may facilitate further engagement and learning about human-emitted greenhouse gases and approaches to CC mitigation and adaptation. Language complexity has several effects. Newspaper complexity, for instance, has been shown to explain the knowledge gap hypothesis as groups with differing socioeconomic status will have less or more knowledge due to their literacy (Kleinnijenhuis 1991). Language is essential as it expresses facts, represents realities and influences attitudes and behaviours, thereby shaping new realities (Fløttum 2016). Finally, language influences the credibility of the message communicated in a text (Tolochko and Boomgaarden

 fmeier@ikp.aau.dk (M. Florian)

 <https://vbn.aau.dk/da/persons/142274> (M. Florian)

 0000-0001-9408-0686 (M. Florian)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

2018). Ultimately, language is a critical factor in citizen involvement and when used properly can motivate people to become more climate-conscious and support climate-friendly policies.

CC communication can be considered an instance of science communication facilitating dialogue between the public and scientific research. Linguistic strategies of CC communication among politicians and CC activists have been studied by eco-linguistics (Cunningham, Foxcroft, and Sauntson 2022). In the media, it is equally important to adopt effective strategies for communicating scientific research (Moser 2016). Previous research in mediated CC communication has mostly looked into effective communication strategies using frame analysis (Schäfer and O’Neill 2017; Robertson 2022), or investigating drivers of media attention (Hase et al. 2021). Less is known about the textual complexity and comprehensibility of CC news articles. However, structural linguistic features of news reporting can be a crucial aspect of journalistic practices, which are often overlooked but potentially important. In this paper, we use a sample of around 32,000 newspaper articles from nine of Denmark’s most prominent news outlets covering the period 1990 to 2021 to study the structural linguistic characteristics of CC news reporting.

1.1. Text complexity and its operationalisation

Structural linguistic characteristics are an essential aspect of written communication. Linguistic complexity is a multidimensional construct for which metrics to measure these concepts fall into the categories of *semantic* and *syntactic* complexity (Tolochko and Boomgaarden 2019). Many previous studies, however, fall short of acknowledging this multidimensional nature and operationalise text complexity only via single-item readability measures, like the Flesch reading ease (Tolochko and Boomgaarden 2018). Readability measures are inherently syntactic measures as they usually – among other aspects – measure the number of words per sentence. However, they are only a weak proxy for text complexity as they exclude the semantic dimension. Semantic complexity is usually represented via metrics of lexical richness or diversity, e.g. the type/token ratio, i.e. the number of unique words (types) vs the total number of words (tokens) in a text (Tolochko, Song, and Boomgaarden 2019). Recently, researchers have begun to account for the multi-faceted nature of text complexity and combine multiple features and data reduction techniques, like factor analysis, to build machine learning models of text complexity (Tolochko and Boomgaarden 2018, 2019; Tolochko, Song, and Boomgaarden 2019).

Applying this approach, previous research successfully links *objective* linguistic complexity to the perceived and more subjective complexity of a text. E.g., Tolochko, Song, and Boomgaarden 2019 found that the more syntactically complex a text is, the stronger decreases the ability of people to recall factual information it conveys (Tolochko, Song, and Boomgaarden 2019). These findings align with research on election ballot wording, which shows that the more semantically and syntactically complex the wording is, the lower the voter’s processing fluency (Shulman et al. 2022). However, while final empirical validation in the context of news articles on CC is still missing, variation in semantic versus syntactic complexity would likely produce differing learning processes and information retention.

1.2. Text complexity and newspaper type

One of our primary interests is the structural linguistic differences between various journalistic outlets. A broad distinction can be made between broadsheet papers, also referred to as quality papers, and tabloid news (Richardson 2007; BBC 2023). It is well known that broadsheet news sometimes covers the same topic in a completely different style than tabloid newspapers (Baker 2010). Tabloid journalists use simpler and shorter sentences characterised by emotive, dramatic, or sensationalist language (BBC 2023). Quality press, on the other hand, features a straightforward and factual style with longer and more complex sentences. Tolochko and Boomgaarden 2019 systematically compared articles from quality newspapers, tabloid newspapers, children's books, and legal texts. They conclude that tabloid articles are semantically complex yet characterised by low syntactical complexity. In comparison, quality newspapers are the direct opposite. Children's books score low on both syntactic and semantic complexity as they use language designed to be understandable. Possible explanations for the differences found in news articles are language standardisation in quality news, i.e. following a strict editorial process, which leads to a lower semantic complexity. In contrast, tabloid articles have a greater variety of styles, featuring emotional language, usage of synonyms, colloquialism, and jargon that is rarely found in the highly standardised language of quality news (Tolochko and Boomgaarden 2018). We are following a desideratum put forward by Tolochko and Boomgaarden that asks for a longitudinal study on the diachronic development of semantic and syntactic complexity that could give further clarification on the *tabloidization of news* (Esser 1999; Skovsgaard 2014). Some researchers consider the tabloidization of news a potential threat to democracy (Skovsgaard 2014). However, in the context of CC communication, a tabloid-style news reporting could be potentially beneficial as it would follow UN recommendations on how to report about CC by making CC stories more engaging and relatable to individuals, thereby paving the way for positive climate action (Nations 2022). Table 1 summarises the main characteristics of semantic and syntactic complexity.

Structural differences between quality newspapers, tabloids, and even citizen journalism (blogs) have been studied before (Tolochko and Boomgaarden 2018). However, how online news articles by public service providers differ in this respect has mostly been overlooked. One exception is a study by Jung 2003 comparing online business news from CNBC, CNN, and CBSMarketWatch with print news from the Wall Street Journal, the New York Times, and USA Today. Jung 2003 finds that online news has a more challenging writing style, with higher average sentence length and low reading ease scores. In our context, we deal with online articles published by public service providers. Research has shown that public service media report differently on political issues than market-driven outlets (Cushion 2022). Thus, our main aim is not to compare print vs online articles but rather the difference between CC news produced by traditional media like national newspapers and CC news provided by public service corporations.

