Language Change in Text Difficulty, Syntactic Complexity and Lexical Diversity in American Newspapers over the Last 50 Years

Nadja Taha

Psychology (BSc) - Bachelor Thesis

University of Twente - Department of Cognition, Data and Education (CODE)

Track - Research Methodology, Measurement and Data Analysis (RMMD/OMD)

First supervisor - Dr. Hans J.W. Luyten

Second supervisor - Prof. Dr. Ir. Bernard P. Veldkamp

Abstract

Newspapers have evolved many times since their emergence with the introduction of new technologies and their inclusion of more nations and classes in their writing style. In the recent age however, people increasingly rely on social media for information. With the rise of the internet, news is spreading faster and in shorter form. Newspapers need to stay competitive and therefore, it is interesting to know whether the language of the newspaper has changed to match this new digital writing style over time. This has been tested in 100 articles from the New York Times starting from 1974 until 2024 with 50 articles each from the politics section and the sports section. Language change was tested in terms of text difficulty within a measure of U.S. grade level necessary to understand the text, syntactic complexity by calculating the mean sentence length and the clauses per sentence, and lexical diversity, counting the number of unique content words within the text. An interaction model was ran to investigate whether there is an effect of time, as well as an effect of genre (politics vs. sports) on these measures and if these variables moderate each other's effect.

Results showed no effect of time on either grade level as a measure of text difficulty, mean sentence length or clauses per sentence to indicate syntactic complexity. An effect of time was found on lexical diversity with increasing diversity, as well as an interaction of article genre for this effect, as the slope is decreased for sports articles. No other interaction effects were found. Article genre had an effect on grade level, mean sentence length and clauses per sentence with sports articles consistently scoring lower on all these measures. However, no effect of genre was found for lexical diversity. These findings show that newspaper language does not significantly change in terms of structure due to the introduction of the internet, the vocabulary however increases with the new internet language.

Keywords: language change, newspapers, New York Times, linguistic features, digital media, language simplification, lexical diversity, text difficulty, syntactic complexity

Introduction

Traditional newspapers have notoriously been in decline the past decades with the rise of social media. During the 19th and 20th century, printed press was the main distributor of information with people solely relying on the daily or weekly newspaper for news updates (Center for Innovation and Sustainability in Local Media, 2019). With the development of the internet, digital media has become more popular in reprising the role of a news distributor. In response, newspapers have developed digital versions of their articles to be accessed on a subscription basis. However, paid subscriptions for news sources as well as buying physical print has become less attractive to the new generation with a 6% decline in the circulation of both printed and digital newspapers in 2021 compared to the previous year (Konopliov, 2024). Most of newspapers' revenue today comes from digital-only news subscriptions (Konopliov, 2024). This movement is reflected by an average decline of 14% in the circulation of printed newspapers from the 25 biggest U.S. titles in the past year (Konopliov, 2024; Majid, 2024). Additionally, since 2017, there has been an overall drop of over 50% in the user base of printed papers across the U.S. (Statista Market Insights, 2024). On the other hand, about 37% of the U.S. population in 2022 instead turned to social media for news consumption (Konopliov, 2024). With the rise of the internet and smartphones, information is spread fast online. This puts pressure on traditional newspapers, as the ease and speed with which information is accessed through the development of digital media is hard competition for traditional printing. Furthermore, with the spread of digital media, the time people invest in reading longer texts has declined (King & Gerisch, 2009). Focus is a scarce resource nowadays as the youth is used to only watching a video clip of a few seconds at a time or only reading 140 characters per message with the limitations that social media poses on them. This trend is not only limited to social media as other areas of life need to accelerate with time as well. Transportation has to be faster and communication now only takes up a fraction of the time it used to (Rosa, 2013). Rosa quotes Conrad (1999) in saying that "modernity is about

the acceleration of time". As society accelerates and time allocated to one single thing becomes shorter, people feel like they have less time for themselves and invest less cognitive effort and concentration into reading. There has been a decline in reading since 2005 with 26% less children and young adults aged 8 to 18 reading daily (National Literacy Trust, 2023) as well as a 6.1% decline, compared to 2012, of U.S. adults (18 to 64) that read at least one book in the year 2022 (Books+Publishing, 2023; Publishers Weekly, 2023). Thus, technology is pressured to correspond to that decline with increased shortening of video and text in media (King & Gerisch, 2009). It is then interesting to investigate if this trend has an effect on traditional newspaper writing as well. Newspapers need to keep competitive and adapt to this new age of fast and simple information. As news already spread over to social media and seemingly hold themselves better in their short text format there, we suspect a change in traditional newspaper language towards a more simple, straightforward and fast-to-read language as well.

Theoretical Background

While technological and economic development might be one driver of change in newspaper language, there are also other potential explanations that facilitate a shift towards a more easy language in news. One theory that is mentioned by linguists is the shift in reader-writer responsibility. Chafe (1982) explains a distinction of writing in which the responsibility lies with the reader to understand the text (as cited in Hinds et al., 1987). He gives an example of selected Japanese and Chinese text writing, in which writers tend to give less clarifications than is typical in western media. Some prose relies more on inference and the responsibility falls partly on the reader to understand the text. This lies in contrast with other texts in which the responsibility of the communication is solely with the writer. If the text is not understandable, it is because the writer did not make it clear enough. The reader should not have to put effort into understanding (Hinds et al., 1987). While technological evolution exists

everywhere, writing did not seem to simplify as much in Asian countries. Hence, technology cannot be the only driver for change in written language and a shift in reader vs. writer responsibility as a potential cause is worth exploring.

Evaluating the writer responsibility in Western countries, this shift seems to have been more pronounced only since around 1998, with it being mentioned in an influential guideline on financial communication (U.S. Securities and Exchange Commission, 1998). Lutz (2012, as cited in Schriver, 2017), one of the handbook's collaborators, who called it a "radical idea at the time" specifically for lawyers and other officials, references the right to intelligibility in official texts, especially those affecting people's lives and therefore puts it on government officials to communicate in a language that is understandable to everyone. It seems to be a new development in governmental literature for the reader to become more influential (Ham et al., 2019). This trend has caught onto other areas of writing as well with the "Easy language movement". There is an ongoing effort since the 1970s to make language in these areas accessible to everyone. This movement spread in different countries with the elicitor being the idea that administration should present their citizens with accessible and inclusive language. In the UK, the program started with a shift in reader-writer responsibility as well, as after the 1980s, people with intellectual disabilities were not responsible for interpreting official texts anymore, this responsibility was now placed on the writer. Prior, separate resources existed for a more specific audience. This has been replaced by producing texts with the widest possible audience. This is achieved by using an active voice in text, avoidance of abstract nouns and jargon, as well as keeping sentences short. The idea of "Easy Language" texts became general, with organizations such as "Easy News" emerging in 2013, an online newspaper for easily readable news (Lindholm & Vanhatalo, 2021). This news trend started in Flanders with their 1985 publication of the "Wablieft" newspaper, which is now also available online (Wablieft, 2014) and spread to other news outlets. Therefore, we might also observe a general shift in newspaper language over the last 50 years due to a shift in writer

responsibility, the rise of inclusion and an effort to widen the target audience by making texts more accessible to minorities and corresponding new legislations in these areas.

In the Netherlands, this has sparked research as well, as online tools analysing the difficulty of a text for people suffering from reading disabilities were a scarce resource. The Dutch Association for Scientific Research started an important Dutch research project on aspects of accessible language. From this, tools for language measurement and international research on text comprehension in Dutch and English have emerged, such as the T-Scan software by Henk Pander-Maat (Lindholm & Vanhatalo, 2021). Our research will pick up on these efforts and we will utilize the results and research tools on text comprehensibility and feature testing that were developed in light of this movement.

Studies conducted previously on shifting language features in newspaper articles, with a focus on diachronic studies, support a change over time. Most often, the features analysed encompassed the lexical diversity, sentence complexity, Average Sentence Length (ASL), Automated Readability Index (ARI), and the Coleman-Liau Index (CLI). Several studies discovered an increase in lexical diversity (Cook, 2004; Juola, 2003; Štajner & Mitkov, 2011; Stajner & Mitkov, 2012; Westin, 2001) which means that the vocabulary became richer over the years as new words are added. The average sentence length in newspapers on the other hand has decreased, two studies found (Leech & Smith, 2009; Štajner & Mitkov, 2011). This is in line with the decreasing sentence complexity found by the studies of Leech & Smith (2006), Mittmann (2011) and Westin & Geisler (2002), comparing grammatical change in English language within the Brown Corpora. Results have specifically been taken from press and newspaper editorials. They define this change by the decrease of abstract language and an increase in the use of proper nouns. Proper nouns refer to a noun that corresponds to a singular entity instead of a class of multiple entities such as ‘continent’ or ‘planet’ as opposed to the proper nouns ‘Africa’ or ‘Jupiter’ (Wikipedia, n.d.-b). This aligns with other studies’

findings of a less narrative style of writing with more argumentation and information (Mittmann, 2011; Westin, 2001; Westin & Geisler, 2002). Furthermore, the passive voice is less used which makes the sentences easier as well. The study by Leech & Smith (2009) however found an increase in sentence complexity, which in this case is defined by the number of finite predicates increasing. Finite predicates are verbs which indicate tense, as well as person and number of their subject. As opposed to non-finite verbs, they morphologically change their form to convey information (StudySmarter, n.d.). An example would be the verb 'eats' in the sentence 'He eats his dinner.' as it indicates present tense, third person and singular number. This is in contrast to non-finite verbs such as 'The girl wants to shop.', in which case 'shop' is non-finite as it does not change to indicate tense or person. Therefore, this difference can be accounted to the differing definitions in what makes a sentence complex. Other measures to analyse the difficulty of texts are the Automated Readability Index (ARI) and the Coleman-Liau Index (CLI), both indicating the level of education/literacy required to understand the text in terms of sentence and word difficulty by number of words and characters. For the ARI, the studies of Leech & Smith (2009) and Štajner & Mitkov (2011) discovered no change over time, while for the CLI the study of Štajner (2011) determined an increase, meaning that the newspaper texts become more complex. The results were all based on American newspapers and change has been identified between the years 1961-1991. Our study will pick up on that research and present more recent changes in textual features.

Limitations of Previous Research

To this date, there are not many diachronic studies that focus on language change within newspapers. Diachronic studies mostly focus on general change in spoken language instead of newspaper language and studies on this topic that do exist most often are taken from a newspaper corpus from the 1960s-1990s (Westin, 2016). Since the majority of these

studies use the same corpus of newspaper reportages and editorials, there is not much variation in the studies. Štajner & Mitkov (2011) remark that previous studies existing on the topic of language change furthermore most often focus on phonetic and lexical (i.e. pronunciation and vocabulary), rather than stylistic and syntactic changes. However, it is just as important to research syntactic changes in newspaper language as syntactically complex texts pose problems in comprehension for people dealing with a language impairment or second language learners, among others (Leech & Smith, 2009). Additionally, with the restriction of existing studies mostly ending in the 90s, recent technological advances have not been taken into account. This is in addition to the rise of the aforementioned Easy Language Movement which we expect to have driven change in recent media language as well. There has been changes everywhere since the early nineties up until now with developments in economy, education, technology especially, and thus in the availability of information via new media, mass communication and in the general rise in literacy, which influences the vocabulary of our time (Herring, 2003; Juola, 2003). Studies have been conducted about the rate of language change in which the post-war period, with veterans' exposure to new experiences and environments, as well as technological advances, were determined to be the strongest accelerators of change in language (Juola, 2003). However, the question then still remains in what form these changes translate within newspapers. Do new technologies primarily drive lexical or also syntactical change and how does the rate of these changes differ?

Research Questions

It is sensical to conduct a new study encompassing specific textual features' development in time for a better understanding of change and its potential drivers in specifically newspaper language. The research question of the current study thus entails "To what extent did language in American newspaper articles change over the past 50 years with

regard to syntactic complexity, lexical diversity and text difficulty?" Additionally, we follow the research question "How does this change differ between formal articles of the Politics section and informal articles of the Sports section in an upscale newspaper like the New York Times?"

Methods

Data Collection

Data was collected from the newspaper "The New York Times". Articles stem from the Politics section, as well as the Sports section of the newspapers. While retrieving articles from 1974 to 2024, special attention was paid to busy periods within news reporting. Upon research, no significantly calm or recommended period could be found for news reporting with search terms such as "news publication trends", "news cycle trends", "newspaper publishing trends political news cycle", "long term trends newspaper industry", "newspaper news trends year periods politics" and variations thereof in Google Scholar, as well as in a general google search. Therefore, busy periods were researched instead for exclusion of certain time frames. For political articles, election periods such as state elections in November, as well as federal elections, which were found to be mostly in spring (U.S. Vote Foundation, n.d.), were avoided for the similarity and regularity of articles and to keep the data as uniform as possible. This is because exceptional periods and special events, as is the case with elections that happen every four years, might result in confounding differences and skew the data. Articles of each year were kept as comparable as possible. Therefore, articles were extracted from June of each year to represent a relatively calm period in U.S. politics for testing. It was suggested by Canada Newswire (n.d.) however that Wednesdays seem to be a calm day of the week for reporting and therefore, the second Wednesday in June of each year was chosen for the sample. Where there was no political news reporting on that particular Wednesday, another article from the same week was selected.

While extracting articles from the New York Times Archive, the search term “Politics and Government” has been used when filtering for genre is not available. For articles after 1980, a filter for section could be applied which was set to “U.S.”. Articles before that year were manually filtered to be U.S. specific instead of general world politics.

For the similarity and structure in the Sports articles sample set, a similar method of circumventing especially busy periods in sports has been applied. Popular sports activities in the U.S. were researched (Gough, 2023), as well as the timeframe of their respective sports seasons through a simple google search. Most competitive seasons range from March to September and October to June. To avoid the beginning and end of these seasons as it is assumed that reporting will be more focused during these periods, the February has been chosen for this sample. Articles were therefore extracted from the first Monday of the February each year to cover weekend sports news and thus ensure enough news coverage for the sampling.

For each excerpt, a length of around 450 words per article were chosen to ensure that the sampled text is large enough for a proper analysis while also taking calculation time into account. Additionally, articles of the year 1980 are not included in both genres as these were inaccessible.

Data Analysis

For the analysis of textual features, different programs have been used. ARTE has been used to determine the difficulty of the texts based on word and sentence length, TAALED is used to analyse the lexical diversity while TAASSC gives information on the syntactic complexity of the articles. R version 4.4.1 was then used to further analyse the data within a linear model.

Within ARTE (Choi & Crossley, 2022), the Flesch-Kincaid-Grade Level has been used to calculate the difficulty level of a text. The Flesch-Kincaid-Grade Level consists of a constant that is subtracted from the word length (number of syllables per words) and the average sentence length with their respective coefficients (Flesch, 1948). The formula is as follows: *Grade Level = .39 (words/sentence) + 11.8 (syllables/word) – 15.59* (Siteimprove, 2024). It is based on the Flesch-Readability Ease, however, simplified for use and immediately giving a grade level without the need to convert the score. It correlates highly with other formulas for text difficulty like the original Flesch-Ease-Readability score as well as the Automated-Readability Index (ARI) (Kincaid et al., 1975), which can be calculated by ARTE. Therefore, a use of multiple formulas would be redundant and the Flesch-Kincaid-Grade Level, due to its ease and reliability, is superior to the other indexes (Kincaid et al., 1975). The score relates to the U.S. grade level necessary to understand the text, thus, a score of 12 indicates 12th grade (18 years), while a higher score indicates college level and above (Siteimprove, 2024). Therefore, the higher the score, the more difficult the text.

TAASSC has been used to calculate the mean length of sentences by dividing the number of words in the text by the number of sentences. Furthermore, the clauses per sentence are calculated by dividing the number of clauses in the text by the number of sentences, thus giving information about the syntactic complexity of the text (Kyle, 2016). A higher score relates to a higher syntactic complexity for both measures.

Lastly, TAALED analyses the lexical diversity of a text (Kyle et al., 2021). To calculate the lexical diversity, i.e. the number of types per tokens, the measure of textual lexical diversity (MTLD) has been used. Lexical diversity is calculated as the ratio of types (unique words) per tokens (total words) in a text (Kyle et al., 2021). MTLD was proven to be superior to other measures of lexical diversity, such as the simple type-token ratio (TTR) given its independence from text length (Kyle et al., 2021). Additionally, it does not rely on an

arbitrary chosen segment of words, proving it superior to the moving average type-token ratio (MATTR) as well (McCarthy & Jarvis, 2010). MATTR uses multiple equally and overlapping sized word windows and takes the average of each window to circumvent the issue of instability across different text length (Kyle et al., 2021). However, as MATTR is dependent on the selected window size, it can skew results in texts with a repetitious internal structure, such as news stories, as windows can coincide with a series of identical tokens (Covington & Mcfall, 2010). MTLD does not use word segments, but utilises TTR factors at the point of stabilization (McCarthy & Jarvis, 2010). As the number of words (i.e. tokens) increases as the text becomes longer, the number of different words (i.e. types) decreases steadily, since less and less new word types are introduced. The point of stabilization refers to the point where no additional types affect the TTR trajectory. It neither falls nor rises in value and reaches a type saturation in the following sequence. MTLD tests how many words it takes to reach this saturation (McCarthy & Jarvis, 2010). Thus, the measure of MTLD is more stable and is tested to have a higher sensitivity and validity (Kyle et al., 2021). The MTLD value therefore consists of "the average number of words required for the text to reach a point of stabilization" with a lower score indicating a less diverse text (McCarthy & Jarvis, 2010, p. 386). For this analysis, we have used the MTLD specifically for content words, as this deemed more meaningful than calculating the MTLD on all words in the text as the stabilization point was reached pre-emptively due to repeating function words. Content words include all words that add meaning to the sentence, whereas function words denote all other words and show grammatical relationships between the content words. Typical examples of function words are therefore prepositions, conjunctions and pronouns such as "in, you, but", while content words include all nouns, verbs, adjectives and adverbs (Wikipedia, n.d.-a). The content word MTLD gives more information about the actual lexical diversity of the text.

For further analysis, R was used with the packages *tidyverse*, *broom*, *GGally* and *modelr*. Firstly, descriptive statistics were calculated to get distributions of the different

results. Following, bivariate correlations between the tested dependent variables were drawn up to investigate meaningful relationships between the variables. Ultimately, all the variables are meant to be a measure of text complexity to validate the research question if newspaper language changed over the past 50 years. A regression model was then used to test whether the dependent variables *Flesch-Kincaid-Grade Level (in our dataset: FKGL)*, *mean length of sentence (MLS)*, *clauses per sentence (CS)* and *measure of textual lexical diversity for content words (MTLD_original_cw)* are influenced by y*ear* and *type of article* (i.e. main effects), and whether there is an interaction effect between *year* and *type of article* (i.e. interaction effect). After investigating assumptions of normality, independence, linearity and equal variance, an interaction model seemed the most fitting for our data. The R script can be found in Appendix A.