1.3. Complexity of climate change texts

While most people pay more attention to TV or movie documentaries, news stories from major news organisations still feature second as the main information source on CC (Robertson 2022).

Table 1

Overview of characteristics of semantic and syntactic complexity known from related work.

	Semantic complexity	Syntactic complexity
Caused by	Emotional tone, colloquialism, jargon	Long sentences with many dependent clauses
Prominent in	Tabloid, blog articles	Broadsheet papers (quality news), legal texts
Metrics	Lexical diversity and lexical richness	Readability measures, Dependency parsing

However, research on the complexity of CC news could be more extensive. Textual material covering the science of CC can be difficult to read, as shown by a study by Barkemeyer et al. 2015 on language use in IPCC's summaries for policymakers (SPM). When comparing IPCC's SPMs with scientific publications and news articles from English language tabloids and quality papers, they identify SPMs as the least readable texts. On the other hand, the readability of CC reporting in newspapers has been steadily increasing. Additionally, the study found these articles have become more emotive over time (Barkemeyer et al. 2015). To sum up, we are addressing the following three research questions:

RQ1 How has semantic and syntactic complexity developed over time, and how does this vary across news outlets?

RQ2 Can structural characteristics of language use explain this development?

RQ3 For which news outlet types do semantic and syntactic complexity differ most significantly?

2. Data and methods

Our analysis is based on a sample of articles collected programmatically by API access to *Infomedia*, one of the largest media archives in Scandinavia. We collected articles on CC and related keywords (greenhouse effect and global warming)¹ between 1990 and 2021, from all printed newspapers currently in circulation as well as two online news sites by Danish public service corporations, Danmarks Radio (dr.dk) and TV2 (tv2.dk), which are the fourth and fifth most visited websites in Denmark (Similarweb 2023). We decided only to use core climate change articles, meaning we focus on all news items that contain one of the keywords at least twice. Table 2 gives an overview of our corpus. One of our main research questions is how textual complexity varies by the type of news outlet. Our sample is dominated by articles from quality newspapers, which are regular board sheet papers intended to be read by the general public. Niche papers appeal more to the intellectual elite. *Information* and *Weekendavisen* belong to this category, also representing both ends of the political spectrum (left-right). Public service providers (DR and TV2) publish online-only news. These stories mostly consist of frequent news updates and breaking news focusing on episodic, less politicized and less controversial

¹The original query in Danish was: klimaforandring* OR "global* opvarmning" OR drivhuseffekt*.

topics. Finally *Ekstra-Bladet* is the only tabloid newspaper in our sample. We study 32,214 articles with an average length of 763.6 words (median=580, min=101, max=13,429).

Table 2

Description of the sample of news articles used in the analysis.

Outlet name	#Print	#Web	Type	Time frame
Berlingske	2,988	3,000	Quality paper	1990-2021
Danmarks Radio (DR)	-	1,523	Public service	2007-2021
Ekstra-Bladet	261	892	Tabloid	1990-2021
Information	2,914	1,466	Niche paper	1997-2021
Jyllands-Posten	2,810	4,061	Quality paper	1996-2021
Kristeligt-Dagblad	1,091	1,786	Niche paper	2001-2021
Politiken	3,724	2,682	Quality paper	1990-2021
TV2	-	1,874	Public service	2008-2021
Weekendavisen	942	200	Niche paper	1990-2021
Total	14,730	17,484	32,214	

2.1. Feature engineering

While syntactic complexity is more clearly defined, a definition of semantic complexity can be more demanding. In this paper, we follow the extensional approach to semantic complexity as introduced by Tolochko and Boomgaarden 2019. Following this definition, the more ambiguity that is associated with a lexical unit (e.g. synonymy, polysemy) the more semantically complex it is. Thus we will mainly use lexical units as proxies for semantic complexity. In our aim to capture as many different characteristics of textual complexity, we calculate multiple measures including the average word length (AWL), the average sentence length (ASL), and various measures of vocabulary richness and lexical diversity, e.g., the word variation index (ordvariationsindex, OVIX), and the mean average type-token ratio (MATTR) (Covington and McFall 2010). Moreover, we calculate LIX (Läsbarhetsindex), a readability index commonly used for Nordic languages (Falkenjack and Jönsson 2014). As previously mentioned, readability measures are commonly seen as measures of syntactic complexity. Finally, we use the Python NLP framework DaCy, an adaption of spaCy for Danish, to perform dependency parsing and extract the height of the dependency tree (Enevoldsen, Hansen, and Nielbo 2021). The tree’s height is the longest dependency chain in the sentence, starting from the root node, the verb.

In the following description of features, S refers to the total number of sentences, N refers to the total number of tokens, V to the number of types, and $f_v(i, N)$ to the numbers of types occurring i times in a sample of length N .

- $AWL = \frac{N}{n_{chars}}$
- $ASL = \frac{N}{S}$
- $LIX = \frac{N}{S} + \left(\frac{N > 6chars}{N} \right) * 100$
- $OVIX = \frac{\log(N)}{\log(2 - \frac{V}{N})}$

- **Yule's I** = $\frac{V^2}{M_2 - V}$ with $M_2 = \sum_{i=1}^V i^2 * f_v(i, N)$
- **MATTR** = Is the mean of multiple TTRs ($\frac{V}{N}$) calculated for a moving window of tokens from the first to the last token of the text ($window\ size = 100$).
- **Syntactic depth**: Height of the dependency tree based on DaCy. We take the average of all sentences for a document.