Results

Descriptive Statistics

Mean scores, standard deviations, as well as minimum and maximum scores to give the range of the dependent variables can be found in Table 1. FKGL scores correspond to American grade levels, MLS and MTLD scores depict word counts informative towards their respective measures, while CS shows the amount of clauses in sentences. All scales range within the middle values resembling a normal distribution. Histograms showcasing the distributions of the dependent variables can be found in Appendix B. Excerpts of the articles corresponding to the minimum and maximum scores of the variables can be found in Appendix C in order to exemplify the meaning of the measures.

Table 1

Means, standard deviations, minimum and maximum of the scales as well as correlations for each dependent variable

Variable	*M*	*SD*	*Min*	*Max*	FKLG	MLS	C_S	MTLD
FKGL	11.75	3.44	5.77	22.01	1	0.89***	0.61***	0.16
MLS	25.39	6.1	14.47	50.11		1	0.76***	0.07
CS	2.12	0.46	1.19	3.89			1	-0.06
MTLD	180.49	63.78	68.38	356.58				1

*Correlations with * are significant at p < .05,*
*correlations with ** significant at p < .01*
*and correlations with *** significant at p < .001*

To measure the bivariate relationships between dependent variables, Pearson's correlations were calculated as can be seen in Table 1. FKGL, MLS as well as CS are all strongly positively correlated. Thus, they all represent a measure of sentence complexity, which correlates with the FKGL measure of reading difficulty. This correlation is sensical given that average sentence length is a component of all three of these measures. MTLD does not correlate significantly with the other variables, as it says something about lexical diversity as opposed to the text's syntax. There is a slight negative correlation with clauses per sentence, meaning that the more clauses the sentences have, the less new words will be introduced.

Inferential Statistics

To investigate the effect of year on reading difficulty (FKGL), syntactic complexity (MLS, CS) and lexical diversity (MTLD), a linear model was run. This model included

interaction and main effects for the different types of newspaper articles (politics and sports) to measure if the year effect is moderated by article type. The results are presented for each dependent variable individually.

Flesch-Kincaid-Grade Level

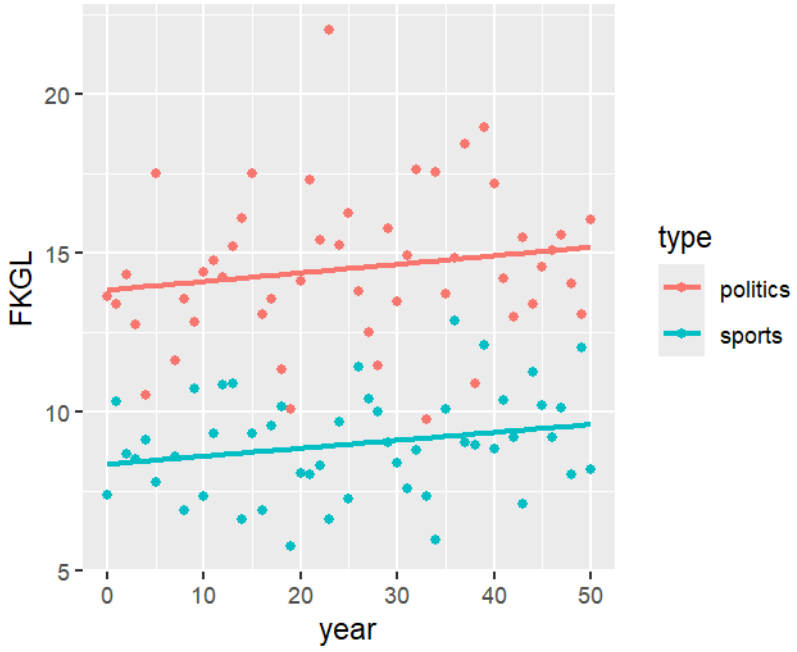
Analysing the change in grade level over the last 50 years, no significant main effect could be detected with $b = 0.03$, $SE = 0.2$, $t(96) = 1.38$, $p = .17$, 95% CI [-0.01, 0.07]. The U.S. grade level needed to understand the articles did not significantly increase over the years.

Type of article has a significant main effect on grade level, as $b = -5.48$, $SE = 0.81$, $t(96) = -6.73$, $p < .001$, 95% CI [-7.09, -3.86] with sports articles requiring a lower grade level than political articles. The grade level required to read a sports article is thus 5.48 grades lower than the grade level required to be able to read a political article.

The interaction effect of article type by year on grade level for readability is non-significant with $b < 0.01$, $SE = 0.03$, $t(96) = -0.08$, $p = .94$, 95% CI [-0.06, 0.05]. Therefore, type of article does not influence the effect of year on grade level. The regression model explained 66.31% of variance on grade level for readability in this sample. The scatterplot visualising the relationship of the Flesch-Kincaid Grade Level with time for sports and political articles with added regression lines can be found in Figure 1.

Figure 1

Scatterplot Showcasing the FKGL Distribution



Mean Length of Sentence

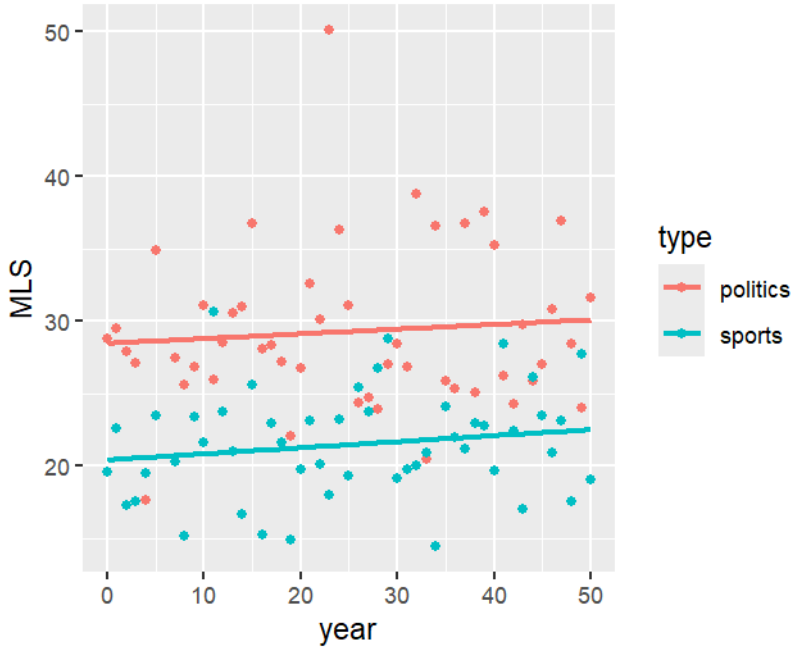
No significant main effect of year on mean length of sentences could be found as b = 0.03, SE = 0.05, $t(96)$ = 0.71, p = .48, 95% CI [-0.06, 0.12]. This means there is no significant increase in the sentence length of New York Times articles over the past 50 years.

The main effect of article type on the mean sentence length is significant with b = -8.05, SE = 1.89, $t(96)$ = -4.26, p < .001, 95% CI [-11.8, -4.3]. The sentences of sports articles are on average 8.05 words shorter than those of political articles.

The interaction effect of article type by year on the mean length of sentences is insignificant with b = 0.01, SE = 0.07, $t(96)$ = 0.15, p = .89, 95% CI [-012, 0.14]. Thus, there is no moderation by article type for the effect of year on mean sentence length. The regression model explained 42.21% of variance on mean length of sentence in this sample. A scatterplot visualising the relationship of year on mean sentence length can be found in Figure 2.

Figure 2

Scatterplot Showcasing the MLS Distribution



Clauses per Sentence

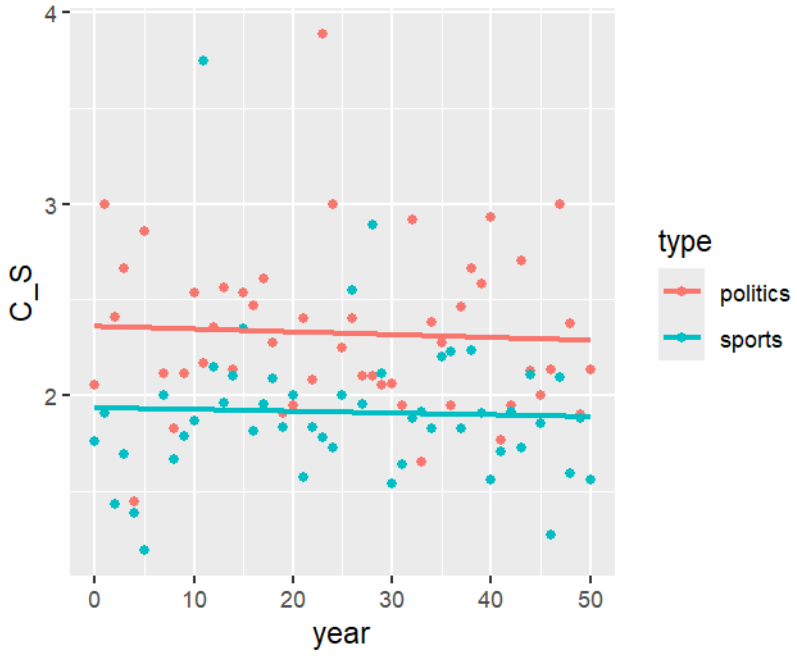
There is no significant main effect of year on clauses per sentence as $b < -0.01$, $SE < 0.01$, $t(96) = -0.36$, $p = .72$, 95% CI [-0.01, 0.01]. Therefore, there is no decrease of clauses per sentence in the newspaper articles over the years.

The type of article main effect on clauses per sentence is significant with $b = -0.43$, $SE = 0.17$, $t(96) = -2.53$, $p = .01$, 95% CI [-0.76, -0.09]. The amount of clauses per sentence are on average 0.43 fewer in sports articles than in political articles.

No significant interaction effect of type of article by year on the clauses per sentence measure could be found with $b < 0.01$, $SE = 0.01$, $t(96) = 0.09$, $p = .93$, 95% CI [-0.01, 0.01]. The type of article is therefore not moderating the relationship of year on clauses per sentence. This model explained 20.21% of the variance in the clauses per sentence. A scatterplot visualising the main effect of year on clauses per sentence can be found in Figure 3.

Figure 3

Scatterplot Showcasing the CS Distribution



Measure of Textual Lexical Diversity (for Content Words)

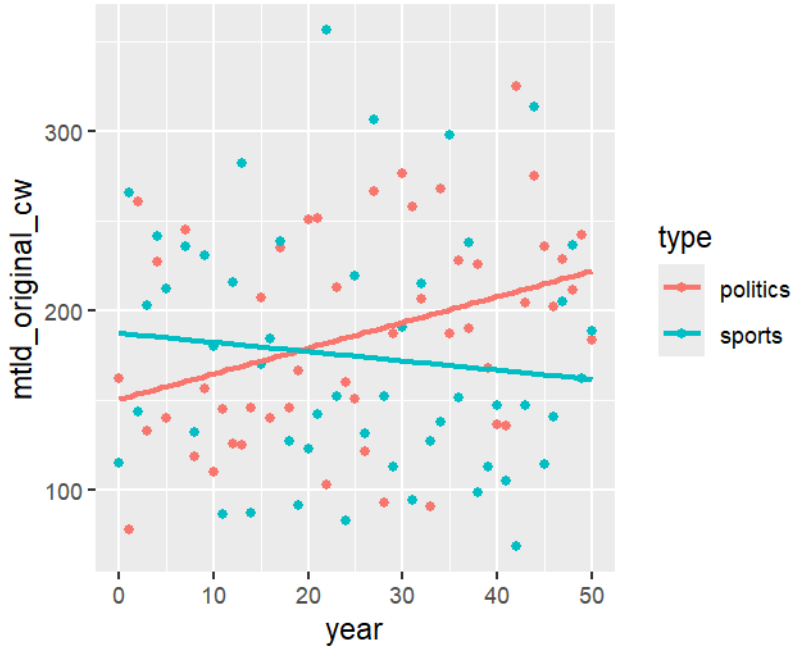
The main effect of year on lexical diversity is significant with b = 1.43, SE = 0.6, t(96) = 2.37, p = .01, 95% CI [-0.01, 0.01]. The lexical diversity and number of unique content words within political articles rises with the year by 1.43 words. No significant effect is found for sports articles (see interaction effect).

Type of article did not have a main effect on lexical diversity and is insignificant with b = 36.9, SE = 0.6, t(96) = 25, p = .14, 95% CI [-12.8, 86.6]. The number of unique words in political articles is not significantly higher than those in sports articles.

The year of article by type interaction effect is significant as b = -1.95, SE = 0.85, t(96) = -2.28, p = .02, 95% CI [-3.64, -0.25]. The type of article moderated the year effect on lexical diversity with -1.95. Thus, the effect of year on lexical diversity is decreased for sports articles than for political articles. The regression model explained 7.13 of the variance in lexical diversity in this sample. The corresponding scatterplot can be found in Figure 4.

Figure 4

Scatterplot Showcasing the MTLD Distribution



Summary

The research question investigated to what extent American newspaper articles change over the last 50 years in regards to their text difficulty, syntactic complexity and lexical diversity. Only a change in lexical diversity was found with political articles becoming more diverse over the years. No significant effect was found for overall text difficulty or syntactic complexity.

It was further investigated if different article types have different effects on text difficulty, syntactic complexity and lexical diversity. An effect of article type was found for text difficulty and syntactic complexity with sports articles scoring lower on each measure. However, article type had no effect on the lexical diversity of the texts.

Lastly, this research also examined how this change over the years is affected by the type of article. Support for the moderator effect of article type with year of publishing on the measures was only found for lexical diversity, in which a significant effect was found for a rise in political articles but none for sports articles. No support was found for moderator effects on text difficulty or syntactic complexity.

Discussion

This study focused on the effect of time on change in American newspaper language in terms of text difficulty, syntactic complexity and lexical diversity. It further investigated the effect of article type on these measures and whether the effect of time would differ depending on the type of article. The results have shown that there is a time effect on lexical diversity with the diversity rising over the years and that this effect is influenced by type of article as well. Sports articles are negatively influencing the rise in diversity, showing a non-significant decrease of unique words over the years. Additionally, an effect of article type was found for text difficulty and syntactic complexity with each of the measures scoring higher for political articles than for sports articles. However, no effect could be found for time on text difficulty or syntactic complexity.

Considering the results, no evidence was found that newspaper language simplifies over time in terms of the literacy level necessary for understanding the text, nor in terms of sentence length or clauses per sentence. This finding in text difficulty is supported by the studies of Leech & Smith (2009) and Štajner & Mitkov (2011) who used ARI to validate their results. No change was found in American newspaper articles of the Brown Corpora (reportages and editorials from 1961 to 1991). However, the study of Štajner (2011) reports positive findings in text difficulty using the Coleman-Liau Index (CLI) of text readability. The same corpora of texts has been used, albeit change was only discovered for editorials specifically. This further validates the distinction of text genre as different types of articles show differing results. It is important to avoid grouping of different text types as results can be masked by heterogeneity of changes, especially if the corpora consists of an unbalanced distribution of genres. A multilinear regression model with an interaction effect yields the most reliable results, accounting for different changes per text group.

Prior research has found mixed results for syntactic complexity, in this study measured by mean sentence length and clauses per sentence. The studies by Leech & Smith (2006), Mittmann (2011) and Westin & Geisler (2002) report a decrease in sentence complexity, while the results of Leech & Smith (2009) show an increase in sentence complexity. However, the studies by Leech & Smith (2009) and by Štajner & Mitkov (2011) found a decrease in the average sentence length, which is also a measure of syntactic complexity. These mixed results might be due to different features being tested. There are multiple ways to measure the complexity of a sentence as there are only scarce studies testing what exactly encompasses the different dimensions of text complexity. Ham et al. (2019) for example lists features that were proven to have an effect on text comprehension consisting of only a selection of the vast measures for lexical complexity and concreteness, as well as syntactic complexity. Leech & Smith (2009) measured their rise in sentence complexity through changes in the number of personal pronouns, noun phrases etc., while the decreasing results were tested on grammatical changes and thus focused more on syntactic complexity (Leech & Smith, 2006; Mittmann, 2011). Thus, the body of research on this topic is not conclusive as no uniform measures have been used to analyse the change in newspaper language.

This research has determined a change in lexical diversity for political articles over time. This aligns with the studies conducted by Cook (2004), Juola (2003), Štajner & Mitkov (2011), Stajner & Mitkov (2012) and Westin (2001). However, it stands in contrast to the findings from Štajner & Mitkov (2011). This, again, shows the inconsistency in measurement method for the same construct, as the same corpora have been used in the listed studies. Some sources use an index for lexical richness, rather than lexical diversity, while others use a different measure for lexical specificity. Nevertheless, lexical diversity is the measure with the least inconsistency in literature, as most prior studies report a positive effect. This effect can be explained by colloquialization. Mittmann (2011) suggests that changes in colloquial speech affect writing. A structural decline was theorised to be part of this colloquialization process by

the authors, however, not supported by this research. Nonetheless, Cook (2004) supports the notion of vocabulary expansion due to colloquialization. Especially the youth, who are named to be accelerators of linguistic change, bring new slang or expressions into written language due to their exposure to technology (Cook, 2004). Examples of this being “unfriend, cyberstalking, etc.”, terms that did not exist before the end of the 20th century. Generally, in an age of technological innovation, new words to describe these innovations accompany its emergence (Herring, 2003; Kristiansen et al., 2011; Matheson, 2000). Additionally, conversations that used to take place face-to-face now exist in written form online (Kristiansen et al., 2011). This causes colloquial speech to enter written media.

Another explanation for this change might be the trend of densification. The study by Mittmann (2011) found that, over the years, press moved towards a more information-packed style, especially by the usage of noun phrases. Since the emergence of digital mass media and the resulting information overload of recent decades (Mittmann, 2011), newspapers have to keep their articles compact to stay competitive. Parsing through the archive during data collection, a trend towards shorter articles in recent years was noticeable, as it became harder finding articles reaching the excerpt size of 450 words. Thus, through the densification and higher usage of noun phrases, an increasing measure of textual lexical diversity for content words is sensical, as a shorter text still has to convey the same information. With a more densely written article, that might barely exert 450 words, the whole scope of the information is given and therefore, more diversity can be found in that excerpt while older less dense texts dwell on a certain topic longer, not introducing new words within the chosen excerpt.