In a next step, we use these features in a principal component analysis (PCA).

2.2. Principal component analysis

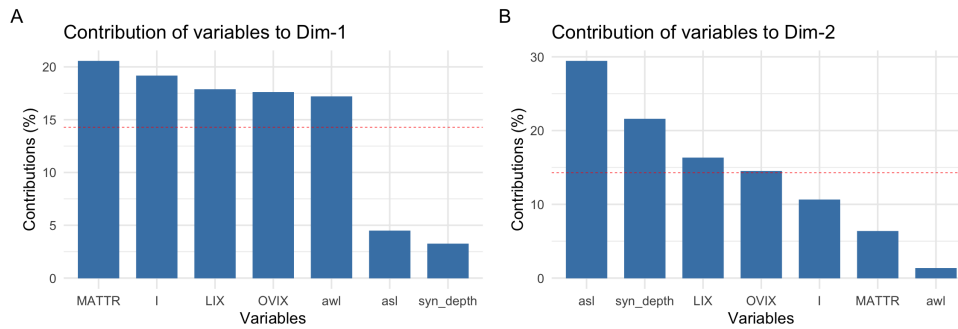


Figure 1: Contribution of features/variables to the two principal components of the PCA. The dashed line in red indicates the expected average contribution. If all variables would contribute equally, the expected value would be $1/7 = 14.3\%$. Contributions above the threshold can be considered important to a dimension.

Our goal is to capture the many facets of textual complexity, characterised by the features presented in the previous section, in a two-dimensional model. With this in mind, we perform dimensionality reduction using PCA to see which features would load onto a model with two main components. The first two principal components explain 70.5% of the variation in the data, which we consider an acceptably large percentage. Next, we determine how much each variable is represented in each component. Figures 1 A) and B) show the variable loadings for the two extracted principal components. In Figure 1 A), we can see that it is primarily measures of lexical diversity and richness like Yule's I, MATTR and OVIX that load on the first component (Dim-1). Thus, we consider Dim-1 to represent semantic complexity. Figure 1 B) shows that the average sentence length, the syntactic depth and LIX load more than average on the second component (Dim-2). Given that these metrics can be considered to measure syntactical characteristics, we consider Dim-2 to represent syntactical complexity. Each article is characterised by a syntactical and semantic complexity value (principal component score), which we analyse now regarding the research questions we presented in section 1.

3. Findings

In the following, we first present the diachronic development of text complexity. Next, we use additional analyses to get further insights into possible explanations for the changes that

semantic complexity undergoes. Finally, in section 3.3, we take a closer look at the effect of outlet type on the complexity dimensions.

3.1. The evolution of text complexity

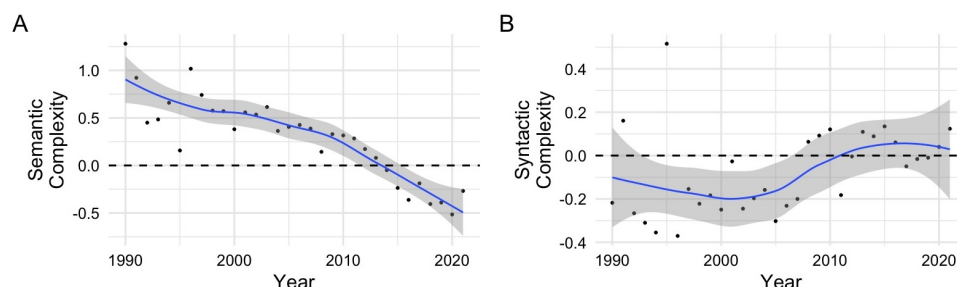


Figure 2: Average A) semantic and B) syntactic complexity over time, with LOESS regression lines indicating the trends.

Our sample of CC articles covers 32 years. Given this long time frame, we assume that writing variations must occur and affect the complexity dimensions. Figures 2 A and 2 B visualise the development of average semantic and syntactic complexity over time. It becomes evident that while semantic complexity decreases steadily from principal component score values of around 1 in the early 90s to -0.5 in 2015 and later, syntactic complexity increases, especially from 2008 onwards. This correlates with seeing more web articles from public service providers.

We use the Mann-Kendall test to assess whether the two complexity dimensions' downward and upward trends are significant. The Mann-Kendall test can be interpreted as a non-parametric rank-correlation test (Kendall's τ) that tests for ordinal association between time and the variable of interest. The test was statistically significant for both semantic and syntactic complexity (semantic complexity: $z = -5.60, n = 32, p \ll 0.00$; syntactic complexity: $z = 3.13, n = 32, p \ll 0.00$). With $\tau = -0.70$, semantic complexity shows a strong monotonic downward trend. The trend for syntactic complexity is not as pronounced. However, syntactic complexity increases slightly over time ($\tau = 0.39$).

To better understand whether variations in the different outlet types can explain these overall trends, we created Figure 3. Figure 3 A) visualises the temporal evolution of semantic complexity for each news outlet category. Interestingly, all four types exhibit somewhat different patterns. While articles from quality papers show a decreasing trend over time, articles from niche papers become more semantically complex and stay constant for a while before declining in the last years of the sample period. Most notable, however, is the development of public service articles, which exhibit a rapid decrease. We consider this evidence of the fact that the overall decrease in semantic complexity can most likely be attributed to the emergence of online articles by public service providers and less due to a strong decrease in semantic complexity among other newspaper types.