Generally, on all measures but lexical diversity, sports articles scored significantly lower than political articles. Sports articles require a lower literacy level to understand them, they have shorter sentences and less clauses per sentences, thus are less syntactically complex. Not many studies investigated the difference in text difficulty of different article types,

however, given the divergent audiences, it is expected that sports articles are simpler than political articles. Newspapers cater towards their audience and as they evolve over the years to keep up with new jargon and the new generation, they also cater to the different readership profiles (Westin, 2016). The study by (Eriksson, 2017) compared sports articles of different sports, while the study by (Ljung, 1997) compared sports articles to news articles. Both studies found that mainstream sports articles followed a more involved writing style to keep readers engaged. The process of colloquialization seemed to be the strongest for sports articles with features often found in spoken interaction instead of a formal writing style (Eriksson, 2017; Ljung, 1997). This might explain the missing difference of lexical diversity between text genres, although a slow, non-significant decrease can be seen for sports article indicating they also become simpler in terms of word usage. Sports articles are much more informal and its topics are less complicated, thus, require less text difficulty than political news. Their aim is to entertain and inform a wide audience, the focus therefore lies on accessibility and readability (Ljung, 1997), resulting in lower scores of text difficulty and syntactic complexity compared to formal political news articles. This is in line with the general finding that sports articles seemed to be shorter than political articles, given that a lot of initial articles needed to be discarded due to them not reaching the 450 word requirement for this sample. However, further research is necessary to deem this difference statistically significant.

No evidence has been found that newspaper language became easier since 1974. This stands in contrast to most earlier conducted studies that found significant effects. These studies however tested an earlier time period as articles were extracted mostly from 1961 to 1991, in some cases even earlier. Several studies have determined that changes in newspaper language occurred in the 70/80s with movements like the Easy Language Movement and the switch of newspapers from being narrative to more matter-of-fact (Lindholm & Vanhatalo, 2021; Matheson, 2000). This has emerged a new way of writing, now also considering wider

audiences. Newspapers moved beyond specific social groups across classes and nations, hence, their language adjusted to that. With now attracting second language speakers, as well as the less educated, newspaper language simplified and became more comprehensive (Matheson, 2000). Matheson (2000) already noted in the 70s that journalism started to become more structed and articles were less of a mere recording of events and more of a uniform, comprehensive news story. This might also be due to the shift in reader-writer responsibility, which started around the same time (Hinds et al., 1987; Lindholm & Vanhatalo, 2021). Thus, since newspaper as we know them now seemed to have emerged just around the start of our testing pool, we do not see a simplification of language. Earlier studies recorded this movement within their results, a more recent study could not replicate those results.

Hence, a larger scope and testing frame might elicit different results and encompass times of language change and newspaper evolution. A larger testing frame can also help in revealing larger patterns of language change, as well as the exact time of change more accurately in order to say more about possible reasons. Additionally, inferences cannot be drawn accurately from this study as the sampling size is too small. Only one article has been extracted per year and per type, which results in a large standard error. Due to the scope of the paper, only a small amount of articles could be collected, which should be corrected for in a replication study.

This study also revealed inconsistencies in measurement of language features. There exists a vast number of measurement formulas for the same linguistic feature resulting in different results. There is no proper operationalisation of the tested features as for some studies, sentence complexity encompassed syntactic features at the sub-sentence level with measurements of the number of finite clauses per sentence, etc., while for other studies, the number of passive sentences and alike have been tested. This causes great variation in results

of different literature and might explain why this study was not successful in replicating the significant results of other research.

However, studies about the influence of the current digital age on written media also revealed that the technological influence is not as grave as expected. The digital writing style that emerged from the recent media evolution used on e.g. social media posts or personal messaging does not translate to institutional language (Kristiansen et al., 2011). Netspeak features do not transfer to other domains, however, a new domain has emerged from the digital revolution (Kristiansen et al., 2011). Thus, the impact of the internet is not a simplification of language in e.g. newspapers, but the creation of new outlets with a different language style. As digital media gives everyone the opportunity to publish and have a voice with a wide range, more unregimented and unregulated writing reaches publicity. Public writing emerges, such as blog entries, reader comments, etc. This now coexists with institutional language, which is still subject to editing and correcting (Kristiansen et al., 2011) and thus, the internet is not accelerating processes of change as has been reflected in this studies' results. Changes and innovation in written digital media have no influence on standard language structure apart from lexical innovations, which has been supported by this research, as they stay within that domain (Kristiansen et al., 2011).

Although this study has revealed important implications for the influence of digital media on newspaper language, as well as problems in measurement and comparison with other studies due to non-conform definitions of language characteristics, much can be learned from its limitations by making recommendations for future research. Firstly, this study only focused on a limited range of features. The scope of language complexity is wide, with many different measures encompassing the actual difficulty. Ham et al. (2019) lists some features that they have proven to influence ease of understanding. This study only tested four different features and neglected many other measures necessary to successfully determine if, and in

what ways, text simplification took place. Additionally, the features tested in this study did not cover a wide range of dimensions. Similar results in the measures of grade level, mean sentence length and clauses per sentence were to be expected as each of these features bases their measures on sentence length itself. Thus, the measure of mean sentence length influences the results of the other features tested as well, as can be seen from their high correlations. Therefore, factually, only two distinct dimensions have been covered in this study. More measures need to be included, such as i.e. the usage of passive sentences, personal pronouns and the concreteness of words to infer about the actual simplicity of the texts. Therefore, future studies should utilise the list provided by Ham et al. (2019) and expand their scope of features in order to more reliably say something about the change in simplicity of newspaper language.

Furthermore, the study is not exhaustive in terms of the tested variables. This can, i.e. also be seen in the R^2 of the positive MTLD effect. As the explained variance is very low, more variables are responsible for the seen effect that have not been tested within this model. Future studies should improve the model and test more moderating variables to investigate the reasons for the seen changes. Previous studies also discovered that there is a difference in effect size between different newspapers. Therefore, we need to be careful in concluding a change across newspaper language generally, as we have only tested a singular newspaper. Different effect sizes have also been found between different English variants. Previous studies discerned a higher effect for American English than for British English, theorising an influence of American English on British English (Štajner & Mitkov, 2011). Future studies should utilize these findings and include more varied articles when replicating this study, as our sample size, limited to one newspaper, is too small to encompass these changes and validate their explanations. Thus, future studies should include more variables that can explain potential language change, more varied measures of text complexity to reliably say

something about changes in newspaper language, as well as utilise a larger sampling pool, extracting multiple articles from different newspapers.

References

Books+Publishing. (2023, November 6). *US survey finds drop in reading participation*. https://www.booksandpublishing.com.au/articles/2023/11/06/240417/us-survey-finds-drop-in-reading-participation/

Canada Newswire. (n.d.). *The Ultimate Guide: Determining the Best Time & Day to Send a Press Release*. Retrieved July 22, 2024, from https://www.newswire.ca/resources/articles/best-time-press-release/

Center for Innovation and Sustainability in Local Media. (2019). *What history teaches us: How newspapers have evolved to meet market demands*. https://www.cislm.org/what-history-teaches-us-how-newspapers-have-evolved-to-meet-market-demands/

Choi, J. S., & Crossley, S. A. (2022). Advances in Readability Research: A New Readability Web App for English. *Proceedings - 2022 International Conference on Advanced Learning Technologies, ICALT 2022*, 1–5. https://doi.org/10.1109/ICALT55010.2022.00007

Conrad, P. (1999). *Modern Times, Modern Places*. Thames & Hudson. https://books.google.de/books?id=6OFkQgAACAAJ

Cook, S. E. (2004). New technologies and language change: Toward an anthropology of linguistic frontiers. *Annual Review of Anthropology*, *33*, 103–115. https://doi.org/10.1146/ANNUREV.ANTHRO.33.070203.143921

Covington, M. A., & Mcfall, J. D. (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

Eriksson, D. (2017). *Using the F-measure to test formality in sports reporting : A comparison of the language used in soccer and horse polo articles in two British newspapers*. https://urn.kb.se/resolve?urn=urn:nbn:se:kau:diva-67982

Flesch, R. (1948). A new readability yardstick. *The Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/H0057532

Gough, C. (2023, February 14). *Most popular live sporting events US 2022*. Statista. https://www-statista-com.ezproxy2.utwente.nl/statistics/830441/live-sports-impact-life/

Ham, L., Lentz, L., Pander Maat, H., & Stolk, F. (2019). Zijn romans en kranten sinds 1950 eenvoudiger geworden? *Tijdschrift Voor Nederlandse Taal- En Letterkunde*, *134*(4), 300–323. https://www.tntl.nl/index.php/tntl/article/view/531

Herring, S. C. (2003). Media and Language Change: Introduction. *Journal of Historical Pragmatics*, *4*(1), 1–17. https://doi.org/10.1075/JHP.4.1.02HER

Hinds, J., Connor, U., & Kaplan, R. B. (1987). Reader versus writer responsibility: A new typology. *Landmark Essays on ESL Writing*, 63–74.

Juola, P. (2003). The time course of language change. *Computers and the Humanities*, *37*(1), 77–96. https://doi.org/10.1023/A:1021839220474/METRICS

Kincaid, P. J., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. *Institute for Simulation and Training*, *56*. http://library.ucf.edu

King, V., & Gerisch, B. (2009). Zeitgewinn und Selbstverlust. Folgen und Grenzen der Beschleunigung. In *Zeitgewinn und {Selbstverlust}. {Folgen} und {Grenzen} der {Beschleunigung}*. Campus Verl. http://www.ciando.com/ebook/bid-35989

Konopliov, A. (2024, June 24). *US Newspaper Industry Statistics & Facts*. Redline Digital. https://redline.digital/us-newspapers-statistics/

Kristiansen, T., Coupland, N., Soukup, B., Moosmüller, S., Gregersen, F., Garrett, P., Selleck, C., Nuolijärvi, P., Vaattovaara, J., Östman, J.-O., Mattfolk, L., Stoeckle, P., Svenstrup, C., Leonard, S. P., Árnason, K., Hifearnáin, T. Ó., Murchadha, N. Ó., Vaicekauskienė, L., Grondelaers, S., … Stuart-Smith, J. (2011). *Language change and digital media: A review of conceptions and evidence*.

Kyle, K. (2016). Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication. *Applied Linguistics and English as a Second Language Dissertations*. https://doi.org/https://doi.org/10.57709/8501051

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the Validity of Lexical Diversity Indices Using Direct Judgements. *Language Assessment Quarterly*, *18*(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Leech, G., & Smith, N. (2006). Recent grammatical change in written English 1961-1992: some preliminary findings of a comparison of American with British English. *The Changing Face of Corpus Linguistics*, 185–204. https://doi.org/10.1163/9789401201797_013

Leech, G., & Smith, N. (2009). Change and constancy in linguistic change: How grammatical usage in written English evolved in the period 1931-1991. *Corpus Linguistics*, 173–200. https://doi.org/10.1163/9789042025981_011

Lindholm, C., & Vanhatalo, U. (2021). Handbook of Easy Languages in Europe. *Handbook of Easy Languages in Europe*, 661. https://doi.org/10.26530/20.500.12657/52628

Ljung, M. (1997). *Text complexity in British and American newspapers*. Brill. https://brill.com/edcollchap/book/9789004653351/B9789004653351_s007.xml

Majid, A. (2024, February 19). *US newspaper circulation 2023: Top 25 titles fall 14%*. Press Gazette. https://pressgazette.co.uk/media-audience-and-business-data/media_metrics/us-newspaper-circulation-2023/

Matheson, D. (2000). The birth of news discourse: Changes in news language in British newspapers, 1880-1930. *Media, Culture and Society*, *22*(5), 557–573. https://doi.org/10.1177/016344300022005002

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Mittmann, B. (2011). Geoffrey Leech, Marianne Hundt, Christian Mair and Nicholas Smith. Change in Contemporary English: A Grammatical Study. *Zeitschrift Für Anglistik Und Amerikanistik*, *59*(4), 425–427. https://doi.org/10.1515/ZAA-2011-0410

National Literacy Trust. (2023, September 4). *Children and Young People's Reading Research Report 2023*. https://literacytrust.org.uk/research-services/research-reports/children-and-young-peoples-reading-in-2023/

Publishers Weekly. (2023, October 18). *NEA Survey Finds Decline in Adult Reading*. https://www.publishersweekly.com/pw/newsbrief/index.html?record=4377

Rosa, H. (2013). *Beschleunigung und Entfremdung: Entwurf einer kritischen Theorie spätmoderner Zeitlichkeit*. Suhrkamp Verlag.

Schriver, K. A. (2017). Plain Language in the US Gains Momentum: 1940-2015. *IEEE Transactions on Professional Communication*, *60*(4), 343–383. https://doi.org/10.1109/TPC.2017.2765118

Siteimprove. (2024). *Readability tests*. https://help.siteimprove.com/support/solutions/articles/80000448325-readability-tests

Štajner, S. (2011). Towards a Better Exploitation of the Brown 'Family' Corpora in Diachronic Studies of British and American English Language Varieties. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011* (pp. 17–24). Association for Computational Linguistics. https://aclanthology.org/R11-2003

Štajner, S., & Mitkov, R. (2011). Diachronic Stylistic Changes in British and American Varieties of 20th Century Written English Language. In *Proceedings of the Second Student Research Workshop associated with RANLP 2011* (pp. 78–85). Association for Computational Linguistics. https://aclanthology.org/W11-4112

Stajner, S., & Mitkov, R. (2012). Using comparable corpora to track diachronic and synchronic changes in lexical density and lexical richness. *The 5th Workshop on Building and Using Comparable Corpora*.

Statista Market Insights. (2024, May). *Print Newspapers & Magazines - US | Market Forecast*. Statista. https://www-statista-com.ezproxy2.utwente.nl/outlook/amo/media/newspapers-magazines/print-newspapers-magazines/united-states#revenue

StudySmarter. (n.d.). *Finite Verbs: Definition, Examples, Types & More*. Retrieved July 22, 2024, from https://www.studysmarter.co.uk/explanations/english/english-grammar/finite-verbs/

U.S. Securities and Exchange Commission. (1998). *A Plain English Handbook: How to create clear SEC disclosure documents*. http://www.sec.gov/news/extra/handbook.htm

U.S. Vote Foundation. (n.d.). *Election Dates & Deadlines*. Retrieved July 22, 2024, from https://www.usvotefoundation.org/state-election-dates-and-deadlines

Wablieft. (2014). *De geschiedenis van Wablieft*. www.wablieft.be/krant/online-krant.

Westin, I. (2001). *The Language of English Newspaper Editorials from a 20th-Century Perspective*. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-801

Westin, I. (2016). Language Change in English Newspaper Editorials. *Language Change in English Newspaper Editorials*. https://doi.org/10.1163/9789004334007

Westin, I., & Geisler, C. (2002). A multi-dimensional study of diachronic variation in British newspaper editorials. *ICAME Journal*.

Wikipedia. (n.d.-a). *Content word*. Retrieved August 12, 2024, from https://en.wikipedia.org/wiki/Content_word

Wikipedia. (n.d.-b). *Proper noun*. Retrieved July 22, 2024, from https://en.wikipedia.org/wiki/Proper_noun