Figure 3 B) confirms some of our expectations about syntactic complexity. For instance, it is above average for articles among niche papers, while tabloid articles are below average across the entire sample period. Contrary to semantic complexity, the temporal patterns for syntactic

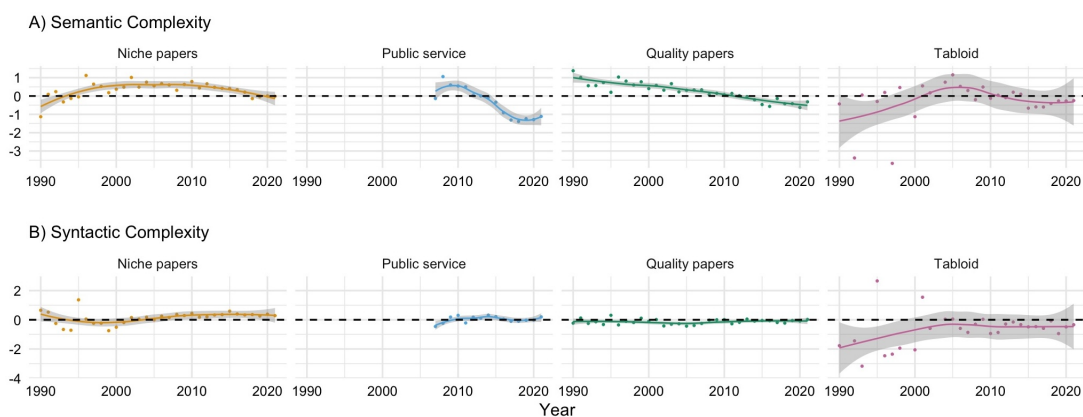


Figure 3: Average A) semantic and B) syntactic complexity over time split by type of news outlet.

complexity are surprisingly stable over time. This is an interesting positive observation as it means that the articles on CC have not become more difficult to read. Nevertheless, again, public service articles stand out as their emergence explains the increase in syntactic complexity observable in Figure 2 B).

To sum up, it is primarily online articles by public service providers that stick out and are responsible for the temporal changes in text complexity. We investigate semantic complexity further as we observed the most striking change for this dimension.

3.2. A closer look at semantic complexity

Besides public service media, other news outlets, like niche papers and quality news, also exhibit a decline in semantic complexity. To find potential explanations for this development, we perform two additional analyses. The first is a study of vocabulary usage, which can shed light on whether word usage has become more generic. The second is a part-of-speech (POS) analysis to study whether writing style has become more conversational and informal.

3.2.1. Analysis of vocabulary usage

A possible explanation for the decrease in semantic complexity could be that language use in media coverage of CC has become more generic over the years. At the beginning of the sampling period, CC reporting was dominated by the notion of *greenhouse effect*, which is mainly associated with technical and scientific terms. Gradually, greenhouse effect was replaced by the term *global warming* during the 2000s and later with the current and broader term *climate change*. We assume that more general vocabulary being used in recent years can be explained by this gradual shift towards the notion of climate change, which is less scientific and more generic. Additionally, this could be paralleled by journalist's efforts to describe climate change in layman's terms and by using more simple language. Either because of editorial requirements, which impose a standard language pattern on each article, reducing the number of synonyms, colloquial language and jargon to a minimum (Tolochko and Boomgaarden

2018), or simply because they want everyone to understand the causes and effects of human-created greenhouse gas emission. We expect that these efforts will decrease vocabulary richness and, thus, decrease features like MATTR and Yule's I, which contribute highly to semantic complexity. We investigate this explanation further by calculating the normalized pointwise mutual information (NPMI) for each word type (Lucy et al. 2023).

As publishing activity in the 1990s is sparse, we group our sample period into five chunks. Using these periods, we assess whether language use has become more generic when communicating CC. Words specific to a certain period will occur more frequently during this period and consequently have a high NPMI. A word whose occurrence is decoupled from a specific period and has an NPMI close to 0 is a more generic word. Following this definition, we can look at the distribution of NPMI values of all words within each period. Periods with many NPMI values of 0 or close to 0 use more generic language than periods with negative/positive NPMI's with a more specialised vocabulary. Finally, we sampled 1000 documents from each period to avoid over-representing articles from periods with high publishing activity. We also set the number of minimum occurrences of a word to 25. The NPMI of a word t in a period p is calculated as follows:

$$NPMI_p(t) = \frac{\log(P(t|p)/P(t))}{-\log P(t,p)}$$

$P(t|p)$ is the probability of a word t occurring in a news article from period p . $P(t,p)$ is their joint probability and $P(t)$ is the probability of the word overall (Lucy et al. 2023).

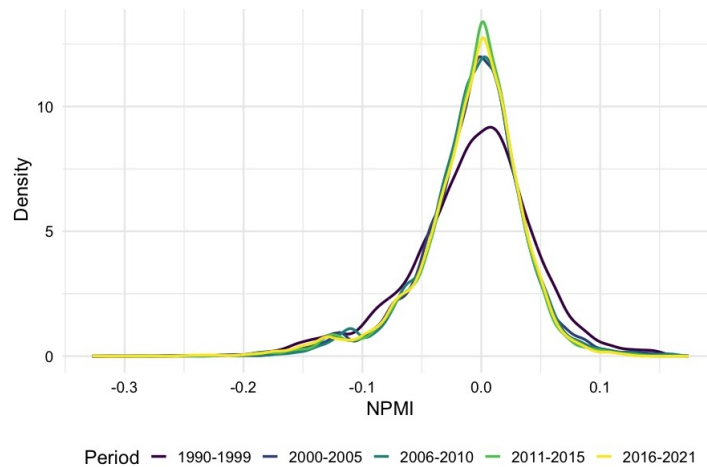


Figure 4: Distribution of NPMI values for each publishing period. During the first period (1990-1999) more specialised vocabulary is used.

Figure 4 shows the distribution of NPMI values for each publishing period. We observe that word usage between the periods is different, as the density curve for the period 1990-1999 is flatter with a slight tendency to the right. The two most recent periods, 2011-2015 and 2016-2021, have high peaks at 0, indicating a higher frequency of words that are used across all periods. Thus, while we can see some differences between the periods and a slight trend of language use becoming more generic, the differences seem only minor and don't allow for final conclusions.

3.2.2. Part-of-speech analysis

To investigate the development of semantic complexity in more detail, we turn to an approach used by Yasseri, Kornai, and Kertész 2012 which compare POS statistics for each publishing periods. For POS tagging we again used DaCy.

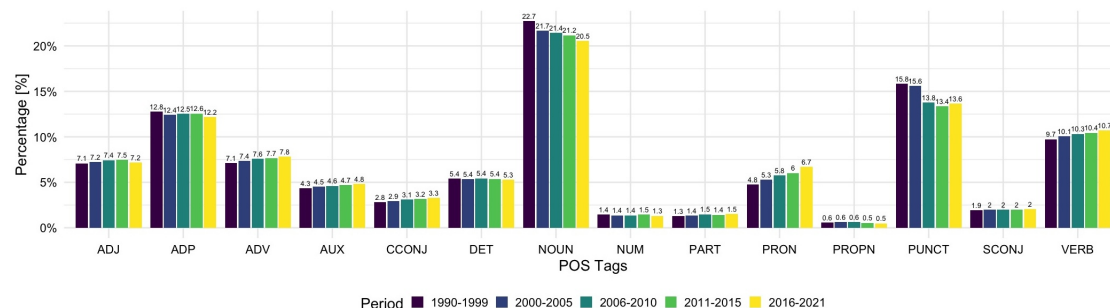


Figure 5: Share of part of speech (POS) tags for each publishing period. The abbreviations are defined as ADJ: adjective; ADP: adposition; ADV: adverb; AUX: auxiliary; CCONJ: coordinating conjunction; DET: determiner; NOUN: noun; NUM: numeral; PART: particle; PRON: pronoun; PROP: proper noun; PUNCT: punctuation; SCONJ: subordinating conjunction; VERB: verb.

Figure 5 shows the share of POS tags for all five periods. With a ratio of around 20% in each period, every fifth word tends to be a noun (NOUN), which seems intuitive as news articles frequently refer to named entities, like persons or organisations. Over time, however, we see a drop in the use of nouns and an increase in pronouns (PRON) from around 4.8% in the 90s to 6.7% in the most recent period. Together with an increase in verbs (VERB; from 9.7% to 10.7%) and adverbs (ADV; from 7.1% to 7.8%) these trends might indicate a movement towards a more conversational, informal writing style.

Finally, while the above analysis has mostly focused on semantic complexity, we also find an increase in sentence length and, thus, indirectly increasing syntactic complexity. This is additionally supported by a slight uptick in the use of conjunctions (CCONJ) (e.g. *and*, *or*, *but*), which increases the average number of words per sentence, and thus contributes to making sentences longer and more complex.

3.3. The influence of news outlet type on text complexity

Finally, we were interested in how text complexity varies by news outlet. Our sample features four different outlet types, which cater to different audiences and are characterized by different properties in terms of language and writing styles. Previous work and common knowledge about outlets characteristics allow us to hypothesise about the picture that might emerge. For instance, as niche papers, are read by the intellectual elite, these news items are most likely characterised by high semantic and high syntactic complexity. On the other hand, tabloid news are probably of lower syntactic complexity and are mainly read by working-class and lower-middle-class readers (Baker 2010; BBC 2023). Less is known about public service news, so it will be interesting to see how our model locates it in relation to other news outlet types.

Figure 6 A) plots all articles along the two complexity dimensions. It is evident that most

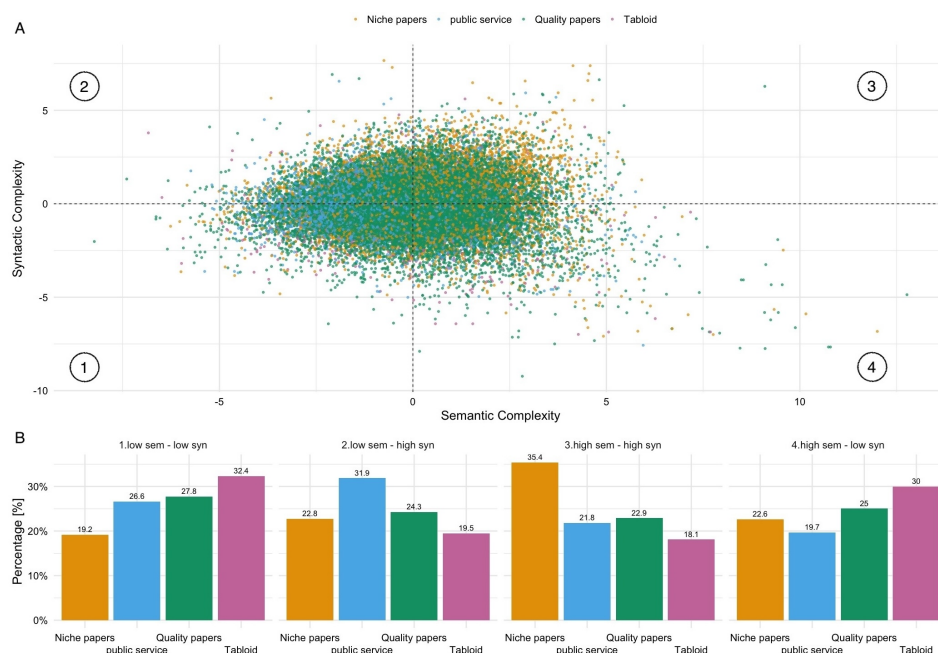


Figure 6: A) Scatter plot of all articles with color representing the different types of outlets. The graph can be divided into four sectors along the syntactic and semantic complexity axes. Figure 6 B) shows the relative share of outlet type in each sector.