Appendix A: R Script

```
library(tidyverse)
library(broom)
library(dplyr)
library(reshape)
library(stringr)
library(GGally)
library(modelr)

setwd("C:/Users/Nadja/Desktop/Results")
data1 = read.csv("ARTE_politics.csv")
data1.5 = read.csv("ARTE_sports.csv")
data2 = read.csv("TAASSC_politics.csv")
data2.5 = read.csv("TAASSC_sports.csv")
data3 = read.csv("TAALED_politics.csv")
data3.5 = read.csv("TAALED_sports.csv")

data1 <-as.data.frame(data1)
data2 <-as.data.frame(data2)
data3 <-as.data.frame(data3)
data1.5 <-as.data.frame(data1.5)
data2.5 <-as.data.frame(data2.5)
data3.5 <-as.data.frame(data3.5)

#add new column
data1$type <- c("politics")
data2$type <- c("politics")
data3$type <- c("politics")
data1.5$type <- c("sports")
data2.5$type <- c("sports")
data3.5$type <- c("sports")

#merge data frames
df_list = list(data1, data1.5)
data1 <- merge_recurse(df_list)
df_list = list(data2, data2.5)
data2 <- merge_recurse(df_list)
df_list = list(data3, data3.5)
data3 <- merge_recurse(df_list)

#rename variables
data1 = data1 %>%
  rename(year = X.File.Name.)
data1 = data1 %>%
  rename(FRE = X.Flesch.Reading.Ease., FKGL = X.Flesch.Kincaid.Grade.Level., ARI =
X.Automated.Readability.Index.)
data2 = data2 %>%
  rename(year = filename)
data3 = data3 %>%
  rename(year = filename)
```

```
#remove prefix cell names
data1 <- data1 %>%
  mutate_at("year", str_replace, """, "")
data1 <- data1 %>%
  mutate_at("year", str_replace, """, "")

data2 <- data2 %>% mutate(year = basename(year))

data1 <- data1 %>%
  mutate_at("year", str_replace, ".txt", "")
data2 <- data2 %>%
  mutate_at("year", str_replace, ".txt", "")
data3<- data3 %>%
  mutate_at("year", str_replace, ".txt", "")

#keep only interesting columns
data1<-data1 %>%select("year", "type", "FKGL")
data2<-data2 %>%select("year", "type", "MLS", "C_S")
data3<-data3 %>%select("year", "type", "mtld_original_aw", "mtld_original_cw")

#merge data frames
df_list = list(data1, data2, data3)
data <- merge_recurse(df_list)

#change classes of variables for lm
data = data %>%
  mutate(year = as.numeric(year))
data = data %>%
  mutate(type = as.factor(type))

data %>% str()

#recode years into range 0 to 50
recode_years <- function(df, year_column) {
  min_year <- min(df[[year_column]], na.rm = TRUE)
  df <- df %>%
    mutate(!!sym(year_column) := (!!sym(year_column)) - min_year)
  return(df)
}
data <- recode_years(data, "year")

#descriptive statistics
summary(data$FKGL)
sd(data$FKGL)
summary(data$MLS)
sd(data$MLS)
summary(data$C_S)
sd(data$C_S)
summary(data$mtld_original_cw)
sd(data$mtld_original_cw)
```

```
#correlation matrix
data %>%
  ggpairs(data[, c("FKGL", "MLS", "C_S", "mtld_original_cw")])


### Flesch-Kincaid-Grade Level
#lm (test)
model1 <- data %>%
  filter(type== "sports") %>%
  lm(FKGL ~ year, data = .)
model1 %>% tidy()

#interaction model FKGL
modelFKGL <- data %>%
  lm(FKGL ~ year + type + year:type, data = .)
modelFKGL %>%
  tidy(conf.int=0.95)

#explained variance + degrees of freedom
sum <- modelFKGL %>%
  summary()
sum$r.squared

modelFKGL$df

#test lm without outlier
data[-c(23), ] %>%
  filter(type== "politics") %>%
  lm(FKGL ~ year, data = .) %>%
  tidy(conf.int=0.95)

#prep assumptions
data <- data %>%
  add_predictions(modelFKGL) %>%
  add_residuals(modelFKGL)
#assumption - normality
data %>%
  ggplot(aes(x = resid)) +
  geom_histogram() +
  facet_wrap(. ~ type)
#assumption - linearity & equal variance
data %>%
  ggplot(aes(x=year, y=resid, colour = type))+
  geom_point()
#independence
data%>%
  ggplot(aes(x=factor(type),y=resid))+
  geom_boxplot()
##independence & equal variance
data %>%
  ggplot(aes(x=FKGL, y=resid, colour = type))+
```

```
  geom_point()
#equal variance
data %>%
  ggplot(aes(x = pred, y = resid, colour = type)) +
  geom_boxplot()


#scatterplot FKGL
data %>%
    ggplot(aes(x = year, y = FKGL, colour = type)) +
    geom_point() +
    geom_smooth(method="lm",se=F)
#boxplot FKGL
data %>%
  ggplot(aes(x = year, y = FKGL, colour = type)) +
  geom_boxplot() +
  ylab('FKGL') +
  xlab('year')
#histogram FKGL
data %>%
  ggplot(aes(x = FKGL))+
  geom_histogram(aes(y= ..density..))


#### Mean Length of Sentence
#lm (test)
model2 <- data %>%
  filter(type== "politics") %>%
  lm(MLS ~ year, data = .)
model2 %>% tidy()

#interaction model MLS
modelMLS <- data %>%
  lm(MLS ~ year + type + year:type, data = .)
modelMLS %>%
  tidy(conf.int=0.95)

#explained variance
sum <- modelMLS %>%
  summary()
sum$r.squared

#test lm without outlier
data[-c(23), ] %>%
  filter(type== "politics") %>%
  lm(MLS ~ year, data = .) %>%
  tidy(conf.int=0.95)

#prep assumptions
data <- data %>%
  add_predictions(modelMLS) %>%
```

```
  add_residuals(modelMLS)
#assumption - normality
data %>%
  ggplot(aes(x = resid)) +
  geom_histogram() +
  facet_wrap(. ~ type)
#assumption - linearity & equal variance
data %>%
  ggplot(aes(x=year, y=resid, colour = type))+
  geom_point()
#independence
data%>%
  ggplot(aes(x=factor(type),y=resid))+
  geom_boxplot()
##independence & equal variance
data %>%
  ggplot(aes(x=MLS, y=resid, colour = type))+
  geom_point()
#equal variance
data %>%
  ggplot(aes(x = pred, y = resid, colour = type)) +
  geom_boxplot()


#scatterplot MLS
data %>%
  ggplot(aes(x = year, y = MLS, colour = type)) +
  geom_point() +
  geom_smooth(method="lm",se=F)
#boxplot MLS
data %>%
  ggplot(aes(x = year, y = MLS, colour = type)) +
  geom_boxplot() +
  ylab('MLS') +
  xlab('year')
#histogram MLS
data %>%
  ggplot(aes(x = MLS))+
  geom_histogram(aes(y= ..density..))


### Clauses per sentence
#lm (test)
model3 <- data %>%
  filter(type== "politics") %>%
  lm(C_S ~ year, data = .)
model3 %>% tidy()

#interaction model C_S
modelC_S <- data %>%
  lm(C_S ~ year + type + year:type, data = .)
```

```
modelC_S %>%
  tidy(conf.int=0.95)

#explained variance
sum <- modelC_S %>%
  summary()
sum$r.squared

#test lm without outlier
data[-c(23), ] %>%
  filter(type== "politics") %>%
  lm(C_S ~ year, data = .) %>%
  tidy(conf.int=0.95)

#prep assumptions
data <- data %>%
  add_predictions(modelC_S) %>%
  add_residuals(modelC_S)
#assumption - normality
data %>%
  ggplot(aes(x = resid)) +
  geom_histogram() +
  facet_wrap(. ~ type)
#assumption - linearity & equal variance
data %>%
  ggplot(aes(x=year, y=resid, colour = type))+
  geom_point()
#independence
data%>%
  ggplot(aes(x=factor(type),y=resid))+
  geom_boxplot()
##independence & equal variance
data %>%
  ggplot(aes(x=C_S, y=resid, colour = type))+
  geom_point()
#equal variance
data %>%
  ggplot(aes(x = pred, y = resid, colour = type)) +
  geom_boxplot()


#scatterplot C_S
data %>%
  ggplot(aes(x = year, y = C_S, colour = type)) +
  geom_point() +
  geom_smooth(method="lm",se=F)
#boxplot C_S
data %>%
  ggplot(aes(x = year, y = C_S, colour = type)) +
  geom_boxplot() +
  ylab('C_S') +
```

```
  xlab('year')
#histogram C_S
data %>%
  ggplot(aes(x = C_S))+
  geom_histogram(aes(y= ..density..))


### Measure of Textual Lexical Diversity
#lm (test)
model4 <- data %>%
  filter(type== "sports") %>%
  lm(mtld_original_cw ~ year, data = .)
model4 %>% tidy()

#interaction model MTLD
modelMTLD <- data %>%
  lm(mtld_original_cw ~ year + type + year:type, data = .)
modelMTLD %>%
  tidy(conf.int=0.95)

#explained variance + degrees of freedom
sum <- modelMTLD %>%
  summary()
sum$r.squared

modelMTLD$df

#prep assumptions
data <- data %>%
  add_predictions(modelMTLD) %>%
  add_residuals(modelMTLD)
#assumption - normality
data %>%
  ggplot(aes(x = resid)) +
  geom_histogram() +
  facet_wrap(. ~ type)
#assumption - linearity & equal variance
data %>%
  ggplot(aes(x=year, y=resid, colour = type))+
  geom_point()
#independence
data%>%
  ggplot(aes(x=factor(type),y=resid))+
  geom_boxplot()
##independence & equal variance
data %>%
  ggplot(aes(x=mtld_original_cw, y=resid, colour = type))+
  geom_point()
#equal variance
data %>%
  ggplot(aes(x = pred, y = resid, colour = type)) +
```

```
  geom_boxplot()


#scatterplot MTLD
data %>%
  ggplot(aes(x = year, y = mtld_original_cw, colour = type)) +
  geom_point() +
  geom_smooth(method="lm",se=F)
#boxplot MTLD
data %>%
  ggplot(aes(x = year, y = mtld_original_cw, colour = type)) +
  geom_boxplot() +
  ylab('MTLD') +
  xlab('year')
#histogram MTLD
data %>%
  ggplot(aes(x = mtld_original_cw))+
  geom_histogram(aes(y= ..density..))
```