articles are positioned around the centre and that extreme principal component scores are scarce. The two axes, representing the two complexity dimensions, divide the articles into four sectors. Interestingly, the sample is divided into almost four equally sized chunks. In sector 1 (low sem - low syn) we find 8238 (25.6%) news items, in sector 2 (low sem - high syn) 7899 (24.5%) articles, in sector 3 (high sem - high syn) 8339 (25.9%) articles, and in sector 4 (high sem - low syn) 7739 (24%) articles.

Figure 6 B) shows the share of news outlet types per sector. It partly confirms our expectations. Sector 3, where both complexity dimensions are above average, is dominated by articles from niche papers (35.4%). Likewise, in sector 1, where the two dimensions are below average, we find that the largest share of news items belongs to tabloid news. Surprisingly, the share of tabloid articles is also highest in sector 4 (30%) where semantic complexity is higher and syntactic complexity is lower than average. Tolochko, Song, and Boomgaarden 2019 describe sector 4 as the ideal scenario in which the basis for knowledge acquisition is highest as this is where complex information (characterised by high semantic complexity) is described easily and concisely (low syntactic complexity). Following this logic, sector 2, where semantic complexity is below average and syntactic complexity is above average, should be avoided. Yet, this is the sector where the share of public service articles is highest.

To better understand these findings, we study whether the observed differences in complexity for newspaper types are statistically significant by assessing if the 95% confidence interval (CI) of the differences in bootstrapped means (1000 re-samples) shows the null effect. Figure 7 visualizes the pairwise comparison between all newspaper types for both complexity measures.

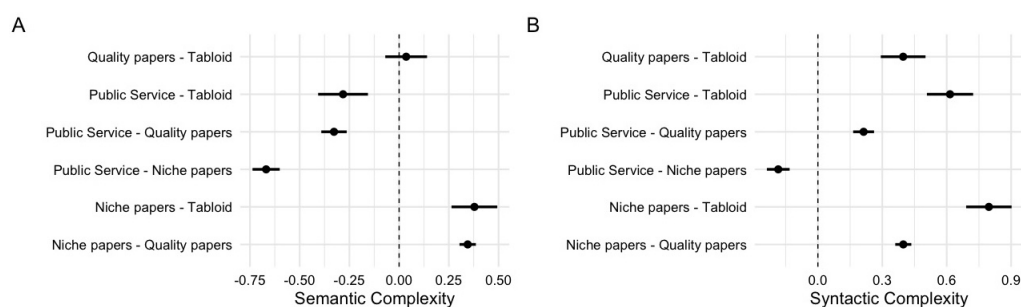


Figure 7: Pairwise comparison between different newspaper types for A) semantic and B) syntactic complexity. The bars indicate the 95% confidence interval.

A significant effect is present if the bars do not overlap the dashed line at 0. The first row in Figures 7 A) and B) show the comparison between articles from quality papers and tabloids. While no difference between the two newspaper types is visible concerning semantic complexity, syntactic complexity is significantly higher in quality papers, as the bars can be found on the right side of the dashed line. On the other hand, semantic complexity is higher for tabloids than public service outlets, hence it can be found on the left side of the dashed line.

We find surprisingly many significant differences, most of them confirming general expectations, for example, that tabloid articles have the lowest syntactic complexity compared to other types of news outlets. The strongest effect is between niche papers and tabloids. The principal component score for syntactic complexity of niche paper articles is, on average, between 0.69 and 0.9 (95% CI) higher than for tabloid articles. Interestingly, public service is the outlet type with the lowest semantic complexity. However, its syntactic complexity is fairly high; only articles published in niche papers are syntactically richer. Thus, the statistical analysis confirms the picture that emerged from the analysis of Figures 3 A) and B) as well as Figures 6 A) and B).

4. Discussion and conclusion

This article presents the analysis of textual complexity in Danish news articles on CC. More concretely, using a sample of around 32,000 news articles from four different outlet types, we build a machine-learning model to capture variations in semantic and syntactic complexity, addressing three research questions.

RQ1 raised the question of how semantic and syntactic complexity have developed over time and how this development varies across news outlet types. We found evidence for a strong decrease in semantic complexity and a slight increase in syntactic complexity. A closer look at how the evolution of the two complexity dimensions varies by outlet types shows that the changes can mainly be attributed to the emergence of articles by public service providers, which impact the overall picture. This finding can be considered both positive but also problematic. First, it means that articles by niche or quality papers have not become more syntactically complex over the sample period. Consequently, their readers are not in need of increased literacy to gain knowledge on CC. However, it also indicates that articles from public service providers have lower semantic complexity and higher syntactic complexity than the average

article. Below, we will discuss the implications of this finding in more detail.

Because syntactic complexity showed little variation, RQ2 focused on further analysing the semantic dimension of text complexity and the changes that, e.g. quality news undergoes. We did so by 1) Identifying whether language use has become more generic over the years and 2) POS analysis to see how the usage of lexical categories has changed. Indeed, we found evidence for less specialised vocabulary being used in more recent times when writing about CC. This aligns with a decrease in semantic complexity as a lack of specialised vocabulary also means a decrease in lexical richness and diversity, measured via MATTR, Yule's I, and OVIX, which all load highly on this principal component. Our POS analysis highlighted an increased usage of pronouns, verbs, and adverbs, which can be interpreted as a trend towards a more informal tone and emotional language. This, however, does not align with the overall trend of decreasing semantic complexity.

Finally, RQ3 studied the differences in complexity between the four types of news outlets in our sample. Many of our expectations are confirmed, and findings by previous studies corroborated (Tolochko and Boomgaarden 2018, 2019). Two of the most relevant findings are: 1) Tabloid articles, being mostly easy to read (low syntactic complexity) and engaging (high semantic complexity), should not be underestimated and disregarded as a channel of telling the climate story with increased urgency. 2) Articles by public service providers are dominant in the category *low semantic complexity - high syntactic complexity*. Indeed, further analysis showed that concerning syntactic complexity, articles by public service providers are only surpassed by articles from niche papers, which cater to an intellectual elite.

Taking all our findings together, it becomes evident that CC reporting in established outlets like tabloids, quality papers, or niche papers has, from a structural linguistic perspective, stayed fairly consistent over the years. We could not detect any evidence for the tabloidization of CC news. However, this picture has changed with the emergence of online articles by public service providers. Consequently, we may consider public service broadcasting a driver of change in terms of public CC communication. Public service articles have, on average, less semantic complexity but higher syntactic complexity, i.e. a combination that, according to Tolochko and Boomgaarden, should be avoided as it limits knowledge acquisition and retention (Tolochko, Song, and Boomgaarden 2019). Indeed, it is problematic that online articles that are daily consumed by millions of Danes, have a higher syntactic complexity, which makes them harder to process, than articles by quality news.

What can explain this tendency in CC news by public service media? Considering the structural constraints of public service corporations in the Danish media system (Willig 2021) is important. As taxpayer's money funds them, they are mandated to stay politically neutral. Consequently, unlike other media outlets, public service media have no opinion pieces, which generally resemble more informal or oral communication strategies. Instead, CC reporting in public service media favours technical and specialized language, which is assumed to be more politically neutral. This is augmented by the tendency of public service news to be rather topical, focusing on news updates and breaking news. As a consequence, CC reporting in public service media is driven by political developments, international summits (e.g. COP meetings), and scientific reports (e.g. by IPCC). However, such news stories also tend to be more technical and scientific, requiring high rather than low syntactic complexity. We are left with a bit of a paradox. While public service media contribute to a generally well-informed public (Curran

et al. 2009), the communicative constraints surrounding CC reporting may prevent a truly engaging modality of CC information.

Our analysis is limited as it is only a quantitative by-proxy account of text complexity. It does not allow us to draw any conclusions about how readers *actually* perceive the complexity (subjective complexity) of public service articles nor how much knowledge retention is possible. Objective and subjective text complexity will likely differ, as studies show that the latter one is being influenced by factors like topic interest (Strømsø, Bråten, and Britt 2010). However, the findings of this study can inform future experiments with reader groups from varying demographics, similar to the ones Tolochko and Boomgaarden 2019 performed, by using the categorisation of the two-dimensional model as a guide in selecting stimulus material. Our study suffers from further limitations that future work should address. For a start, it is not yet well studied which combination of linguistic features and machine learning models have the highest accuracy in capturing semantic and syntactic complexity. Moreover, to enhance generalizability of findings in this area cross national studies would be necessary. Finally, we have to acknowledge that due to the lack of a meaningful baseline, we can not say whether our findings are specific to CC news or document an overall, topic independent trend in news reporting. Addressing these limitations would, without doubt, give us a more holistic picture of text complexity in CC news reporting and thus advance communications and journalism research not only in Denmark but at a global scale.

References

- Baker, Paul. 2010. "Representations of Islam in British broadsheet and tabloid newspapers 1999–2005." *Journal of Language and Politics* 9 (2): 310–338. ISSN: 1569-2159. <https://doi.org/https://doi.org/10.1075/jlp.9.2.07bak>.
- Barkemeyer, Ralf, Suraje Dessai, Beatriz Monge-Sanz, Barbara Gabriella Renzi, and Giulio Napolitano. 2015. "Linguistic analysis of IPCC summaries for policymakers and associated coverage." *Nature Climate Change* 6 (3): 311–316. ISSN: 1758-6798. <https://doi.org/10.1038/nclimate2824>.
- BBC. 2023. *Journalism Analysis - Differences between tabloid and quality press*. <https://www.bbc.co.uk/bitesize/guides/zc3nmnb/revision/2>. Accessed: 2024-01-11.
- 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. <https://doi.org/10.1080/09296171003643098>.
- Cunningham, Clare, Charlotte Foxcroft, and Helen Sauntson. 2022. "The divergent discourses of activists and politicians in the climate change debate: An ecolinguistic corpus analysis." *Language & Ecology*, 1–18. https://www.ecoling.net/_files/ugd/ae088a_0c7b300e9f564506b50f2e2c6eb56bfa.pdf.