Appendix B : Histograms

Figure 1

Histogram Showcasing the FKGL Distribution

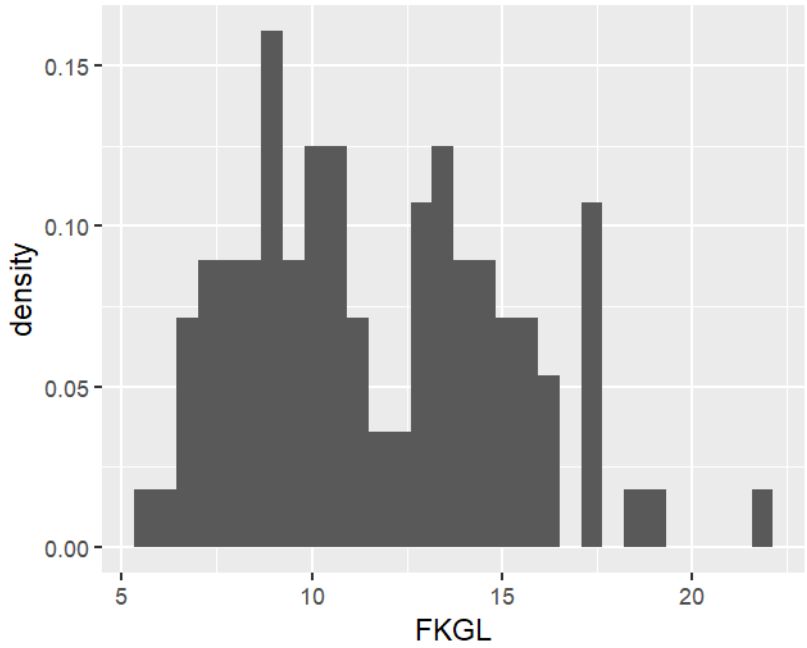


Figure 2

Histogram Showcasing the MLS Distribution

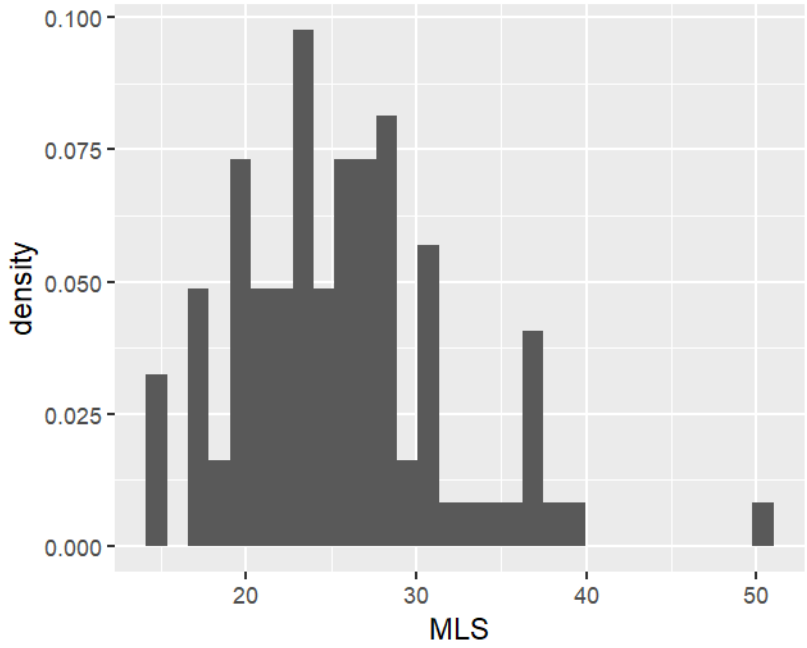


Figure 3

Histogram Showcasing the CS Distribution

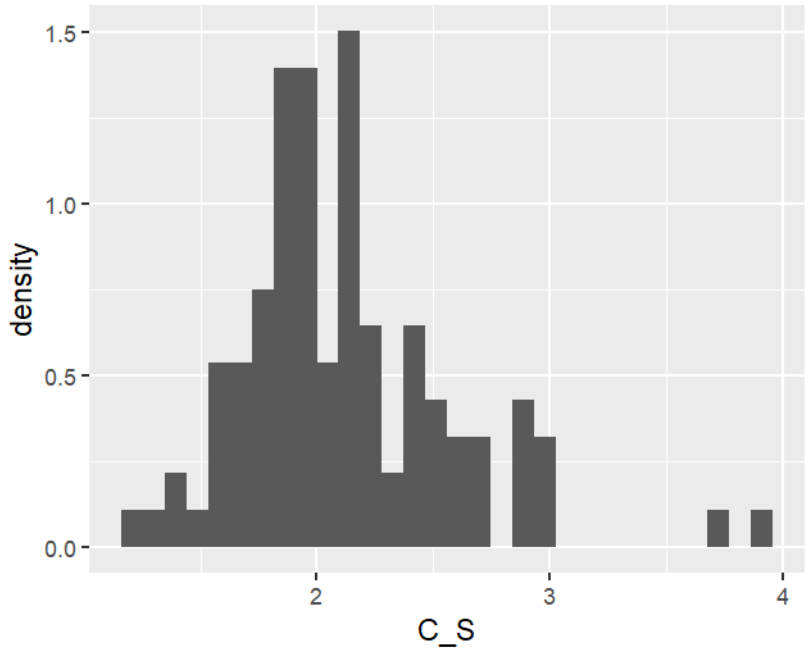
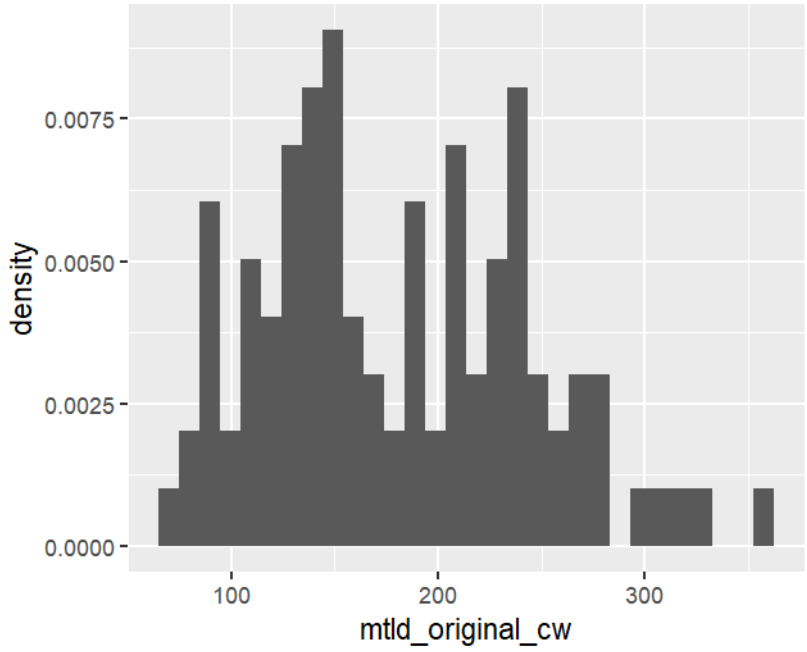


Figure 4

Histogram Showcasing the MTLD Distribution



Appendix C: Article Excerpts

Excerpt 1

Minimum FKGL (1993, Sports)

After Troy Aikman threw two touchdown passes to Michael Irvin today, he passed him his Most Valuable Player trophy. "Promise I won't drop it," Irvin said.

Aikman is the reticent one who this week turned down Jay Leno so he could take a nap. Irvin is the bombastic one who once said, "My biggest asset is my ego."

But tonight, they were the tag-team that eliminated the Buffalo Bills. Within a ludicrously quick span of 18 seconds, the two hooked up on a pair of second-quarter touchdowns that helped turn Super Bowl XXVII into a 52-17 bore. Aikman said he had been squeamish early in the game. "Had to talk myself into relaxing," Aikman said.

But the Buffalo Bills' defense had something to do with his anxiety. He said the Bills started in a "two-deep" zone defense, designed to make Irvin and his fellow wide receiver Alvin Harper disappear. Where to Go? "They were forcing us to go underneath," Aikman said.

Excerpt 2

Maximum FKGL (1997, Politics)

In their effort to save his life, lawyers for Timothy J. McVeigh today showed a Federal jury the magazine articles and videotapes about the Federal raid near Waco, Tex., that they said convinced Mr. McVeigh that the Government had declared war on the American people.

This morning the jury of seven men and five women, which will decide whether Mr. McVeigh should receive the death penalty for the Oklahoma City bombing, were read an affidavit in which he listed the videotapes he had seen and articles he had read about the raid and other articles that formed his political beliefs.

Prosecutors have said Mr. McVeigh planned and executed the Oklahoma City bombing, which killed 168 people on the second anniversary of the Texas raid, as an act of revenge against the Federal Government and in hopes of provoking a general uprising.

Excerpt 3

Minimum MLS (2008, Sports)

Plaxico Burress strode to the podium with an infant on his arm and his significant other standing by his side. He wore a championship hat and T-shirt, his eye black smeared across his face.

“Whew,” is all Burress, the Giants receiver and resident Nostradamus, could manage initially Sunday night.

What a week it had been. First, Burress predicted the Giants would do something almost no one thought possible: Beat the New England Patriots and ruin the first undefeated season in 35 years. Not only did he predict victory, but he also offered a final score.

Giants 23, Patriots 17.

Excerpt 4

Maximum MLS (1997, Politics)

See Excerpt 2

Excerpt 5

Minimum CS (1978, Sports)

Far from Madison Square Garden, perhaps a light-year away, one of the top college basketball teams in the East plays in its 1,200-seat gymnasium in Brooklyn tonight. The school is St. Francis, coached by Lou Rossini, and the opponent is the oldest on St. Francis’

schedule—City College. The rivalry began, during the 1921-22 season, when Nat Holman was in his third year as City's coach, and City has the edge in victories, 28-23. Rossini, former coach at New York University, is hoping to lead his team beyond the five boroughs when the time for postseason tournaments arrives. This game, at 8 o'clock, is preceded by one between the women's basketball teams from the two schools, at 6. Admission to the doubleheader is $3, with students admitted for $2. St. Francis is at 180 Remsen Street, in the Brooklyn Heights section. Call 522-2300 for information.

Excerpt 6

Maximum CS (1997, Politics)

See Excerpt 2

Excerpt 7

Minimum MTLD (2016, Sports)

After two one-year deals with CBS to televise eight “Thursday Night Football” games, the N.F.L. announced a two-year deal Monday with CBS and NBC worth an estimated $450 million annually.

CBS will show five Thursday night games starting in Week 2 of next season, and NBC will follow with five games starting in Week 11, a stretch that includes a Thanksgiving Day game that is not a part of the new contract.

“We had two priorities,” Sean McManus, chairman of CBS Sports, said. “It was very important to get the beginning half of the season as a terrific platform to launch our prime-time schedule.”

CBS paid $300 million for its Thursday night rights this season. Under the new contracts, CBS and NBC will pay $225 million each, or $45 million a game, up from the $37.5 million CBS paid this season.

Excerpt 8

Maximum MTLD (1996, Sports)

The Super Bowl may be over, but this is crunch time for Mike Reid. He is about to premiere his treasured new work on Broadway, the musical-theater capital of the country. And after a long career of trying to figure out what he wants to do with his life, he has chosen to set his music to a story about a football player who is trying to find himself.

At a recent rehearsal, Reid walked among the singers and production staff of his new football opera, "Different Fields," like a Gulliver among the Lilliputians. Reid has slimmed down from his 255-pound playing weight, but at 6 feet 3 inches he still has the bull neck and massive thighs of a National Football League defensive tackle.

An all-American, Reid co-captained Penn State's undefeated 1968 and 1969 teams, and was inducted into the College Football Hall of Fame in 1988. He went on to become an All-Pro pass rusher with the Cincinnati Bengals. "Different Fields," which he conceived and scored with lyrics by Sarah Schlesinger, was commissioned by the Metropolitan Opera Guild and Opera Memphis of Tennessee. It will begin a two-week run Wednesday at the New Victory Theater on 42d Street.