- Curran, James, Shanto Iyengar, Anker Brink Lund, and Inka Salovaara-Moring. 2009. "Media System, Public Knowledge and Democracy: A Comparative Study." *European Journal of Communication* 24 (1): 5–26. ISSN: 0267-3231, 1460-3705, accessed September 28, 2023. <https://doi.org/10.1177/0267323108098943>.
- Cushion, Stephen. 2022. "Are public service media distinctive from the market? Interpreting the political information environments of BBC and commercial news in the United Kingdom." *European Journal of Communication* 37, no. 1 (February): 3–20. Accessed October 13, 2023. <https://doi.org/10.1177/02673231211012149>.
- Enevoldsen, Kenneth, Lasse Hansen, and Kristoffer L. Nielbo. 2021. "DaCy: A Unified Framework for Danish NLP." In *Ceur Workshop Proceedings*, 2989:206–216. CEUR Workshop Proceedings. ceur workshop proceedings.
- Esser, Frank. 1999. "'Tabloidization' of News: A Comparative Analysis of Anglo-American and German Press Journalism." *European Journal of Communication* 14 (3): 291–324. <https://doi.org/10.1177/0267323199014003001>.
- Falkenjack, Johan, and Arne Jönsson. 2014. "Classifying easy-to-read texts without parsing." In *Proc. of the 3rd Workshop on P1TR*, edited by Sandra Williams, Advait Siddharthan, and Ani Nenkova, 114–122. Gothenburg, Sweden: ACL, April. <https://doi.org/10.3115/v1/W14-1213>.
- Fløttum, Kjersti. 2016. *Linguistic Analysis in Climate Change Communication*. <https://doi.org/10.1093/acrefore/9780190228620.013.488>.
- Hase, Valerie, Daniela Mahl, Mike S. Schäfer, and Tobias R. Keller. 2021. "Climate change in news media across the globe: An automated analysis of issue attention and themes in climate change coverage in 10 countries (2006–2018)." *Global Environmental Change* 70. <https://doi.org/10.1016/j.gloenvcha.2021.102353>.
- Jung, Jaemin. 2003. "Business News Web Sites Differ from Newspapers in Business Content." *Newspaper Research Journal* 24 (2): 114–119. <https://doi.org/10.1177/073953290302400209>.
- Kleinnijenhuis, Jan. 1991. "Newspaper Complexity and the Knowledge Gap." *European Journal of Communication* 6 (4): 499–522. <https://doi.org/10.1177/0267323191006004006>.
- Lucy, Li, Jesse Dodge, David Bamman, and Katherine Keith. 2023. "Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications." In *Findings of the ACL 2023*, edited by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, 6929–6947. Toronto, Canada: ACL. <https://doi.org/10.18653/v1/2023.findings-acl.433>.
- Moser, Susanne C. 2016. "Reflections on climate change communication research and practice in the second decade of the 21st century: what more is there to say?" *WIREs Climate Change* 7 (3): 345–369. <https://doi.org/10.1002/wcc.403>.
- Nations, United. 2022. *Five ways media and journalists can support climate action while tackling misinformation*. <https://news.un.org/en/story/2022/10/1129162>. Accessed: 2024-01-11.
- Richardson, John E. 2007. *Analysing Newspaper. An Approach from Critical Discourse Analysis*. Houndmills: Palgrave MacMillan. ISBN: 1-4039-3565-3.

- Robertson, Craig. 2022. *How people access and think about climate change news*. <https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022/how-people-access-and-think-about-climate-change-news>. Accessed: 2024-01-05.
- Schäfer, Mike S., and Saffron O’Neill. 2017. “Frame Analysis in Climate Change Communication.” In *Oxford Research Encyclopedia of Climate Science*. Oxford University Press, September. ISBN: 978-0-19-022862-0. <https://doi.org/10.1093/acrefore/9780190228620.013.487>.
- Shulman, Hillary C., Matthew D. Sweitzer, Olivia M. Bullock, Jason C. Coronel, Robert M. Bond, and Shannon Poulsen. 2022. “Predicting Vote Choice and Election Outcomes from Ballot Wording: The Role of Processing Fluency in Low Information Direct Democracy Elections.” *Political Communication* 39 (5): 652–673. <https://doi.org/10.1080/10584609.2022.2092920>.
- Similarweb. 2023. *Top Websites Ranking in Denmark*. <https://www.similarweb.com/top-websites/denmark/>. Accessed: 2024-01-11.
- Skovsgaard, Morten. 2014. “A tabloid mind? Professional values and organizational pressures as explanations of tabloid journalism.” *Media, Culture & Society* 36 (2): 200–218. <https://doi.org/10.1177/0163443713515740>.
- Strømsø, Helge I., Ivar Bråten, and M. Anne Britt. 2010. “Reading multiple texts about climate change: The relationship between memory for sources and text comprehension.” *Learning and Instruction* 20 (3): 192–204. ISSN: 0959-4752. <https://doi.org/https://doi.org/10.1016/j.learninstruc.2009.02.001>.
- Tolochko, Petro, and Hajo G. Boomgaarden. 2018. “Analysis of Linguistic Complexity in Professional and Citizen Media.” *Journalism Studies* 19 (12): 1786–1803. <https://doi.org/10.1080/1461670X.2017.1305285>.
- Tolochko, Petro, and Hajo G. Boomgaarden. 2019. “Determining Political Text Complexity: Conceptualizations, Measurements, and Application.” *International Journal of Communication* 13 (0): 21. ISSN: 1932-8036.
- Tolochko, Petro, Hyunjin Song, and Hajo Boomgaarden. 2019. ““That Looks Hard!”: Effects of Objective and Perceived Textual Complexity on Factual and Structural Political Knowledge.” *Political Communication* 36 (4): 609–628. <https://doi.org/10.1080/10584609.2019.1631919>.
- Willig, Ida. 2021. “Mediesystemteori.” In *Klassisk og moderne medieteori*, edited by Mikkel Fugl Eskjær and Mette Mortensen. Hans Reitzels Forlag. ISBN: 9788741272610.
- Yasseri, Taha, András Kornai, and János Kertész. 2012. “A Practical Approach to Language Complexity: A Wikipedia Case Study.” *PLOS ONE* 7, no. 11 (November): 1–8. <https://doi.org/10.1371/journal.pone.0048386>.

5. Online Resources

The data and R code for this paper are available on GitHub.